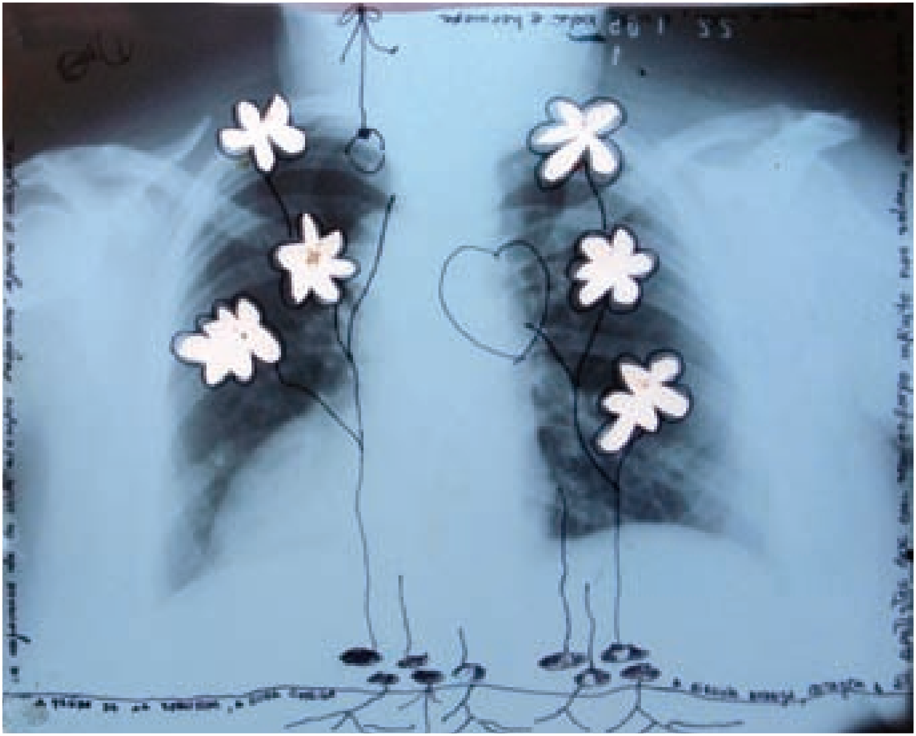
CHAPTER 21

Bioinformatics makes possible the interaction between very distant researchers. The large number of databases produced, sometimes with the use of very expensive technologies and processing programs, once kindly made freely available online, benefited researchers from all parts of the world, even those from low-resourced labs



Untitled – Lili Sarmiento
Mixed/X-ray film

‘… we are like gods, inventing
from the solitude of the world these signs
as bridges hugging the distances....’

***The Words are New* (Poem) – José Saramago**

Lidia Sarmiento Garcia San Miguel
La Habana, Cuba

Lidia Sarmiento Garcia San Miguel studied and graduated in Architecture from CUJAE (La Habana, Cuba) in 1979. She did her Doctorate Course in restoration and monument conservation at the Catedra Gaudi in Barcelona. Her Master Course in restoration and museology was from CENCREM (La Habana, Cuba). She lead the restoration project of historical areas in Old Havana from 1979 to 1995. She serves as Adviser of the Department of Patrimony of the State Secretariat, Ceara State, Brazil. She also works as a researcher in Patrimonial and Museologic Education. She did an educational brochure on TB, ‘Tuberculosis se cura, si’, for the TB Vaccines/TB Art 07 at Varadero, Cuba in 2007.

CHAPTER 21

Bioinformatics and Tuberculosis Vaccine Development: A Comparative Genomics Approach

Marcos Catanho and Antonio Basílio de Miranda

The Foundation of a New Era: The Emergence of Bioinformatics and Computational Biology

The origins of Bioinformatics and Computational Biology (1) can be traced back to the 1960s, when computers became essential tools in the field of Molecular Biology (as well as in all other areas). We can cite at least three main factors responsible for the emergence of these new disciplines: (i) the growing number of available protein sequences, providing at the same time an important source of new data and a set of relevant challenges impossible to cope without computer assistance; (ii) the idea that macromolecules carry information had become a fundamental part of the Molecular Biology conceptual framework; and, (iii) the availability of more powerful computers in the main universities and research centres (2).

Indeed, several computational techniques (algorithms and computer programs) for the analysis of structure, function, and evolution of nucleotide and protein sequences, as well as rudimentary protein databases, were already available towards the end of the 1960s (2, 3). New methods and approaches were introduced in the following decades, such as algorithms for sequence alignments, public domain databases, efficient data search and retrieval systems, more sophisticated methods for protein structure prediction, tools for the annotation and comparison of genes and genomes, and systems for functional genome analysis (4).

By the 1980s, when efficient algorithms were developed to cope with the ever-increasing amount of biological information, and computer implementations of these algorithms became available to the wider scientific community, Bioinformatics and Computational Biology could finally be recognized as independent disciplines, with their own challenges and achievements (3).

The consolidation of these disciplines occurred in the 1990s, with the emergence of supercomputers, powerful personal computers, and computer networks at global scale (Internet), as well as with the emergence of huge biological databases and the so-called *ome* projects: genome, transcriptome, and proteome, supported by the continuous progress in DNA sequencing techniques, the development of microarrays and biochip technologies, and mass spectrometry.

New Challenges, New Approaches: Genomics and the Comparative Genome Analysis

The pioneering initiative of the US 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 to 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 (5).

Hence, since the 1990s, the complete genetic code of almost 1,000 living organisms has been deciphered, and more than 3,000 genome projects are ongoing representing a huge variety or organisms of medical, commercial, environmental, and industrial interest, or comprising model organisms, important for the development of scientific research (6). Furthermore, with the time-to-finish of these projects becoming increasingly shorter, due to dramatic new developments in sequencing techniques and instrumentation (7), and the recent feasibility to obtain and analyse (complete or partial) genomic sequences of entire microbial communities recovered directly from uncultured environmental samples (the so-called metagenomics, also known as environmental genomics, ecogenomics or community genomics), new and important scientific breakthroughs and technological advances can be anticipated for the future.

Concomitantly, the achievement and analysis of numerous complete genome sequences (genomics), gene and protein expression data of cells, tissues and organs (supported by other high-throughput technologies such as transcriptomics and proteomics, respectively), combined with the development of high-throughput computing technologies and more efficient algorithms (provided by the emergence and consolidation of sciences such as Computation, Bioinformatics and Computational Biology), allowed *new* 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 (8), in differential expression analyses of genes and proteins (9), in protein three-dimensional structure predictions (10), in the process of metabolic reconstruction, and in the functional prediction of genes (11–16). Among these *new* approaches, we can distinguish the comparative genome analysis (also known as comparative genomics or genome comparison), which involves analysing and comparing genetic materials from diverse species or strains, aiming at investigating the structure, organization, and evolution of the compared genomes (and the corresponding species), as well as revealing the function of genes and non-coding regions in these genomes. In fact, microbial comparative genome analyses have undoubtedly made an important contribution to the elucidation of fundamental aspects of the genetics, biochemistry, and evolution of numerous species (8, 11, 17–32).

From a practical point of view, high-throughput technologies and approaches provide the research community the opportunity to not only expand our knowledge on the biology of living beings, but also enable us to develop new diagnostic systems, new drugs, more efficient vaccines, new prognostic markers, and a range of biotechnological applications. 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 efficient vaccines), diagnosis (development of faster and more accurate methods), and treatment (development of new drugs) of TB and other mycobacterial diseases (33). Some of these applications include: identification of unique genes and virulence factors, and metabolism reconstruction (34); characterization of pathogens and identification of new diagnostic and therapeutic targets (35); investigation of the molecular basis of pathogenesis and host range, and differences in phenotypes between clinical isolates and natural populations of pathogens (36–39); and, investigation of the genetic basis of virulence and drug resistance in TB-causing bacteria (40).

Comparing Microbial Genomes: Available Computational Resources for Comparative Genome Analysis of Prokaryotic Species

Currently, numerous databases and computational tools for microbial comparative genome analysis are publicly available as online services and/or stand-alone applications, comprising a range of functionalities and particular purposes (summarized in Table 21.1). With the exception of a few organism or group-specific databases presented in Table 21.1, all computational resources discussed here comprise (or can be applied to) mycobacterial genomes, particularly MTB complex species, and can potentially be applied to the identification of new drug targets, vaccine antigens, and diagnostics for TB.

Table 21.1 Main Databases and Computational Tools for Comparative Analysis of Prokaryotic Genomes

Category	Description	Examples
	Databases	
Generic and multifunctional	Dedicated to cover the universe of prokaryotic species genomes which have been completely sequenced, and to offer the required resources to search/retrieve precomputed and/or experimentally achieved data available for each species	BacMap, CMR, Genome Atlas, BLASTatlas, IMG, MBGD, Microbes Online, PLATCOM, PUMA2
Organism or group-specific	Dedicated to comparative analyses of particular microbial genomes, offering the required resources to search/retrieve precomputed and/or experimentally achieved data available for each species	GenoList, xBASE, GenoMycDB, BioHealthBase, TBDB, MGDD, MyBASE, MycoperonDB, LEGER, MolliGen, ShiBASE, Burkholderia Genome Database, Strepto-DB
Specialized	Dedicated to comparative analyses of particular features of genomes and their components (genes, proteins, protein domains, and other genomic regions)	COG, HAMAP, Hogenom, OMA Browser, OrthoMCL-DB, Round-Up, ATGC, FusionDB, IslandPath, ProtRepeatsDB, ORFanage, OrphanMine, SEED, TransportDB, STRING, KEGG, MetaCyc
Phylogenomic	Provide visualization and comparison of phylogenetic profiles, phylogeny reconstruction on the basis of conserved gene content or conservation of gene order across species, or analyses of phylogenetic orthologous groups	Phydbac, SHOT, PHOG
Genomic metadata	Dedicated to comparative studies of genomic metadata	Genome Properties, GenomeMine, SACSO

Category	Description	Examples
Computational Tools		
Interactive genome browsing programs	Provide interactive comparative visualization and browsing of pairs or groups of genomes (or genomic sequences) in different graphical environments, or interactive visual investigation of multiple alignments of genomic sequences	ACT, Cinteny, DNAVis, GeneOrder3.0, G-InforBIO, inGeno, SynBrowse, AutoGRAPH, GECO, GenColors, GenomeViz, MuGeN, SynView, CGView Server, ABC, CGAT, ComBo
Large-scale genomic sequences comparison programs	Based on large-scale sequence comparison involving multiple genomes using local or global alignment algorithms, or physical and genetic positions of specified groups of genes in whole genomes (or genomic sequences) and their similarity matrices	BioParser, BSR, COMPAM, GenomeBlast, GenomeComp, PSAT, M-GCAT, MUMmer, PipMaker/PipTools/MultiPipMaker/zPicture, VISTA, PyPhy, GenomePixelizer

Source: Catanho et al. 2007 (41).

Overall, databases for comparative analyses of prokaryotic genomes can be divided into five main categories, according to their principles and functionalities: (i) generic and multifunctional; (ii) organism or group-specific; (iii) specialized; (iv) phylogenomic; and, (v) genomic metadata (Table 21.1). In contrast, the computational tools can be grouped into (i) interactive genome browsing programs and (ii) large-scale genomic sequence comparison programs (Table 21.1). Certainly, these classifications are not definitive or perhaps the most suitable, since the purposes of and the analysis tools offered by these systems are naturally overlapping. Alternative classification schemes are therefore feasible and equally valid (42, 43).

Most generic and multifunctional databases presented in this section are dedicated to cover the universe of prokaryotic species (and sometimes eukaryotic species as well) whose genomes have been completely sequenced, and to offer the required resources to search/retrieve precomputed (mostly) and/or experimentally achieved data available for each species (BacMap, CMR, Genome Atlas, BLASTatlas, IMG, MBGD, Microbes Online, PLATCOM, PUMA2). The accessible information and the available searching/retrieval and analysis tools vary significantly from one database to another. They may comprise, for instance, physico-chemical, structural, statistical, functional, evolutionary, taxonomic, and/or phenotypical features associated to entire genomes or to their coding and/or non-coding regions, and searching/retrieval mechanisms based on keywords, gene/coding sequences and/or species names/identification numbers, or based on pairwise comparison of entire genomes, genomic sequences, or coding regions using local or global alignment algorithms. All these features also apply to organism or group-specific databases which are dedicated to particular microbes (GenoList, xBASE, GenoMycDB, BioHealthBase, TBDB, MGDD, MyBASE, MycoperonDB, LEGER, MolliGen, ShiBASE, Burkholderia Genome Database, Strepto-DB). Among these organism or group-specific databases we can find several resources fully dedicated to mycobacterial species:

GenoList. The GenoList (44) is a collection of databases dedicated to microbial genome analysis, providing a complete data set of protein and nucleotide sequences for selected species, as well as annotation and functional classification of these sequences. The **TubercuList**, **BoviList**, **BCGList**, **Leproma**, **BuruList**, and **MarinoList** databases are devoted to collect and integrate various aspects of the genomic information from MTB H37Rv, *M. bovis* AF2122/97, *M. bovis* BCG Pasteur 1173P2, *M. leprae* TN, *M. ulcerans* Agy99, and *M. marinum*, respectively.

xBASE. The xBASE (45) is another collection of databases, this one dedicated to bacterial comparative genome analyses. It 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, codon adaptation index (CAI), hydropathy, and aromaticity of the protein coding sequences in these bacteria. As part of this multi-microbial system, the **MycoDB** currently comprises comparative data from 20 completely sequenced or unfinished mycobacterial genomes—*M. avium* 104, *M. avium* subsp. *paratuberculosis* K-10, *M. bovis* AF2122/97, *M. bovis* BCG Pasteur 1173P2, *M. gilvum* PYR-GCK, *M. leprae* TN, *M. marinum* ATCC BAA-535, *M. smegmatis* MC2 155, *Mycobacterium sp.* (strains JLS, KMS, and MCS), MTB (strains C, CDC1551, F11, H37Ra (two representatives), H37Rv, and Haarlem), *M. ulcerans* Agy99, and *M. vanbaalenii* PYR-1.

GenoMycDB. The GenoMycDB (46) is a relational database for large-scale comparative analysis of completely sequenced mycobacterial genomes based on their predicted protein content. Currently, the database comprises six mycobacteria—MTB (strains H37Rv and CDC1551), *M. bovis* AF2122/97, *M. avium* subsp. *paratuberculosis* K10, *M. leprae* TN, and *M. smegmatis* MC2 155—providing for each of their encoded protein sequences the predicted subcellular localization, the assigned cluster of orthologous groups (COGs), features of the corresponding gene, and links to several important databases; in addition, pairs or groups of homologues between selected species/strains can be dynamically inferred based on user-defined criteria.

BioHealthBase. The BioHealthBase (47) provides a comprehensive genomic data repository for five different pathogenic organism groups considered a threat to public health. It also provides an analysis platform with suitable computational tools to assist genomic studies of these pathogens. One of these repositories is entirely dedicated to the available Mycobacterium-related data (combining *in silico* achieved and curated data in several instances) on genes and protein sequences, predicted structure, predicted orthologous groups, assigned gene ontology, protein function, protein localization, domains, motifs, metabolic pathways, and immunological epitopes. This repository also comprises experimental data on MTB essential genes and transposon mutants.

TBDB. Similarly to the BioHealthBase Mycobacterium database, TBDB (48) provides a comprehensive genomic data repository for MTB and related bacteria, combining (*in silico*) genome sequence and annotation data and (experimental) gene-expression data. It also provides an analysis platform with suitable computational tools to assist (comparative) genomic and gene-expression studies of these microorganisms. Annotated features of genes and genomes, predicted orthologous groups, operons and synteny blocks, as well as predicted and curated immunological epitopes and gene-expression patterns are accessible.

MGDD. The MGDD (49) comprises a data repository of genetic variations among different organisms belonging to the MTB complex. The MGDD system provides quick searches for precomputed SNPs, insertions, deletions, repeat expansions, and divergent sequences (inversions, duplications, and changes in synteny) in genomic regions of fully sequenced MTB complex species and strains genomes.

MyBASE. The MyBASE (50) is an integrated platform for functional and evolutionary genomic study of the genus *Mycobacterium*, comprising extensive literature review and data annotation on mycobacterial genome polymorphism, virulence factors, and essential genes.

MycoperonDB. The MycoperonDB (51) is a repository of known and computationally predicted operons and transcriptional units of (currently) five different mycobacteria—MTB (strains H37Rv and CDC1551), *M. bovis* AF2122/97, *M. avium* subsp. *paratuberculosis* K10, and *M. leprae* TN—whose genomes have been completely sequenced. Presently, it comprises 18,053 genes organized as 8,256 predicted operons and transcriptional units, providing literature links for experimentally characterized operons, and access to known promoters and related information.

On the other hand, there are an increasing number of databases dedicated to comparative analyses of particular features of genomes and their components (genes, proteins, protein domains, and other genomic regions). Among the features explored by these specialized databases, one may distinguish: conservation of orthologous genes (or proteins) across species (COG, HAMAP, Hogenom, OMA Browser, OrthoMCL-DB, RoundUp, ATGC); gene fusion/

fission events (FusionDB); occurrence of genomic islands (IslandPath); incidence of amino acid repetitions in proteins (ProtRepeatsDB); 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); configuration of protein interaction networks (STRING); incidence and conservation of metabolic pathways (KEGG, MetaCyc).

In the last twelve 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 for instance: visualization and comparison of phylogenetic profiles (co-occurrence of genes across species) (Phydbac); phylogeny reconstruction on the basis of conserved gene content or conservation of gene order (SHOT) across species; analysis of phylogenetic orthologous groups, that is, 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 information achieved from genomes and particular groups of genes in hundreds of microbial species, and also partially based on information compiled from published scientific researches. 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 (Table 21.1). They offer different graphical environments for visual comparison and browsing of pairs (ACT, Cinteny, DNAVis, GeneOrder3.0, G-InforBIO, inGeno, SynBrowse) or groups (AutoGRAPH, GECO, GenColors, GenomeViz, MuGeN, SynView, CGView Server) of genomes (or genomic sequences), and for visual investigation of multiple alignments of genomic sequences (ABC, CGAT, ComBo). Another group of tools is based on large-scale sequence comparison involving multiple genomes using local (BioParser, BSR, COMPAM, GenomeBlast, GenomeComp, PSAT) or global (M-GCAT, MUMmer, Pip-Maker/PipTools/MultiPipMaker/zPicture, VISTA, PyPhy) alignment algorithms, or using physical and genetic positions of specified groups of genes in whole genomes (or genomic sequences) and their similarity matrices (GenomePixelizer) (Table 21.1). 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: searching/retrieval mechanisms based on keywords, gene/coding sequence and/or species name/identification number; acquisition of functional gene annotations; phylogenetic reconstruction; detection of collinearity, synteny, gene duplication, orthologous and paralogous clusters, rearrangements, repetitions, inversions, insertions, deletions, restriction sites, motifs, and profiles, among others. These tools are available as online services and/or stand-alone applications.

Other Non-Comparative Mycobacterial Resources

Finally, there are other important non-comparative mycobacterial resources that could be helpful for the identification of new drug targets, vaccine antigens, and diagnostics.

The **TB Structural Genomics Consortium** (TBSGC) (52) is an organization devoted to support the determination and analysis of structures of proteins from MTB. Presently, 603 structures are accessible on the TBSGC website.

The **MTBreg** provides a database of conditionally regulated proteins in MTB, which includes information on proteins that are regulated by selected transcription factors or other regulatory proteins. Another database, **MTBRegList** (53), is dedicated to the analysis of gene expression and regulation data in MTB, containing predicted and characterized regulatory motifs cross-referenced with their respective transcription factor(s), experimentally identified transcription start sites, and DNA binding sites.

The **Proteome Database System for Microbial Research** at the Max Planck Institute for Infection Biology provides two-dimensional gel electrophoresis and mass spectrometry data of diverse microorganisms, including *M. bovis* and MTB, as well as comparative isotope-coded affinity tag–liquid chromatography/mass spectrometry (ICAT–LC/MS) data between MTB and *M. bovis* BCG.

Internet Resources

Generic and multifunctional databases
BacMap: http://wishart.biology.ualberta.ca/BacMap/
BLASTatlas: http://www.cbs.dtu.dk/ws/BLASTatlasCMR: http://cmr.tigr.org/
Genome Atlas: http://www.cbs.dtu.dk/services/GenomeAtlas/
IMG: http://img.jgi.doe.gov/
MBGD: http://mbgd.genome.ad.jp/
Microbes Online: http://www.microbesonline.org/

PLATCOM: http://platcom.informatics.indiana.edu/platcom/
PUMA2: http://compbio.mcs.anl.gov/puma2/

Genomic metadata databases
GenomeMine: http://www.genomics.ceh.ac.uk/GMINE/
Genome Properties: http://www.tigr.org/Genome_Properties/
SACSO: http://www.pasteur.fr/~tekaia/sacso.html

Interactive genome browsing programs
ABC: http://mendel.stanford.edu/sidowlab/downloads.html
ACT: http://www.sanger.ac.uk/Software/ACT/
AutoGRAPH: http://genoweb.univ-rennes1.fr/tom_dog/AutoGRAPH/
CGAT: http://mbgd.genome.ad.jp/CGAT/
CGView Server: http://stothard.afns.ualberta.ca/cgview_server/
Cinteny: http://cinteny.cchmc.org/
ComBo: http://www.broad.mit.edu/annotation/argo/
DNAVis: http://www.win.tue.nl/dnavis/
GECO: http://bioinfo.mikrobio.med.uni-giessen.de/geco2/GecoMainServlet
GenColors: http://gencolors.imb-jena.de/
GeneOrder3.0: http://binf.gmu.edu/genometools.html
GenomeViz: http://www.uniklinikum-giessen.de/genome/genomeviz/intro.html
G-InforBIO: http://wdcm.nig.ac.jp/inforbio/
inGeno: http://ingeno.bioapps.biozentrum.uni-wuerzburg.de/
MuGeN: http://genome.jouy.inra.fr/MuGeN/
SynBrowse: http://www.synbrowse.org/
SynView: http://www.ApiDB.org/apps/SynView/

Large-scale genomic sequences comparison programs
BioParser: http://www.dbbm.fiocruz.br/BioParser
BSR: http://www.microbialgenomics.org/BSR/
COMPAM: http://bio.informatics.indiana.edu/projects/compam/
GenomeBlast: http://bioinfo-srv1.awh.unomaha.edu/genomeblast/
GenomeComp: http://www.mgc.ac.cn/GenomeComp/
GenomePixelizer: http://www.atgc.org/GenomePixelizer/
M-GCAT: http://alggen.lsi.upc.es/recerca/align/mgcat/intro-mgcat.html
MUMmer: http://www.tigr.org/software/mummer/
PipMaker/PipTools/MultiPipMaker/zPicture: http://bio.cse.psu.edu/
PSAT: http://www.nwrce.org/psat
PyPhy: http://www.cbs.dtu.dk/staff/thomas/pyphy/
VISTA: http://www-gsd.lbl.gov/vista/

Organism or group-specific databases
BioHealthBase: http://www.biohealthbase.org/

Burkholderia Genome Database: http://www.burkholderia.com/GenoList: http://genolist.pasteur.fr/
GenoMycDB: http://www.dbbm.fiocruz.br/GenoMycDB
LEGER: http://leger2.gbf.de/cgi-bin/expLeger.pl
MGDD: http://mirna.jnu.ac.in/mgdd/
MolliGen: http://cbi.labri.fr/outils/molligen/
MycoperonDB: http://www.cdfd.org.in/mycoperondb/index.html
ShiBASE: http://www.mgc.ac.cn/ShiBASE/
Strepto-DB: http://oger.tu-bs.de/strepto_db
TBDB: http://www.tbdb.org/
xBASE: http://xbase.bham.ac.uk/

Phylogenomic databases
PHOG: http://bioinf.fbb.msu.ru/phogs/index.html
Phydbac: http://igs-server.cnrs-mrs.fr/phydbac/
SHOT: http://www.Bork.EMBL-Heidelberg.de/SHOT

Specialized databases
ATGC: http://atgc.lbl.gov/
COG: http://www.ncbi.nlm.nih.gov/COG
FusionDB: http://igs-server.cnrs-mrs.fr/FusionDB/
HAMAP: http://www.expasy.org/sprot/hamap/
Hogenom: http://pbil.univ-lyon1.fr/databases/hogenom.html
IslandPath: http://www.pathogenomics.sfu.ca/islandpath/
KEGG: http://www.genome.jp/kegg
MetaCyc: http://metacyc.org/
OMA Browser: http://omabrowser.org/
ORFanage: http://www.cs.bgu.ac.il/~nomsiew/ORFans/
OrphanMine: http://www.genomics.ceh.ac.uk/orphan_mine/faq.php
OrthoMCL-DB: http://orthomcl.cbil.upenn.edu/
ProtRepeatsDB: http://bioinfo.icgeb.res.in/repeats/
RoundUp: http://roundup.hms.harvard.edu/roundup/
SEED: http://theseed.uchicago.edu/FIG/index.cgi
STRING: http://string.embl.de/
TransportDB: http://www.membranetransport.org/

Other non-comparative mycobacterial resources
MTBreg: http://www.doe-mbi.ucla.edu/Services/MTBreg/
MTBRegList: http://www.usherbrooke.ca/vers/MtbRegList
Proteome Database System for Microbial Research: http://www.mpiib-berlin.mpg.de/2D-PAGE/
TB Structural Genomics Consortium: http://www.doe-mbi.ucla.edu/TB/

❐ References

1 BISTIC Definition Committee. NIH working definition of bioinformatics and computational biology. 2000. Available at: <http://www.bisti.nih.gov/docs/CompuBioDef.pdf> Accessed: 05 May 2009.

2 Hagen JB. The origins of bioinformatics. *Nat Rev Genet*, 2000; 1(3): 231–6.

3 Ouzounis CA and Valencia A. Early bioinformatics: the birth of a discipline—a personal view. *Bioinformatics*, 2003; 19(17): 2176–90.

4 Ouzounis C. Bioinformatics and the theoretical foundations of molecular biology. *Bioinformatics*, 2002; 18(3): 377–8.

5 HGP. HUMAN GENOME PROGRAM (USA). US Department of Energy. Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001.

6 GOLD. Genomes Online Database. Available at: <http://www.genomesonline.org/> Accessed: 05 May 2009.

7 Shendure JA, Porreca GJ, and Church GM. Overview of DNA sequencing strategies. *Curr Protoc Mol Biol*, 2008; Chapter 7: Unit.

8 Abby S and Daubin V. Comparative genomics and the evolution of prokaryotes. *Trends Microbiol*, 2007; 15(3): 135–41.

9 Patterson SD and Aebersold RH. Proteomics: the first decade and beyond. *Nat Genet*, 2003; 33 (Suppl): 311–23.

10 Ginalski K. Comparative modeling for protein structure prediction. *Curr Opin Struct Biol*, 2006; 16(2): 172–7.

11 Galperin MY and Koonin EV. Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol*, 2000; 18(6): 609–13.

12 Stein L. Genome annotation: from sequence to biology. *Nat Rev Genet*, 2001; 2(7): 493–503.

13 Gabaldon T and Huynen MA. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci*, 2004; 61(7–8): 930–44.

14 Francke C, Siezen RJ, and Teusink B. Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol*, 2005; 13(11): 550–8.

15 Lee D, Redfern O, and Orengo C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*, 2007; 8(12): 995–1005.

16 Skrabanek L, Saini HK, Bader GD, and Enright AJ. Computational prediction of protein-protein interactions. *Mol Biotechnol*, 2008; 38(1): 1–17.

17 Cordwell SJ. Microbial genomes and 'missing' enzymes: redefining biochemical pathways. *Arch Microbiol*, 1999; 172(5): 269–79.

18 Galperin MY and Koonin EV. Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. *Genetica*, 1999; 106(1–2): 159–70.

19 Huynen MA, Dandekar T, and Bork P. Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol*, 1999; 7(7): 281–91.

20 Kondrashov AS. Comparative genomics and evolutionary biology. *Curr Opin Genet Dev*, 1999; 9(6): 624–9.

21 Fraser CM, Eisen J, Fleischmann RD, Ketchum KA, and Peterson S. Comparative genomics and understanding of microbial biology. *Emerg Infect Dis*, 2000; 6(5): 505–12.

22 Koonin EV, Aravind L, and Kondrashov AS. The impact of comparative genomics on our understanding of evolution. *Cell*, 2000; 101(6): 573–6.

23 Wei L, Liu Y, Dubchak I, Shon J, and Park J. Comparative genomics approaches to study organism similarities and differences. *J Biomed Inform*, 2002; 35(2): 142–50.

24 Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol*, 2003; 1(2): 127–36.

25 Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, et al. Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat Biotechnol*, 2003; 21(7): 790–5.

26 Peregrin-Alvarez JM, Tsoka S, and Ouzounis CA. The phylogenetic extent of metabolic enzymes and pathways. *Genome Res*, 2003; 13(3): 422–7.

27 Coenye T, Gevers D, Van de PY, Vandamme P, and Swings J. Towards a prokaryotic genomic taxonomy. *FEMS Microbiol Rev*, 2005; 29(2): 147–67.

28 Delsuc F, Brinkmann H, and Philippe H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, 2005; 6(5): 361–75.

29 Huynen MA, Gabaldon T, and Snel B. Variation and evolution of biomolecular systems: Searching for functional relevance. *FEBS Lett*, 2005; 579(8): 1839–45.

30 Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, La T, et al. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics*, 2006; 6(3): 165–85.

31 Ochman H and Davalos LM. The nature and dynamics of bacterial genomes. *Science*, 2006; 311(5768): 1730–3.

32 Dutilh BE, van Noort V, van der Heijden RT, Boekhout T, Snel B, and Huynen MA. Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics*, 2007; 23(7): 815–24.

33 Ernst JD, Trevejo-Nunez G, and Banaiee N. Genomics and the evolution, pathogenesis, and diagnosis of tuberculosis. *J Clin Invest,* 2007; 117(7): 1738–45.

34 Gordon SV, Brosch R, Eiglmeier K, Garnier T, Hewinson RG, and Cole ST. Royal Society of Tropical Medicine and Hygiene Meeting at Manson House, London, 18 January 2001. Pathogen genomes and human health. Mycobacterial genomics. *Trans R Soc Trop Med Hyg*, 2002; 96(1): 1–6.

35 Fitzgerald JR and Musser JM. Evolutionary genomics of pathogenic bacteria. *Trends Microbiol*, 2001; 9(11): 547–53.

36 Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science*, 1999; 284(5419): 1520–3.

37 Brosch R, Pym AS, Gordon SV, and Cole ST. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol*, 2001; 9(9): 452–8.

38 Cole ST. Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J Suppl*, 2002; 36: S78–S86.

39 Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N, et al. Comparing genomes within the species Mycobacterium tuberculosis. *Genome Res*, 2001; 11(4): 547–54.

40 Randhawa GS and Bishai WR. Beneficial impact of genome projects on tuberculosis control. *Infect Dis Clin North Am*, 2002; 16(1): 145–61.

41 Catanho M, Miranda AB, and Degrave W. Comparing genomes: databases and computational tools for comparative analysis of prokaryotic genomes. *RECIIS: Elet J Commun Inf Innov Health*, 2007; 1(2): Sup334–Sup355.

42 Field D, Feil EJ, and Wilson GA. Databases and software for the comparison of prokaryotic genomes. *Microbiology*, 2005; 151(Pt 7): 2125–32.

43 Galperin MY. The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Res*, 2005; 33 (Database issue): D5–D24.

44 Fang G, Ho C, Qiu Y, Cubas V, Yu Z, Cabau C, et al. Specialized microbial databases for inductive exploration of microbial genome sequences. *BMC Genomics*, 2005; 6(1): 14.

45 Chaudhuri RR and Pallen MJ. xBASE, a collection of online databases for bacterial comparative genomics. *Nucleic Acids Res*, 2006; 34 (Database issue): D335–D337.

46 Catanho M, Mascarenhas D, Degrave W, and Miranda AB. GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. *Genet Mol Res*, 2006; 5(1): 115–26.

47 Greene JM, Collins F, Lefkowitz EJ, Roos D, Scheuermann RH, Sobral B, et al. National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. *Infect Immun*, 2007; 75(7): 3212–19.

48 Reddy TB, Riley R, Wymore F, Montgomery P, DeCaprio D, Engels R, et al. TB database: an integrated platform for tuberculosis research. *Nucleic Acids Res*, 2008.

49 Vishnoi A, Srivastava A, Roy R, and Bhattacharya A. MGDD: Mycobacterium tuberculosis genome divergence database. *BMC Genomics*, 2008; 9: 373.

50 Zhu X, Chang S, Fang K, Cui S, Liu J, Wu Z, et al. MyBASE: a database for genome polymorphism and gene function studies of Mycobacterium. *BMC Microbiol*, 2009; 9: 40.

51 Ranjan S, Gundu RK, and Ranjan A. MycoperonDB: a database of computationally identified operons and transcriptional units in Mycobacteria. *BMC Bioinformatics*, 2006; 7(Suppl 5): S9.

52 Terwilliger TC, Park MS, Waldo GS, Berendzen J, Hung LW, Kim CY, et al. The TB structural genomics consortium: a resource for Mycobacterium tuberculosis biology. *Tuberculosis (Edinb)*, 2003; 83(4): 223–49.

53 Jacques PE, Gervais AL, Cantin M, Lucier JF, Dallaire G, Drouin G, et al. MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis. *Bioinformatics*, 2005; 21(10): 2563–5.