INSTITUT PASTEUR

Research in Microbiology 158 (2007) 724–736

www.elsevier.com/locate/resmic

Annotation, comparison and databases for hundreds of bacterial genomes

Claudine Médigue [a,b,*], Ivan Moszer [c]

[a] CNRS UMR8030, Génomique Métabolique, 2 rue Gaston Crémieux, 91057 Evry Cedex, France
[b] Commissariat à l'Energie Atomique (CEA), Direction des Sciences du Vivant, Institut de Génomique, Genoscope, Laboratoire de Génomique Comparative, 2 rue Gaston Crémieux, 91057 Evry Cedex, France
[c] Pasteur Genopole Île-de-France, Plate-forme Intégration et Analyse Génomiques, Département Génomes et Génétique, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

Received 29 July 2007; accepted 26 September 2007
Available online 6 October 2007

Abstract

The multitude of bacterial genome sequences being determined has opened up a new field of research, that of comparative genomics. One role of bioinformatics is to assist biologists in the extraction of biological knowledge from this data flood. Software designed for the analysis and functional annotation of a single genome have, in consequence, evolved towards comparative genomics tools, bringing together the information contained in numerous genomes simultaneously. This paper reviews advances in the development of bacterial annotation and comparative analysis tools, and progress in the design of novel database structures for the integration of heterogeneous biological information.

Keywords: Genome annotation; Databases and servers; Computational tools; Comparative genomics; Data integration

1. Introduction

Large-scale genome sequencing has revolutionized biology over the past ten years, generating a vast amount of new information that has completely transformed our understanding of hundreds of species. At the time of writing, there are 570 publicly listed complete bacterial and archaeal genomes (GOLD database — http://www.genomesonline.org/). Beyond additional species, multiple strains of some bacteria are being sequenced, opening up the opportunity for detailed studies of genome evolution over the smallest time scales. Due to the importance of infectious diseases and the desire to maximize human health and wealth, biases towards the sequencing of pathogens and organisms of economic consequence are evident (ca. 80% according to the GOLD database). This bias is now being balanced by interest in isolates from the environment [71], and by projects that aim to cover the tree of life more extensively and to discover new biological functions.

Interpretation of raw DNA sequence data involves the identification and annotation of genes, proteins, and regulatory and/or metabolic pathways. This process is typically performed using sequence annotation pipelines (i.e. a variety of software modules) and, in some cases, human expertise to handle the annotations generated automatically. The reference databases, computational methods and knowledge that form the basis of these pipelines are constantly being developed. In addition, the rapid increase in new sequence data has necessitated the evolution of software resources from functional annotation of a single genome towards simultaneous analysis of information from multiple genomes [7]. Therefore, there is a natural shift towards the creation of tools for viewing and manipulating data in a comparative genomics context. Also, genome annotations need to be reprocessed on a regular basis to take into account the identification of newly characterized functions. Furthermore, large-scale functional analyses

* Corresponding author. CNRS UMR8030, Génomique Métabolique, 2 rue Gaston Crémieux, 91057 Evry Cedex, France.
E-mail addresses: cmedigue@genoscope.cns.fr (C. Médigue), moszer@pasteur.fr (I. Moszer).

0923-2508/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2007.09.009

generate additional data that contribute to the interpretation of genomic data. These considerations are driving the community to think about how to manage public collections of genomes in novel ways.

In this review, we discuss progress in the development of databases and computational tools for organizing and extracting biological meaning from the comparison of large sets of genomes. We will show that, despite all the benefits of continually growing collections of genomes and the development of powerful bioinformatics approaches, full interpretation of the content of these genomes remains challenging. We need to improve the quality and the speed of annotation, and to combine computational analysis with results of experimental studies (both large-scale investigations and focused assays), especially to elucidate the functions of the large number of hypothetical and orphan genes still found in genome databases. One solution to this challenge would be the development of integrated environments that combine and standardize information from a variety of sources and apply uniform (re-)annotation techniques.

2. What is genome annotation?

The process of sequencing and annotating bacterial genomes has become highly automated in these last years. Annotation, broadly speaking, is the extraction of biological knowledge from raw nucleotide sequences. Two main levels of genome annotation have been identified: the first corresponds to a static view of the genome whereas the second is associated with a more dynamic view [59] (Fig. 1).

2.1. Static view of genome annotation

In the initial step of the process, several bioinformatics methods are automatically linked up to predict the location of genes and to describe the cellular function of gene products. First, gene prediction programs, reviewed in [62], are executed to find regions that are likely to encode proteins or functional RNA products. Although very accurate for prokaryotes, gene calling programs are still liable to miss small genes or genes of atypical nucleotide composition. In addition, an increasing number of genomes are being released in "draft" form (i.e. before the finishing stage of a sequencing project) with high sequencing error rates, thus leading to errors in gene predictions. These initial predictions are then used for sequence similarity searches against generalist or specialized databases (Tables 1 and 2): information from hits above a similarity threshold is used to assign functions to proteins. The accuracy of this step depends not only on the software used for the

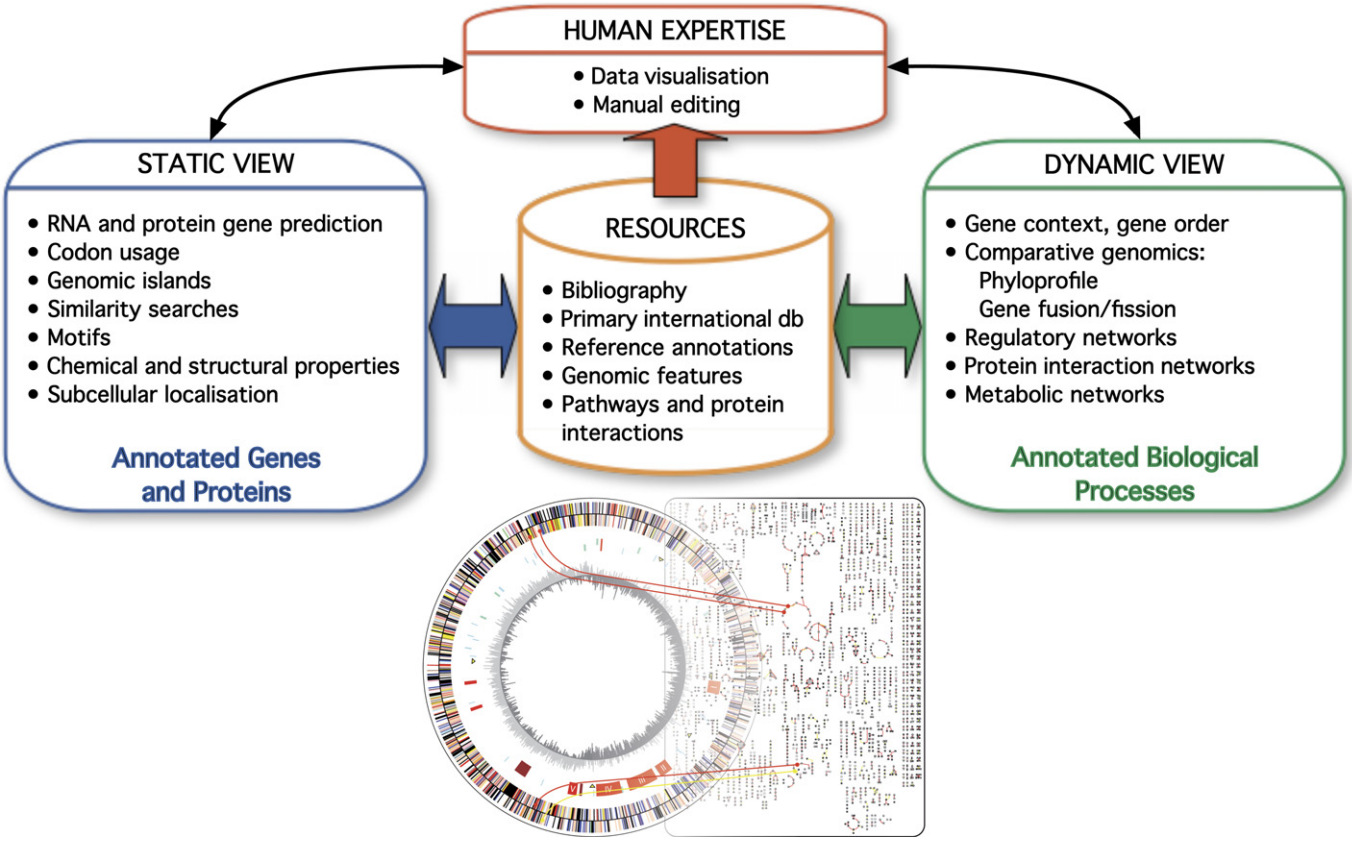

Fig. 1. General procedure used to annotate bacterial genome sequences. Automatic gene prediction and functional assignment, mainly based on sequence similarity and domain profiles ("STATIC VIEW" panel), provide information that can be used to identify interactions between the genomic components and to annotate biological processes ("DYNAMIC VIEW" panel). User-friendly interfaces are required for manual input of human expertise into these automatic predictions, an essential step to increase the specificity of the assigned biological functions ("HUMAN EXPERTISE" panel). These various levels of the annotation process are based on a wide range of generalist and thematic databases ("RESOURCES" panel; "db": databases).

Table 1
Software commonly used for bacterial genome annotation and comparison

DNA level annotation		
GeneMark	http://exon.gatech.edu/genemark/	Protein gene prediction
Glimmer	http://www.genomics.jhu.edu/Glimmer/	Protein gene prediction
SHOW	http://genome.jouy.inra.fr/ssb/SHOW/	Protein gene prediction
tRNAscan-SE	http://lowelab.ucsc.edu/tRNAscan-SE/	tRNA gene prediction
RNAmmer	http://www.cbs.dtu.dk/services/RNAmmer/	rRNA gene prediction
RepSeek	http://www.abi.snv.jussieu.fr/%98public/RepSeek/	Search for approximate repeats in complete DNA sequences
IslandPath	http://www.pathogenomics.sfu.ca/islandpath/	Identification of genomic islands
Protein level annotation		
BLAST	http://www.ebi.ac.uk/blast/	Compare a novel sequence with those contained in nucleotide and protein databases
InterProScan	http://www.ebi.ac.uk/InterProScan/	Search for domains/motifs in the InterPro database
COGNITOR	http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html	Compare a query sequence to the COG (Cluster of Orthologous Groups of proteins) database
PRIAM	http://bioinfo.genopole-toulouse.prd.fr/priam/	Detection of enzymatic function in a fully sequenced genome, based on all sequences available in the ENZYME database
GOAnno	http://bips.u-strasbg.fr/GOAnno/	BLAST search on the Gene Ontology database
PSORTb	http://www.psort.org/psortb/	Prediction of bacterial protein subcellular localization
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	Prediction of transmembrane helices in protein sequences
SignalP	http://www.cbs.dtu.dk/services/SignalP/	Prediction of signal peptide cleavage sites in protein sequences
Comparative genomic tools		
Mauve	http://gel.ahabs.wisc.edu/mauve/	Multiple genome alignments in the presence of large-scale evolutionary events
MOSAIC	http://mig.jouy.inra.fr/mig/mig_eng/presentation/project/mosaic	Define the set of backbones and loops in closely related bacterial genomes
ACT	http://www.sanger.ac.uk/Software/ACT/	Comparative genome analysis and visualization tools for multiple genome alignments
CGAT	http://mbgd.genome.ad.jp/CGAT/	
MaGe	http://www.genoscope.cns.fr/agc/mage/	Computation of gene order conservation (syntenies) between available bacterial genomes
Pathologic	http://biocyc.org/	Metabolic network reconstruction and comparative pathway analysis
PUMA2	http://compbio.mcs.anl.gov/puma2/	Metabolic pathway reconstruction
The SEED	http://theseed.uchicago.edu/FIG/	Comparative analysis and annotation tools using the subsystem approach
STRING	http://string.embl.de/	Search Tool for the Retrieval of Interacting Proteins
PyPhy	http://www.cbs.dtu.dk/staff/thomas/pyphy/	Reconstruction of phylogenetic relationships of complete microbial genomes
HoSeqI	http://pbil.univ-lyon1.fr/software/HoSeqI/	Automatically assign sequences to homologous gene families from the HOGENOM database

automatic annotation (see Section 3), but also on the quality of the primary resources, i.e. the annotation already stored in the databases, and on the quality of the sequence itself. To increase the value of functional annotation, some automatic procedures give a priority to the similarity results obtained with reference annotations of model organism(s) (see Sections 4 and 6).

In addition to the general prediction of gene functions, annotation pipelines can provide other types of information about the encoded proteins: chemical and structural properties (e.g. isoelectric point and molecular mass are important information for proteomic studies), subcellular localization (which has implications for both the function of the protein and its interactions with other proteins) and modular organization (Table 1). It is indeed important for the different domains of a protein to be characterized so as to avoid a well-known annotation error: a function is transferred to another protein that only shares one single module [26]. Consequently, sequence similarity search tools for thematic databases, such as motif/pattern or enzyme family databases (Tables 1 and 2), are also used. Assignment of Gene Ontology terms [31] can be directly obtained from these results. This classification, although not originally defined for bacterial genome annotation, is useful for the consideration of individual proteins in the context of the cell: what they do, i.e. the molecular function that describes the biochemical role of the protein (transporter, regulator, enzyme, structural protein, etc.); where they are found in the cell, i.e. their subcellular localization (cytoplasm, periplasm, cytoplasmic membrane, etc.); and what larger processes they participate in, i.e. the biological function that describes the role of the protein in the cell (metabolic pathway, signalling cascade, etc.).

2.2. Dynamic view of genome annotation

At this stage, information obtained by the first round of annotation is placed into a biological context to identify interactions between the genomic components: these are mainly protein-protein interactions, regulatory interactions and metabolite transformations. This leads to reconstruction of networks and subsequently to more complex dimensions of annotation [56]. The dynamic view of genome annotation is also useful to correct or to increase the specificity of the assigned functional annotations [16] (Fig. 1).

Metabolic-network reconstruction is an active field of research [22], and some annotation platforms include automatic prediction of these kinds of network (see Section 3). The annotated proteins that are characterized by an enzyme

Table 2
Main resources used for bacterial genome annotation and comparison

DNA sequence and annotation databases		
DDBJ	http://www.ddbj.nig.ac.jp/	Generalist International Nucleotide Sequence Databases
EMBL	http://www.ebi.ac.uk/embl/	Generalist International Nucleotide Sequence Databases
GenBank (Entrez)	http://www.ncbi.nlm.nih.gov/Genbank/	Generalist International Nucleotide Sequence Databases
GenomeReviews (Integr8)	http://www.ebi.ac.uk/GenomeReviews/	Integrated views of complete genomes and proteomes
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/	Curated non-redundant sequence database of genomes, transcripts and proteins
RNA sequence and annotation databases		
RDP-II	http://rdp.cme.msu.edu/	Sequences and tools for high-throughput rRNA analysis
Rfam	http://www.sanger.ac.uk/Software/Rfam/	Non-coding RNAs in complete genomes
NONCODE	http://noncode.bioinfo.org.cn/	Integrated knowledge database of non-coding RNAs
NcRNAdb	http://biobases.ibch.poznan.pl/ncRNA/	Non-coding RNAs database
FRNAdb	http://www.ncrna.org/	Platform for mining functional RNA candidates from non-coding RNAs
Protein sequence and annotation databases		
UniProt	http://www.expasy.uniprot.org/	Comprehensive catalogue of information on proteins (Swiss-Prot + TrEMBL + PIR)
HAMAP	http://expasy.org/sprot/hamap/	Identification and semi-automatic annotation of proteins that are part of well-conserved families in prokaryotes
Protein domain and motif databases		
Pfam	http://www.sanger.ac.uk/Software/Pfam/	Database of protein domain families based on seed alignments
Prosite	http://www.expasy.ch/prosite/	Documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles
SMART	http://smart.embl-heidelberg.de/	Domain-based sequence annotation resource
ProDom	http://prodom.prabi.fr/	Database of protein domain families automatically generated from Swiss-Prot and TrEMBL
PRINTS	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/	Database of protein fingerprints (group of conserved motifs used to characterise a protein family)
InterPro	http://www.ebi.ac.uk/interpro/	Integrative database of protein families, domains and functional sites (gathers databases listed above and others)
Protein family and classification databases		
COG	http://www.ncbi.nlm.nih.gov/COG/	Clusters of Orthologous Groups of proteins
TIGRFAMs	http://www.tigr.org/TIGRFAMs/	Database of protein families based on Hidden Markov Models
HOGENOM	http://pbil.univ-lyon1.fr/databases/hogenom.html	Database of homologous genes from fully sequenced organisms
GeneTrees	http://genetrees.vbi.vt.edu/	Database of pre-compiled alignments and gene phylogenies
Gene Ontology	http://www.geneontology.org/	Controlled vocabulary to describe gene and gene product attributes
KEGG BRITE	http://www.genome.ad.jp/kegg/brite.html	Functional hierarchies and binary relationships of biological entities
PANTHER	http://www.pantherdb.org/	Protein families subdivided into functionally related subfamilies
Protein and RNA structure databases		
WwPDB	http://www.wwpdb.org/	Deposition, data processing and distribution of protein structure data
MSD	http://www.ebi.ac.uk/msd/	Macromolecular structure database
RNABase	http://www.rnabase.org/	Annotated database of RNA structures
Protein interaction databases		
STRING	http://string.embl.de/	Database of known and predicted protein-protein interactions — direct (physical) and indirect (functional) associations
IntAct	http://www.ebi.ac.uk/intact/site/	Database system and analysis tools for protein interaction data
DIP	http://dip.doe-mbi.ucla.edu/	Database of experimentally determined interactions between proteins
Subcellular localization databases		
PSORTdb	http://db.psort.org/	Protein subcellular localization database for bacteria
TransportDB	http://www.membranetransport.org/	Predicted cytoplasmic membrane transport protein systems
Enzyme, metabolism and regulatory network databases		
ENZYME	http://www.expasy.ch/enzyme/	Repository of information relative to the nomenclature of enzymes
BRENDA	http://www.brenda.uni-koeln.de/	Enzyme functional data information system
MetaCyc	http://metacyc.org/	Multiorganism database of non-redundant, experimentally elucidated metabolic pathways
KEGG (PATHWAY)	http://www.genome.jp/kegg/	Collection of manually drawn pathway maps
PUMA2	http://compbio.mcs.anl.gov/puma2/	Grid-based high-throughput comparative and evolutionary analysis of genomes and metabolic pathways
PRODORIC	http://prodoric.tu-bs.de/	Prokaryotic database of gene regulation networks

(*continued on next page*)

Table 2 (*continued*)

Microbial genome databases and browsers		
IMG	http://img.jgi.doe.gov/	Integrated data management and analysis platform for Microbial Genomes
CMR	http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi	Comprehensive Microbial Resource displaying information on complete prokaryotic genomes
MBGD	http://mbgd.genome.ad.jp/	Microbial genome database for comparative analysis based on the automated construction of orthologous groups
MicrobesOnLine	http://www.microbesonline.org/	Browsing and comparing prokaryotic genomes
PEDANT	http://pedant.gsf.de/	Exhaustive automatic analysis of genomic sequences
GenoList	http://genolist.pasteur.fr/	Integrated environment for the analysis of microbial genomes
BacMap	http://wishart.biology.ualberta.ca/BacMap/	Interactive picture atlas of annotated bacterial genomes
XBASE	http://xbase.bham.ac.uk/	Collection of online databases for bacterial comparative genomics
Pathema	http://pathema.tigr.org/	Curatorial analysis of target organisms from the list of NIAID category A–C pathogens
PATRIC	http://patric.vbi.vt.edu/	The VBI PathoSystems Resource Integration Center
NMPDR	http://www.nmpdr.org/	National Microbial Pathogen Data Resource based on subsystem annotation
Phydbac	http://www.igs.cnrs-mrs.fr/phydbac/	Phylogenomic profiles of bacterial protein sequences
Generalist functional genomics databases		
ArrayExpress	http://www.ebi.ac.uk/arrayexpress/	Public database of microarray experiments and gene expression profiles
CIBEX	http://cibex.nig.ac.jp/	Gene expression database system
GEO	http://www.ncbi.nlm.nih.gov/geo/	Gene expression/molecular abundance data repository
SMD	http://genome-www5.stanford.edu/	Database of raw and normalized data from microarray experiments
SWISS-2DPAGE	http://expasy.org/ch2d/	Data on proteins identified on various 2-D PAGE and SDS-PAGE reference maps

A large number of biological databases are described in the annual database issue of the journal Nucleic Acids Research, and in the accompanying Molecular Biology Database Collection (http://www.oxfordjournals.org/nar/database/c/) — see also the CMS Molecular Biology Resource (http://restools.sdsc.edu/) and the newly established MetaBase (http://biodatabase.org/).

commission number (EC number) and/or an enzyme name are used to predict the set of metabolic pathways present in the organism, by searching in a pathway and reaction reference database (Table 2). The value of this similarity-based reconstruction is highly dependent on the quality of annotation, the completeness of the metabolic database and the criteria used for assessing the presence of a pathway. Although these automated reconstructions provide an overview of the metabolic capabilities of the studied microorganisms, detailed evaluation of these networks remains essential. Indeed, issues potentially leading to error include incorrect substrate specificity, multifunctional enzymes, reaction reversibility, cofactor usage and missing reactions that have no assigned gene [8].

Chromosomal context methodologies are also being employed to improve the accuracy of genome annotation (co-occurrence, gene order, gene fusion: see Section 5). They provide information for functional characterization of genes from their genome environment, e.g. by allowing a more confident identification of orthologues when sequence similarities are low. In addition, the analysis of co-localized genes provides clues about the functional interactions between the corresponding proteins, which is a first step towards the description of protein interaction networks.

3. Annotation platforms for bacterial genomes

The genome annotation process is complex and requires strong bioinformatics support including at least three main elements: (i) a pipeline for fully automated sequence annotation, which includes a large spectrum of bioinformatics tools; (ii) a coherent data management system (i.e. advanced biological data models and integrated databases); and (iii) several interactive graphical interfaces to organize and present the results in a helpful manner. These required features are not necessarily all available in the most common annotation platforms (Table 3).

The degree to which sequence annotation pipelines are automated varies. The first systems developed more than ten years ago, MAGPIE [25] and GeneQuiz [32], were strictly automatic, with original ‘‘reasoning’’ capabilities enabling the analysis of results produced by the different tools and the assignment of a specific biological function. They also provided an assessment of the annotation accuracy. Automatic tools with new functionalities continue to be developed, e.g. AutoFACT [38] which classifies sequences into various functional protein classes, and BASys [70], a web server that performs completely automated annotation of bacterial chromosomes. Indeed, online services are clearly the most convenient tools for research groups who lack the computing resources and expertise required to install or implement the software necessary for bacterial genome annotation. In most of these systems, user-friendly interfaces, which are essential for the manual input of expertise concerning automatic predictions to ensure high-quality annotation, are not available. This observation led to the development of annotation browsers and editors such as Artemis [6]. This system is a useful tool for reviewing and editing annotation, although the annotations themselves are necessarily provided by other programs.

Most of the existing systems offer two complementary functionalities: they generate automatic annotations and they provide graphical facilities for subsequent manual review of the predictions (Table 3). Examples of comprehensive

Table 3
Systems and platforms commonly used for bacterial genome annotation

Mainly automatic annotation systems		
MAGPIE	http://magpie.ucalgary.ca/	Software package for the automated annotation and presentation of DNA and protein sequences
GenQuiz	http://jura.ebi.ac.uk:8765/ext-genequiz/	Automatic annotation of protein sequences
AutoFACT	http://megasun.bch.umontreal.ca/Software/AutoFACT.htm	Automatic Functional Annotation and Classification Tool
BASys	http://wishart.biology.ualberta.ca/basys	Web accessible annotation system that is fully automatic
Annotation browsers/editors		
Artemis	http://www.sanger.ac.uk/Software/Artemis/	DNA sequence viewer and annotation tool
GENQUIRE	http://bioinformatics.org/Genquire/	Genome browser and annotation tool
Bluejay	http://bluejay.ucalgary.ca	Software for viewing sequence annotations
ASAP	http://asap.ahabs.wisc.edu/asap/home.php	Annotation system used to store, update and distribute genome sequence and gene expression data collected from enterobacteria primarily
Automatic and manual annotation systems		
ERGO	http://ergo.integratedgenomics.com/ERGO_supplement/	Provides automatic annotations and tools to analyse the conservation of gene context (commercial)
Pedant-Pro	http://www.biomax.de/products/pedantpro.php	Genome analysis package for comprehensive annotation and curation of DNA and protein sequences (commercial)
CAAT-Box	http://genopole.pasteur.fr/PF4/logiciels.html	Software package containing methods for the follow-up of the assembly phases and for initiating annotation during the finishing stage
IOGMA	http://www.genostar.com/fr/products.html	Integrated, interactive bioinformatics tool dedicated to systems biology (commercial)
GenDB	http://www.cebitec.uni-bielefeld.de/groups/brf/software/gendb_info/	Automatic annotation of genes and proteins using a large collection of software tools and user interfaces for expert annotation
SABIA	http://www.sabia.lncc.br/	System for Automated Bacterial Integrated Annotation
Manatee	http://manatee.sourceforge.net/	Web-based gene evaluation and genome annotation tool (automatic annotation provided by TIGR's annotation engine)
AGMIAL	http://genome.jouy.inra.fr/agmial/	Integrated system for bacterial genome annotation
MaGe	http://www.genoscope.cns.fr/agc/mage/	Annotation system that uses conservation of gene order to support predictions

annotation platforms include commercial systems such as ERGO [50] or Pedant-Pro (successor of PEDANT [24]), and open-source systems, such as GenDB [42], Manatee (TIGR, unpublished), SABIA [1] and AGMIAL [9]. However, the installation of these systems is not straightforward because they make use of various bioinformatics tools that must be installed independently. Finally, the availability of a large collection of genomes led, more recently, to the development of comparative annotation and analysis environments, such as MaGe, which enables the annotation of microbial genomes using genomic context and synteny results obtained using known bacterial genome sequences [69]. Indeed, the predictive power of chromosomal clustering has proven to be very effective for assigning putative functions (see Section 5).

4. Resources for genome annotation

Genome annotation relies on numerous data collections to infer knowledge by means of similarity analysis (Table 2). International nucleotide sequence databanks (DDBJ/EMBL/GenBank, joined in the International Nucleotide Sequence Database Collaboration — http://www.insdc.org/) are reference archives that play a key role in the construction of derived repositories containing a subset of sequences or for computational analysis based upon DNA information. However, these data collections, established almost 30 years ago, have faced new challenges since the onset of the genomic era. Many of these challenges have not been satisfactorily resolved: the maximum DNA sequence length that can be handled is not adapted to genome sequences, the huge number of associated features cannot be efficiently queried, updating annotation is not straightforward, etc. Parallel resources have therefore been developed with the aim of providing solutions to these issues. Genome Reviews at the EBI presents an up-to-date, standardized and comprehensively annotated version of complete genomes [37]. RefSeq at the NCBI provides an integrated and non-redundant set of nucleotide and protein sequences for organisms widely used in research [53]. These significant efforts remain, however, very generalist and are mostly driven by automatic procedures; as a result, they cannot replace specialized and expertly curated microbial resources (see below and Section 6).

A large collection of databases are available for proteins. The most widely accepted reference is UniProt [68], which originated from a merger between the Swiss-Prot and PIR protein databanks. The quality targets of this knowledge base are many: expertly curated annotations, inclusion of the relevant literature, numerous cross-references, etc. However, the exponential increase in sequence data has made the curation of all information an impossible task. Therefore, two separate sections have been defined: UniProtKB/Swiss-Prot contains manually validated protein entries, and UniProtKB/TrEMBL provides access to computationally annotated records. Also, the Swiss Institute of Bioinformatics leads a project — HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes) — to define rules which aim to semi-automatically annotate proteins

in the Swiss-Prot database that are part of well-conserved families in prokaryotic organisms [27].

Proteins are key players in cellular processes, hence the presence of a huge number of resources derived from the primary sequence databanks. Protein motifs and domains identified by various methods are made accessible in a variety of data collections (listed in Table 2). The InterPro site brings together many of these databases, thus providing a unified resource to search for protein signatures [45]. Most of these databases organize proteins into families according to their motifs and domains. Other levels of classification are defined using clustering procedures that lead to groups of proteins: for instance, the COG (Clusters of Orthologous Groups of proteins) database is built upon systematic BLASTP comparison of all proteins against all (from selected genomes), and subsequent construction of clusters containing at least three proteins from distantly-related species [66].

Many proteins exhibit catalytic activities; therefore the definition of enzymatic information provided by databases like BRENDA [4] can be particularly useful. Similarly, the formal description of metabolic pathways — either computationally predicted or experimentally described — can also be valuable, as exemplified by the KEGG [36] and BioCyc/MetaCyc [11] databases. In addition, the STRING online resource [73] gives an access to protein interaction data, by integrating known and predicted interactions from a large variety of organisms. Finally, recent large-scale endeavours in structural genomics are enhancing data collections such as PDB [5], used to store protein three-dimensional structures.

Functional RNAs that do not code for proteins are now known to be more frequent in genomes than originally thought and play important roles in various cellular processes. Initially termed small RNAs (sRNAs) in bacteria [61], and more generally non-coding RNAs (ncRNAs, including microRNAs, siRNAs, etc.) in higher organisms, these molecules have led to the creation of several dedicated databases (see Table 2).

Finally, organism-specific databases are an important data resource for the annotation of new bacterial genomes and the re-annotation of ‘‘old’’ genomes (see Section 6). Indeed these databases are usually carefully curated, with an emphasis on what makes the species of particular interest, and include support from manual investigation of the relevant literature. Obviously, model organisms are prominent references for related species: this is especially true for *Escherichia coli* [57] and *Bacillus subtilis* [44], the most extensively studied Gram-negative and Gram-positive bacteria, respectively.

5. Analysis and comparison of many bacterial genomes

Comparative genomics require the development of novel methods, databases and graphical interfaces for organizing and extracting biological information from the comparison of a large collection of complete and unfinished genome sequences [21]. Progress towards the development of such computational tools includes: (i) the development of online resources widely accessible to the scientific community; (ii) the improvement of comparative annotation methods, which have a significant impact on the accuracy of the functional annotation of genes; and (iii) the analysis of multiple isolates to describe and characterize species diversity.

5.1. Databases and servers for multiple genomes

It is of paramount importance that the data sets described in this article (primarily genomes and annotations) be made accessible to biologist users appropriately, to allow hypotheses about specific genomes or sets of genes to be tested. Several online resources provide the ability to view and manipulate pre-computed analyses and to access a large range of tools for genome analysis and comparative studies (Table 2).

Two main requirements trigger the development of specialized databases and associated integrated environments. First, the data should be logically and consistently organized in a non-redundant way: this can be achieved by the definition of ad hoc data models and the use of efficient database management systems. Secondly, specialized graphical user interfaces must be implemented in a community-driven way. Data access, querying, browsing and analysis have to be user-friendly and intuitive, and this is possible only if the interfaces are designed according to the reasoning of biologist users, not excluding innovative features thought of by database developers. Moreover, such tools should be coupled with rigorous procedures allowing curators to correctly update information and users to extract accurate data.

From a functional perspective, these environments may propose various levels of basic and advanced features. Several microbial genome sites provide consistent and comprehensive sets of annotations, together with a series of tools for genome querying, analysis and comparative studies [20]. Features range from simple selections of organisms and genes to complex multigenome queries, e.g. allowing the user to identify specific or shared proteins among a set of selected genomes. Examples of such specialized software environments include the Integrated Microbial Genomes (IMG) [41] and the Comprehensive Microbial Resource (CMR) [51] databases (Table 2). Recently, the GenoList genome browser (http://genolist.pasteur.fr/GenoList) was upgraded to provide an intuitive yet powerful multi-genome user interface, primarily designed to address biologists' requirements, and including original functionalities such as subtractive proteome analysis [13].

5.2. Genome annotation in a comparative genomics background

Techniques based on genomic context use the co-localization of genes in several genomes at various levels of proximity (chromosomal, metabolic, co-citation, etc.) and do not require sequence similarity for the genes to be annotated. In particular, genes co-localized on the chromosome tend to be functional neighbors, either in terms of expression patterns or network neighborhood [49]. Combined with similarity-based predictions, such information can be used to elucidate protein function [19]. This technique has been exploited recently and proved valuable for the accurate identification of candidates

for the missing genes of the lysine fermentation pathway in anaerobic organisms [39].

Novel methods are emerging for the annotation of a set of functionally related genes across a set of genomes, rather than the usual ''gene-by-gene'' annotation of one genome at a time [48]. Indeed, exploration of biological processes is more effective when realized at the scale of the global system, and gathering biological knowledge about several organisms simultaneously allows biologists to detect discrepancies and identify exceptions (i.e., the lack of a key enzymatic reaction in a pathway in several organisms may suggest the existence of an alternative route). Systems such as SEED [48] and Genome Properties [30] define a set of biological processes (e.g. metabolic pathways, secretion systems) and a set of functional roles that are necessary to complete a process. For each process, a two-dimensional matrix is obtained in which columns describe roles and rows describe organisms. A cell defines (or not) which gene(s) encode(s) a particular role in a particular organism. This matrix can be used as a starting point to identify variants concerning a process by gathering organisms sharing the same profile (i.e. the same functional roles). In addition, Genome Properties proposes rules, mainly based on role essentiality, to determine automatically whether or not a process exists in a given organism [30]. The integration of these approaches into annotation platforms should improve annotations such that automatic analysis of functional variants is possible, including mapping missing genes and locating gene candidates for experimental validation [75].

5.3. Analysis of closely related species and multiple strains

Analysis of genomes from closely related species can help in the identification of novel genes and other features such as gene fusion/fission and pseudogenes, which only become apparent in a comparative genomics context. These phenomena also include lineage-specific genes, which can be characterized efficiently if many related sequences are exploited. Identification of genetic differences between entire genomes allows correlation of the differences with biological function, providing insight into selective evolutionary pressures and patterns of gene transfer or loss. This has proven to be particularly pertinent for virulence analysis of pathogenic strains [55]. For example, a comparative analysis of several extraintestinal pathogenic *E. coli* (ExPEC) strains has shown that the ability to accumulate and express a variety of virulence-associated genes distinguishes ExPEC from many commensals and that different pathogenic strategies exist in ExPEC [10].

One important discovery of the genomic era is that the gene pool of several strains of a microbial species, called the ''pan-genome'', is far larger than the number of genes present in any single genome [67]. This observation, which was shown to be particularly noticeable for *E. coli* [12], could be linked to bacterial niche adaptations. Indeed, bacterial genomes consist of a backbone called ''core genome'' (which contains the genetic information for basic cellular functions) supplemented with adaptive or selfish genome modules (which are not necessarily all present in a given strain of the species) [23]. This ''flexible gene pool'' is often localized in mobile genetic elements (plasmids, transposons, bacteriophages, etc.); the integration of these elements into bacterial chromosomes can lead to permanent genomic elements called Genomic Islands (GIs) [7].

A number of methods to detect GIs have been developed, and rely mostly on atypical sequence composition (including GC content and codon usage) and associated annotation features (for example tRNA and integrase genes) [33]. More recently, a new method based on compositional biases using variable order motif distributions has been published [72]. A comparative genomics approach can also be used: two or more sequences are aligned to identify unique genomic regions that are putative GIs [14]. Once identified, the nature of the GIs (pathogenicity, symbiosis, metabolic islands, or other [76]) and the nature of the donor genome, must be characterized. One strategy is to calculate the local signature of identified GIs and look for similar signatures in a database of genomic signatures [18]; another makes use of phylogenetic tools to date the transfer event during evolution and to identify the origin of the suspected alien DNA fragments [29]. However, the identification of Gis' origin remains a difficult task and there are few validated examples.

A multistrain project might also involve Single Nucleotide Polymorphism (SNP) analyses to address evolutionary issues. SNPs are very short sequence differences between highly similar sequences; they may affect either coding or non-coding sequences. SNPs are important features of a genomic sequence: for example, they may reveal genes that contribute to the adaptation of the bacteria to different environmental stimuli, allowing them to shift from commensalism to pathogenicity. For example, SNP analysis in *Staphylococcus epidermidis* showed that virulence factors and surface proteins evolved quickly, in parallel with the pathogenicity environment, whereas translation, ribosomal structure and biogenesis-related proteins, which were submitted to considerable structural and functional constraints, evolved slowly [74].

In the near future, innovative sequencing technologies [40] will deliver a huge number of new sequences, both finished and draft genomes. Analysis of SNPs of thousands of genome sequences from the same species will raise new questions about the organization and interpretation of these data. Similarly, the availability of numerous draft genome sequences will complement standard multi-locus sequence typing (MLST) techniques for population analysis (''population genomics'').

6. Automatic versus expert annotation: does it still make sense?

Genomes were generally subjected to careful curation and review in the early days of sequencing, but this is no longer true. High-throughput and low-cost sequencing methods have resulted in ever-increasing sequencing capabilities, and most often the resulting genomes receive only automatic annotation, with very little input from human expertise. Consequently,

although bioinformatics tools are continually improving, some genomes remain poorly annotated even today, especially in terms of functional annotation. Against the background of the avalanche of bacterial genome sequences now being published, the question arises as to the value of thorough manual review of genome annotations. Expert input is most likely still required, although the amount of work greatly depends on the nature of the sequencing project: either a genome for which close relatives are already known or a new representative of a taxonomic group.

In the case of projects involving the annotation of one or several genomes for which a closely related species, considered as a reference genome, is already available in public databanks or in thematic databases, an ideal strategy would involve two steps: first, an update of the reference genome annotation, i.e. a re-annotation process, and second, the annotation of the new genome based on the re-annotated one.

The re-annotation step is very important: even when a number of annotation sources are very accurate, continual updating of genome annotations for a large number of species is not straightforward [58]. Indeed, databases and computational methods are constantly evolving and the re-processing of automatic functional annotations should be performed on a regular basis. In addition, new experimentally derived functional information is being continually generated, and can prove useful, for example, for modifying the annotation of genes of ''putative'' or unknown function. This requires systematic exploration of bibliographic references (http://www.pubmed.gov/), an element of paramount importance for collecting sound fundamental knowledge about model organisms.

The second step involves the transfer of the reliable up-to-date reference annotation to ''strong'' orthologues in the newly sequenced genomes. Incidentally, annotation platforms will soon include procedures that rely on the ability to cluster proteins from related genomes into orthologous groups and new interfaces allowing annotators to view evidence associated with each protein in the cluster and make annotation decisions about the group as a whole (see Section 5.2 and Table 3). Then, the manual annotation process only needs to be performed for the specific genes or regions, which generally do not represent more than 20% of the gene products when the new and reference genome sequences belong to the same species.

There is also the interesting issue of handling projects related to the annotation of bacterial genomes that are evolutionarily distant, and very different, from the minuscule fraction of microbial species we know today [34]. Examples are studies of prokaryotic species belonging to novel bacterial genus (e.g. *Herminiomonas arsenoxidans*, which is able to metabolize arsenic and to efficiently colonize toxic environments [46]), and some metagenome projects enabling the reconstruction of complete genomes. The case of an anaerobic ammonium oxidation (anammox) community dominated by *Kuenenia stuttgartiensis* is interesting and illustrative: genome annotation has led to novel candidate genes for hydrazine and ladderane metabolism [64]. Obviously, it is impossible to implement, in a computer, the rules that a manual annotator would follow because the biology of such organisms presents numerous exceptions or novel features. The meticulous work of expert annotation is thus very often the only way to discover novelties.

Contributing to the dispute about automatic vs. expert annotation, Raes and collaborators recently published a surprising result [54]: they estimated that, in completely sequenced genomes, the fraction of proteins to which at least some functional features can be automatically assigned is close to 75% using similarity searches alone, and 85% if genomic context methods are also used. However, for non-housekeeping genes, these automatic annotations might often be erroneous and of limited use for biologists (e.g. unknown enzyme/transporter substrate, very short domain, etc.): manual curation undoubtedly remains necessary.

7. Integrated databases and analysis: the key for new discoveries

Integration of annotation data and functional genomics information is a major issue in tackling a systems-level understanding of biology [47] (Fig. 2). High-throughput genomics technologies are briefly reviewed in reference [35] and can be summarized as follows: (i) transcriptomics: using high-density DNA chips or sequence-based technologies such as SAGE (Serial Analysis of Gene Expression), gene expression profiling generates descriptions of overall transcriptional changes according to various conditions; (ii) proteomics: two-dimensional gel electrophoresis or liquid chromatography techniques, coupled with mass spectrometry, are used to unravel the complement of proteins present in the cell; (iii) interactomics: techniques such as yeast two-hybrid (protein-protein interactions) and ChIP-chip (protein-DNA interactions) screenings reveal relationships between cellular components; (iv) localisomics: this term defines the subcellular localization of individual biological objects or complexes, in particular secreted proteins in bacteria (the secretome); (v) metabolomics: this describes the analysis of the concentration and dynamics (the fluxome) of the full set of metabolites (e.g. by NMR spectrometry) to advance towards a complete description of the cell physiology; (vi) phenomics: this is the comprehensive study of phenotypes, made possible by improvements in techniques for generating mutant strains and determining cell fitness or viability.

These technologies share common features: they operate on a large-scale with high-throughput performances, and they rely primarily on information concerning the genome. These approaches, commonly designed ''omics'' studies, generate a flood of data sets which require both dedicated management systems and ad hoc strategies for integrating and mining heterogeneous information (Fig. 2). Although a few general repositories have been set up for storing this type of information, such as Gene Expression Omnibus (GEO) [3] for microarray experiments (Table 2), they provide views that are less integrated and comprehensive than organism-focused resources. Moreover, functional genomics experiments are prone to yielding findings whose validity and interpretation are open to question: they may give false-positive and/or false-negative

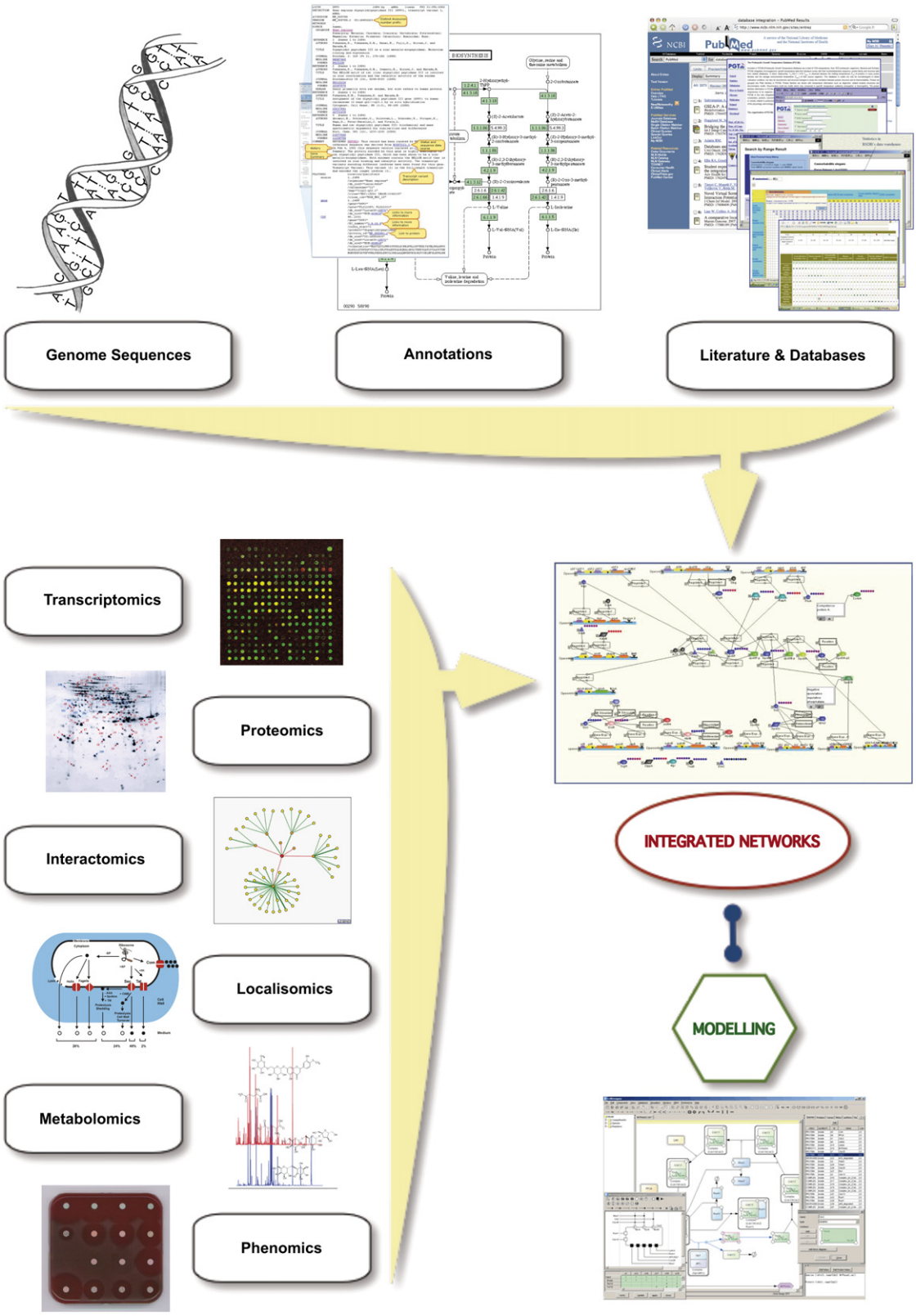

Fig. 2. Integration of genome and functional information. A full interpretation and modelling of cellular processes requires the integration of heterogeneous data sets: genome sequence, annotations (components and pathways) and elements from the associated literature, and also high-throughput "omics" data (e.g. transcriptome, proteome, interactome, metabolome, etc.). The ultimate goal of data integration is to use computational methods to model biological processes, and thereby combine the different networks to construct a comprehensive description of nearly all components and interactions within the cell.

results and their reproducibility is not always extremely high (although this also depends on the level of confidence used in the analysis of the results) [2]. Therefore, as many correlations as possible should be performed to better elucidate the meanings of the outcomes of ''omics'' experiments. Furthermore, genome annotation and databases provide information for their interpretation and validation. Symmetrically, results from functional genomics experiments may contribute to the refinement and investigation of genome annotation, hence the need for data integration.

The concept of integration is threefold. First, data standardization is a key requirement for overall interpretation of experimental results obtained at different sites and times by independent researchers [65]. One emblematic example is in the field of transcriptomics, where the organization MGED (Microarray Gene Expression Data) has worked to define standards and ontologies for microarray data (see a review in Ref. [63]).

Second, data have to be interconnected in such a way that simultaneous and coordinated access is made possible for establishing complex correlations between heterogeneous pieces of information. Data federation allows distributed management and curation in specialized structures, whereas data warehousing provides unified information through a centralized entry point [60]. The weakness of these two approaches include hindered communications and regular updates between distant data resources, respectively. A few integrative systems are emerging for bacteria, in particular the SYSTOMONAS platform dedicated to *Pseudomonas* species, which combines both data warehouse and database interoperability concepts [15]. Other examples include the EchoBASE system for post-genomics data on *E. coli* [43] and the Insieme software package for *Staphylococcus aureus* [52].

Finally, data integration also means coordinated data mining. This involves the design and the implementation of adequate modelling procedures, and the development of computational methods by which genes or proteins that behave similarly under various conditions can be grouped [28]. Approaches developed in this field usually address three specific tasks: identifying the connections between cellular components, decomposing this network scaffold into modules, and developing models to simulate and predict network behaviour that gives rise to cellular phenotypes [17,35].

8. Conclusion

The amount of genomic data available on prokaryotic organisms will continue to increase in an exponential manner in the short and medium term, both in quantity and in complexity. Facing this deluge of data, specialized tools for annotating, integrating and mining the information are absolutely essential if new biological knowledge and understanding is to be extracted from the primary data.

The interpretation of the raw sequence and of the genomic components identified remains the first and necessary step in deciphering cell functioning, and this is based on automatic pipelines and/or expertly curated knowledge. Novel approaches and new tools to tackle these issues are emerging, based upon innovative concepts which take multiple dimensions of genome annotation into account (comparative genomics, functional modules, etc.). Genomic databases also act as information hubs that connect logically organized high-quality data and relevant search and analysis tools, accessible through graphical user interfaces designed to respond to the needs of biologists. An added value is generated by the synergy between the individual components of an integrated software environment, and this contributes to knowledge discovery through fine data exploration, guided by suitable human-computer interactions and visual representations.

References

[1] Almeida, L.G., Paixao, R., Souza, R.C., Costa, G.C., Barrientos, F.J., Santos, M.T., Almeida, D.F., Vasconcelos, A.T. (2004) A System for Automated Bacterial (genome) Integrated Annotation—SABIA. Bioinformatics 20, 2832–2833.

[2] Bammler, T., Beyer, R.P., Bhattacharya, S., Boorman, G.A., Boyles, A., Bradford, B.U., Bumgarner, R.E., Bushel, P.R., Chaturvedi, K., Choi, D., Cunningham, M.L., Deng, S., Dressman, H.K., Fannin, R.D., Farin, F.M., Freedman, J.H., Fry, R., Harper, A., Humble, M.C., Hurban, P., Kavanagh, T.J., Kaufmann, W.K., Kerr, K.F., Jing, L., Lapidus, J.A., Lasarev, M.R., Li, J., Li, Y., Lobenhofer, E.K., Lu, X., Malek, R.L., Milton, S., Nagalla, S.R., O'malley, J.P., Palmer, V.S., Pattee, P., Paules, R.S., Perou, C.M., Phillips, K., Qin, L.X., Qiu, Y., Quigley, S.D., Rodland, M., Rusyn, I., Samson, L.D., Schwartz, D.A., Shi, Y., Shin, J.L., Sieber, S.O., Slifer, S., Speer, M.C., Spencer, P.S., Sproles, D.I., Swenberg, J.A., Suk, W.A., Sullivan, R.C., Tian, R., Tennant, R.W., Todd, S.A., Tucker, C., Van Houten, B., Weis, B.K., Xuan, S., Zarbl, H. (2005) Standardizing global gene expression analysis between laboratories and across platforms. Nat. Methods 2, 351–356.

[3] Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res. 35, D760–D765.

[4] Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I., Schomburg, D. (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. Nucleic Acids Res. 35, D511–D514.

[5] Berman, H., Henrick, K., Nakamura, H., Markley, J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 35, D301–D303.

[6] Berriman, M., Rutherford, K. (2003) Viewing and annotating sequence data with Artemis. Brief. Bioinform. 4, 124–132.

[7] Binnewies, T.T., Motro, Y., Hallin, P.F., Lund, O., Dunn, D., La, T., Hampson, D.J., Bellgard, M., Wassenaar, T.M., Ussery, D.W. (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct. Integr. Genomics 6, 165–185.

[8] Borodina, I., Nielsen, J. (2005) From genomes to *in silico* cells via metabolic networks. Curr. Opin. Biotechnol 16, 350–355.

[9] Bryson, K., Loux, V., Bossy, R., Nicolas, P., Chaillou, S., van de Guchte, M., Penaud, S., Maguin, E., Hoebeke, M., Bessières, P., Gibrat, J.-F. (2006) AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res. 34, 3533–3545.

[10] Brzuszkiewicz, E., Brüggemann, H., Liesegang, H., Emmerth, M., Olschläger, T., Nagy, G., Albermann, K., Wagner, C., Buchrieser, C., Emody, L., Gottschalk, G., Hacker, J., Dobrindt, U. (2006) How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. Proc. Natl. Acad. Sci. U.S.A. 103, 12879–12884.

[11] Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C.,

Zhang, P., Karp, P.D. (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. 34, D511–D516.

[12] Chen, S.L., Hung, C.S., Xu, J., Reigstad, C.S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R.R., Ozersky, P., Armstrong, J.R., Fulton, R.S., Latreille, J.P., Spieth, J., Hooton, T.M., Mardis, E.R., Hultgren, S.J., Gordon, J.I. (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. Proc. Natl. Acad. Sci. U.S.A. 103, 5977–5982.

[13] Chetouani, F., Glaser, P., Kunst, F. (2002) DiffTool: building, visualizing and querying protein clusters. Bioinformatics 18, 1143–1144.

[14] Chiapello, H., Bourgait, I., Sourivong, F., Heuclin, G., Gendrault-Jacquemard, A., Petit, M.-A., El Karoui, M. (2005) Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. BMC Bioinformatics 6, 171.

[15] Choi, C., Münch, R., Leupold, S., Klein, J., Siegel, I., Thielen, B., Benkert, B., Kucklick, M., Schobert, M., Barthelmes, J., Ebeling, C., Haddad, I., Scheer, M., Grote, A., Hiller, K., Bunk, B., Schreiber, K., Retter, I., Schomburg, D., Jahn, D. (2007) SYSTOMONAS—an integrated database for systems biology analysis of *Pseudomonas*. Nucleic Acids Res. 35, D533–D537.

[16] Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., Palsson, B.Ø. (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429, 92–96.

[17] De Keersmaecker, S.C., Thijs, I.M., Vanderleyden, J., Marchal, K. (2006) Integration of omics data: how well does it work for bacteria? Mol. Microbiol. 62, 1239–1250.

[18] Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., Deschavanne, P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res. 33, e6.

[19] Enault, F., Suhre, K., Claverie, J.-M. (2005) Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics 6, 247.

[20] Field, D., Feil, E.J., Wilson, G.A. (2005) Databases and software for the comparison of prokaryotic genomes. Microbiology 151, 2125–2132.

[21] Field, D., Wilson, G., van der Gast, C. (2006) How do we compare hundreds of bacterial genomes? Curr. Opin. Microbiol. 9, 499–504.

[22] Francke, C., Siezen, R.J., Teusink, B. (2005) Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol. 13, 550–558.

[23] Fraser-Liggett, C.M. (2005) Insights on biology and evolution from microbial genome sequencing. Genome Res. 15, 1603–1610.

[24] Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., Mewes, H.W. (2001) Functional and structural genomics using PEDANT. Bioinformatics 17, 44–57.

[25] Gaasterland, T., Sensen, C.W. (1996) Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. Biochimie 78, 302–310.

[26] Galperin, M.Y., Koonin, E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol. 1, 55–67.

[27] Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A.H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C.J., Lachaize, C., Veuthey, A.-L., Gasteiger, E., Bairoch, A. (2003) Automated annotation of microbial proteomes in SWISS-PROT. Comput. Biol. Chem. 27, 49–58.

[28] Ge, H., Walhout, A.J., Vidal, M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet. 19, 551–560.

[29] Gogarten, J.P., Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3, 679–687.

[30] Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N., White, O. (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. Bioinformatics 21, 293–306.

[31] Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258–D261.

[32] Hoersch, S., Leroy, C., Brown, N.P., Andrade, M.A., Sander, C. (2000) The GeneQuiz web server: protein functional analysis through the Web. Trends Biochem. Sci. 25, 33–35.

[33] Hsiao, W.W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B.B., Brinkman, F.S. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS Genet. 1, e62.

[34] Hugenholtz, P. (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol. 3. REVIEWS0003.

[35] Joyce, A.R., Palsson, B.Ø. (2006) The model organism as a system: integrating 'omics' data sets. Nat. Rev. Mol. Cell Biol. 7, 198–210.

[36] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354–D357.

[37] Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I., Apweiler, R. (2005) Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res. 33, D297–D302.

[38] Koski, L.B., Gray, M.W., Lang, B.F., Burger, G. (2005) AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics 6, 151.

[39] Kreimeyer, A., Perret, A., Lechaplais, C., Vallenet, D., Médigue, C., Salanoubat, M., Weissenbach, J. (2007) Identification of the last unknown genes in the fermentation pathway of lysine. J. Biol. Chem. 282, 7191–7197.

[40] Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M., Chen, Y., Chen, Z., Dewell, S., Du, L., Fierro, J., Gomes, X., Godwin, B., He, W., Helgesen, S., Ho, C., Ho, C., Irzyk, G., Jando, S., Alenquer, M., Jarvie, T., Jirage, K., Kim, J., Knight, J., Lanza, J., Leamon, J., Lefkowitz, S., Lei, M., Li, J., Lohman, K., Lu, H., Makhijani, V., McDade, K., McKenna, M., Myers, E., Nickerson, E., Nobile, J., Plant, R., Puc, B., Ronan, M., Roth, G., Sarkis, G., Simons, J., Simpson, J., Srinivasan, M., Tartaro, K., Tomasz, A., Vogt, K., Volkmer, G., Wang, S., Wang, Y., Weiner, M., Yu, P., Begley, R., Rothberg, J. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

[41] Markowitz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N., Kyrpides, N.C. (2006) The integrated microbial genomes (IMG) system. Nucleic Acids Res. 34, D344–D348.

[42] Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., Puhler, A. (2003) GenDB—an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 31, 2187–2195.

[43] Misra, R.V., Horler, R.S., Reindl, W., Goryanin, I.I., Thomas, G.H. (2005) EchoBASE: an integrated post-genomic database for *Escherichia coli*. Nucleic Acids Res. 33, D329–D333.

[44] Moszer, I., Jones, L.M., Moreira, S., Fabry, C., Danchin, A. (2002) SubtiList: the reference database for the *Bacillus subtilis* genome. Nucleic Acids Res. 30, 62–65.

[45] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P.S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C.,

McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Orengo, C., Petryszak, R., Selengut, J.D., Sigrist, C.J., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H., Yeats, C. (2007) New developments in the InterPro database. Nucleic Acids Res. 35, D224–D228.

[46] Muller, D., Médigue, C., Koechler, S., Barbe, V., Barakat, M., Talla, E., Bonnefoy, V., Krin, E., Arsène-Ploetze, F., Carapito, C., Chandler, M., Cournoyer, B., Cruveiller, S., Dossat, C., Duval, S., Heymann, M., Leize, E., Lieutaud, A., Lièvremont, D., Makita, Y., Mangenot, S., Nitschke, W., Ortet, P., Perdrial, N., Schoepp, B., Siguier, P., Simeonova, D.D., Rouy, Z., Segurens, B., Turlin, E., Vallenet, D., Van Dorsselaer, A., Weiss, S., Weissenbach, J., Lett, M.-C., Danchin, A., Bertin, P.N. (2007) A tale of two oxidation states: bacterial colonization of arsenic-rich environments. PLoS Genet. 3, e53.

[47] Ng, A., Bursteinas, B., Gao, Q., Mollison, E., Zvelebil, M. (2006) Resources for integrative systems biology: from data through databases to networks and dynamic system models. Brief. Bioinform 7, 318–330.

[48] Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., Vonstein, V. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 33, 5691–5702.

[49] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N. (1999) Use of contiguity on the chromosome to predict functional coupling. In Silico Biol. 1, 93–108.

[50] Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov Jr., E., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., Bhattacharyya, A., Burd, H., Gardner, W., Hanke, P., Kapatral, V., Mikhailova, N., Vasieva, O., Osterman, A., Vonstein, V., Fonstein, M., Ivanova, N., Kyrpides, N. (2003) The ERGO genome analysis and discovery system. Nucleic Acids Res. 31, 164–171.

[51] Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K., White, O. (2001) The comprehensive microbial resource. Nucleic Acids Res. 29, 123–125.

[52] Plikat, U., Voshol, H., Dangendorf, Y., Wiedmann, B., Devay, P., Müller, D., Wirth, U., Szustakowski, J., Chirn, G.W., Inverardi, B., Puyang, X., Brown, K., Kamp, H., Hoving, S., Ruchti, A., Brendlen, N., Peterson, R., Buco, J., Oostrum, J., Peitsch, M.C. (2007) From proteomics to systems biology of bacterial pathogens: approaches, tools, and applications. Proteomics 7, 992–1003.

[53] Pruitt, K.D., Tatusova, T., Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D65.

[54] Raes, J., Harrington, E.D., Singh, A.H., Bork, P. (2007) Protein function space: viewing the limits or limited by our view? Curr. Opin. Struct. Biol. 17, 362–369.

[55] Raskin, D.M., Seshadri, R., Pukatzki, S.U., Mekalanos, J.J. (2006) Bacterial genomics and pathogen evolution. Cell 124, 703–714.

[56] Reed, J.L., Famili, I., Thiele, I., Palsson, B.Ø. (2006) Towards multidimensional genome annotation. Nat. Rev. Genet. 7, 130–141.

[57] Riley, M., Abe, T., Arnaud, M.B., Berlyn, M.K., Blattner, F.R., Chaudhuri, R.R., Glasner, J.D., Horiuchi, T., Keseler, I.M., Kosuge, T., Mori, H., Perna, N.T., Plunkett, G., Rudd, K.E., Serres, M.H., Thomas, G.H., Thomson, N.R., Wishart, D., Wanner, B.L. (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. Nucleic Acids Res. 34, 1–9.

[58] Salzberg, S.L. (2007) Genome re-annotation: a wiki solution? Genome Biol. 8, 102.

[59] Stein, L. (2001) Genome annotation: from sequence to biology. Nat. Rev. Genet. 2, 493–503.

[60] Stein, L.D. (2003) Integrating biological databases. Nat. Rev. Genet. 4, 337–345.

[61] Storz, G., Haas, D. (2007) A guide to small RNAs in microorganisms. Curr. Opin. Microbiol. 10, 93–95.

[62] Stothard, P., Wishart, D.S. (2006) Automated bacterial genome analysis and annotation. Curr. Opin. Microbiol. 9, 505–510.

[63] Strömbäck, L., Hall, D., Lambrix, P. (2007) A review of standards for data exchange within systems biology. Proteomics 7, 857–867.

[64] Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., Taylor, M.W., Horn, M., Daims, H., Bartol-Mavel, D., Wincker, P., Barbe, V., Fonknechten, N., Vallenet, D., Segurens, B., Schenowitz-Truong, C., Médigue, C., Collingro, A., Snel, B., Dutilh, B.E., Op den Camp, H.J., van der Drift, C., Cirpus, I., van de Pas-Schoonen, K.T., Harhangi, H.R., van Niftrik, L., Schmid, M., Keltjens, J., van de Vossenberg, J., Kartal, B., Meier, H., Frishman, D., Huynen, M.A., Mewes, H.W., Weissenbach, J., Jetten, M.S., Wagner, M., Le Paslier, D. (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature 440, 790–794.

[65] Swertz, M.A., Jansen, R.C. (2007) Beyond standardization: dynamic software infrastructures for systems biology. Nat. Rev. Genet. 8, 235–243.

[66] Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

[67] Tettelin, H., Masignani, V., Cieslewicz, M., Donati, C., Medini, D., Ward, N., Angiuoli, S., Crabtree, J., Jones, A., Durkin, A., Deboy, R., Davidsen, T., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J., Hauser, C., Sundaram, J., Nelson, W., Madupu, R., Brinkac, L., Dodson, R., Rosovitz, M., Sullivan, S., Daugherty, S., Haft, D., Selengut, J., Gwinn, M., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K., Smith, S., Utterback, T., White, O., Rubens, C., Grandi, G., Madoff, L., Kasper, D., Telford, J., Wessels, M., Rappuoli, R., Fraser, C. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA 102, 13950–13955.

[68] The UniProt Consortium (2007) The Universal Protein Resource (UniProt). Nucleic Acids Res. 35, D193–D197.

[69] Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., Médigue, C. (2006) MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. 34, 53–65.

[70] Van Domselaar, G.H., Stothard, P., Shrivastava, S., Cruz, J.A., Guo, A., Dong, X., Lu, P., Szafron, D., Greiner, R., Wishart, D.S. (2005) BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res. 33, W455–W459.

[71] Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., Smith, H.O. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74.

[72] Vernikos, G.S., Parkhill, J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics 22, 2196–2203.

[73] von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B., Bork, P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. 35, D358–D362.

[74] Wei, W., Cao, Z., Zhu, Y.L., Wang, X., Ding, G., Xu, H., Jia, P., Qu, D., Danchin, A., Li, Y. (2006) Conserved genes in a path from commensalism to pathogenicity: comparative phylogenetic profiles of *Staphylococcus epidermidis* RP62A and ATCC12228. BMC Genomics 7, 112.

[75] Ye, Y., Osterman, A., Overbeek, R., Godzik, A. (2005) Automatic detection of subsystem/pathway variants in genome analysis. Bioinformatics 21(Suppl. 1), i478–i486.

[76] Zhang, R., Zhang, C.T. (2004) A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. Bioinformatics 20, 612–622.