Chapter 6
Bioinformatics Approaches in Studying Microbial Diversity

Mohammad Tabish, Shafquat Azim, Mohammad Aamir Hussain, Sayeed Ur Rehman, Tarique Sarwar, and Hassan Mubarak Ishqi

Abstract Proper understanding of molecular sequences, identification and phylogenetics of microorganisms are very important in many branches of biological science. Generation of genomic DNA sequence data from different organisms including microbes requires the application of bioinformatics tools for their analysis. Bioinformatics tools including BLAST, multiple sequence alignment tools etc. are used to analyze nucleic acid and amino acid sequences for phylogenetic affiliation. The emerging fields of comparative genomics and phylogenomics require the substantial knowledge and understanding of computational methods to handle the large scale data involved. The introduction of comparative rRNA sequence analysis represents a major milestone in the history of microbiology. Also single gene based phylogenetic inference and alternative global markers including elongation and initiation factors, RNA polymerase subunits, DNA gyrases, heat shock and RecA proteins are of immense importance. The analysis of the sequence data involves four general steps including: (i) selection of a suitable molecule or molecules, (ii) acquisition of molecular sequences, (iii) multiple sequence alignment and (iv) phylogenetic evaluation. This chapter explains in detail how raw data of molecular sequences from any microbe may be used for the detection and identification of microorganisms using computer based bioinformatics tools.

Keyword Bioinformatics tool • BLAST • Sequence alignment • DNA Chips • Gene identification

M. Tabish (✉) • S. Azim • M.A. Hussain • S.U. Rehman • T. Sarwar • H.M. Ishqi
Department of Biochemistry, Faculty of Life Sciences, Aligarh Muslim University,
Aligarh, Uttar Pradesh 202002, India
e-mail: tabish.biochem@gmail.com

A. Malik et al. (eds.), *Management of Microbial Resources in the Environment*,
DOI 10.1007/978-94-007-5931-2_6, © Springer Science+Business Media Dordrecht 2013

1 Introduction

There has been a flood of nucleic acid sequence information, bioinformatics tools and phylogenetic inference methods in public domain databases, literature and World Wide Web space. Last 20 years has seen the rapid development of prokaryotic genomics. Since the sequencing of *Haemophilus influenzae* in 1985 (Johnston 2010; Fleischmann et al. 1995), currently over 11,364 whole genome sequences organized in three major groups of organisms i.e. eukaryota, prokaryota (archaea and bacteria) and viruses are available in Genome database of NCBI including complete chromosomes, organelles and plasmids as well as draft genome assemblies. Out of 11,364 whole genome sequences, 7,473 genome projects running across the world belong only to microbes with 1,696 completed microbial genomes projects whereas assembly is being done for 2,247 organisms and 3,531 genome project are still unfinished (Benson et al. 2002). The developing technology of nucleic acid sequencing, together with the recognition that sequences of building blocks in informational macromolecules (nucleic acids, proteins) can be used as ‘molecular clocks’ that contain historical information, led to the development of the three-domain model in the late 1970s, primarily based on small subunit ribosomal RNA sequence comparisons. The information currently accumulating from complete genome sequences of an ever increasing number of prokaryotes are now leading to further modifications of our views on microbial phylogeny. Prokaryotic genomics has had a revolutionary impact on our view of the microbial world and also on the methodologies for microbiological studies.

2 Complexity of Microbial Genomes

Analysis of genomic sequences has revealed that microbial genomes are very diverse. This is due to the complicated nature of microbial evolution. Mutations play a key role in evolution of eukaryotic genomes whereas, the contents of prokaryotic genomes are also changed by gene losses, gene rearrangements, horizontal gene transfer, and so on (McHardy et al. 2007; Doolittle 1999; Woese 1987). This means that even strains from the same species can differ significantly. For example, two *Escherichia coli* strains O157:H7 and K-12 have more than 1,000 different genes (Perna et al. 2001). The dynamic nature of microbial genomes complicates several tasks in microbiological studies. One of these is the development of strategies to prevent and treat microbe related diseases. Since microbe related diseases are common threats to the public health, microbes especially bacteria have been studied for many years. One point of progress was the introduction of antibiotics to treat bacterial infections. However, the use of antibiotics has been challenged by the emergence of antibiotic resistance among bacteria.

Plasmids play an important role in conferring antibiotic resistance in microbes. It is believed that antibiotic resistance evolves via natural selection. However, antibiotic resistance can also be introduced to bacteria via horizontal gene transfer

(Boerlin and Reid-Smith 2008). Plasmids are extra-chromosomal genetic elements that constitute upto 10% of the total DNA found in many species of bacteria (Mølbak et al. 2003; Thomas 2000). Because plasmids are capable of cell-to-cell transfer between bacterial species, genes harboured by plasmids are widely shared, playing a critical role in the evolution of bacteria (Feinbaum 2001; Summers 1996). Establishing accurate relationships between plasmids will help us to understand an important factor in the dissemination of antibiotic resistance genes, and establishing accurate relationships between bacteria will help us to identify the factors that cause diseases, the risks of outbreaks, and methods for preventing disease transmission. Unfortunately, the complexity of microbial genomes is apparent when we try to compare the genetic contents of strains and to build a phylogeny tree from them (McHardy et al. 2007; Doolittle 1999).

3 Obtaining Data (Wet Lab Approach)

One characteristic of microbiological studies in the genomics era is that we can generate a huge amount of data efficiently. Numerous different genomics based experimental methods are available. These methods are usually called molecular methods since they are often based on genetic characteristics. Compared to traditional phenotype-based methods, molecular methods are cost effective, easy to implement, and generate highly discriminatory data (Foxman et al. 2005; Tenover et al. 1997). Of these methods, the most widely used method for nucleic acid amplification is the polymerase chain reaction assay i.e., PCR. This assay includes a specific primer pair to amplify a unique genomic target nucleotide sequence for analysis. Following PCR, a variety of post-amplification methods are used to evaluate the product such as direct sequence analysis, use of genus or species specific probes, and utilization of restriction enzymatic analysis of the product, e.g., restriction fragment length polymorphism analysis (RFLP). Pulse-field gel electrophoresis (PFGE) is also considered as the gold standard. Multiple locus variable-number tandem repeat analysis (MLVA) assays are also a potentially powerful alternative or complementary tool. Another most powerful technique is DNA microarrays which provide a powerful high-throughput genomic method that has been widely used in biological studies. To construct a DNA microarray, single-strand fragments of DNA (also called probes) representing the genes of an organism are attached to a surface of glass or plastic. Each fragment can bind to a complementary DNA or RNA strand. Typically, more than 30,000 spots can be put on one slide, and it is possible to create a microarray representing every gene in a genome. Thus, microarrays can provide genome wide information which allows a comprehensive genetic analysis of an organism or a sample. DNA microarrays have been used for genotyping, expression analysis, and studies of protein-DNA interactions (Bilitewski 2009). When used for assessing the genetic relationships of bacterial strains, microarrays may be prepared for whole genome composed of open reading frames (ORFs) of one complete genome sequence (Zhou 2003). However, this type of microarray is limited by the requirement of representing one complete reference sequence which may not contain genetic content specific to nonsequenced strains. One possible improvement is to include specific genes from multiple whole-genome sequences or to use mixed-genome microarrays (MGMs) which use randomly-selected gene fragments from many strains of bacteria as probes (Wan et al. 2007; Borucki et al. 2004; Call et al. 2003).

From the enormous data to knowledge of microbial genomic information makes it possible to study microorganisms systematically. Sequence-based identification requires the recognition of a molecular target that is large enough to allow discrimination of a wide variety of microbes. One such target area that has been recognized is the rDNA gene complex which is present in all microbial pathogens. In bacteria, this complex is composed of a 16S rRNA gene and a 23S rRNA gene separated by a genomic segment called the internal transcribed spacer (ITS). Within fungi there are three genes (18S, 5.8S, and 28S) with spacers located between the genes (ITS1 and ITS2). Located in the rDNA gene complex are highly variable sequences that provide unique signatures for the identification of species and also conserved regions that contain genomic codes for the structural restrains that are present within organism groups. It has been shown that the ITS regions contain the most variability and that these regions are useful under most circumstances for species recognition. The availability of these variable sequence regions (ITS) surrounded by conserved sequences (16S/23S and 18S/5.8S/28S) allows for the utilization of an amplification system using universal (or consensus) bacterial or fungal primers. Once amplification has occurred using the consensus primers, the sequence is determined and comparison analysis of the unknown sequence to known sequences contained within a large database (such as the National Center for Biological Information (NCBI), GenBank databases) can be done to determine similarity and subsequently may lead to species identification. However, how to manipulate the massive amount of available data, how to retrieve genomic information effectively, and how to process the large scale data efficiently are all challenging problems. Because of these problems, the field of bioinformatics has emerged and has become an integral part of microbial studies (Foster et al. 2012).

4 Bioinformatics

Bioinformatics has evolved into a full-fledged multidisciplinary subject that integrates developments in information and computer technology as applied to biotechnology and biological sciences. Bioinformatics uses computer software tools for database creation, data management, data warehousing, data mining and global communication networking. In this, knowledge of many branches are required like biology, mathematics, computer science, laws of physics & chemistry, and sound knowledge of information technology to analyze the data. Bioinformatics is not limited to the computing data, but in reality it can be used to solve many biological problems and find out how living things work. It is the comprehensive application of mathematics (e.g., probability and statistics), science including biochemistry, molecular biology and a core set of problem-solving methods e.g. computer algorithms to the understanding of living systems. Bioinformatics is the recording, annotation, storage, analysis, and searching/retrieval of nucleic acid sequence genes, RNAs, protein sequences and structural information. This includes databases of the sequences and structural information as well methods to access, search, visualize and retrieve the information.

Functional genomics, biomolecular structure, proteome analysis, cell metabolism, biodiversity, drug and vaccine designs are some of the areas in which bioinformatics is an integral component. Bioinformatics concern the creation and maintenance of databases of biological information whereby researchers can both access existing information and submit new entries. The most pressing tasks in bioinformatics involve the analysis of sequence information. Computational Biology is the name given to this process.

5 Bioinformatics and Its Scope

Bioinformatics has evolved into a full-fledged scientific discipline over the last decade. The definition of Bioinformatics is not restricted to computational molecular biology and computational structural biology. It now encompasses fields such as comparative genomics, structural genomics, transcriptomics, proteomics, cellunomics and metabolic pathway engineering. Developments in these fields have direct implications to healthcare, medicine, discovery of next generation drugs, development of agricultural products, renewable energy, environmental protection etc.

Bioinformatics integrates the advances in the areas of computer science, information science and information technology to solve complex problems in life sciences. The core data comprises of the genomes and proteomes of human to microbes, 3-D structures and functions of proteins, microarray data, metabolic pathways, cell lines, hybridoma and biodiversity etc. The sudden growth in the quantitative data in biology has rendered data capture, data warehousing and data mining as major issues for biotechnologists and biologist. Availability of enormous data has resulted in the realization of the inherent biocomplexity issues which call for innovative tools for synthesis of knowledge. Information technology, particularly the internet, is utilized to collect, distribute and access ever-increasing data which are later analyzed with mathematics and statistics-based tools. Bioinformatics has a key role to play in the cutting edge research and development areas such as functional genomics, proteomics, protein engineering, pharmacogenomics, discovery of new drugs and vaccines, molecular diagnostic kits, agro-biotechnology etc. This has attracted attention of several companies and entrepreneurs. As a result, a large number of bioinformatics based start-ups have been launched and the trend is likely to continue. A Bioinformatician must acquire/possess expertise in the essential multidisplinary fields that comprise the core of this new science. Quality research and education in bioinformatics are vital not only to meet the existing challenges but also to set and accomplish new goals in life sciences.

6 The Potential of Bioinformatics

The potential of bioinformatics in the identification of useful genes leading to the development of new gene products, drug discovery and drug development has led to a paradigm shift in biology and biotechnology. These fields are becoming more and more computationally intensive. The new paradigm, now emerging, is that all the genes will be known "in the sense of being resident in database available electronically", and the starting point of biological investigation will be theoretical and a scientist will begin with a theoretical conjecture and only then turning to experiment to follow or test the hypothesis. With a much deep understanding of the biological processes at the molecular level, the bioinformatics scientist have developed new techniques to analyze genes on an industrial scale resulting in a new area of science known as 'Genomics'. This is the science that deals with the study of whole genome, largely encompasses biology of genetics at molecular level i.e., the constitution of DNA and RNA, its analysis, translation of the chemical information carried over by these materials into biological data and digitizing that huge biological data through computational means.

The shift from gene biology has resulted in the development of strategies from lab techniques to computer programmes to analyze whole batch of genes at once. Genomics is revolutionizing drug development, gene therapy, and our entire approach to health care and human medicine. The genomic discoveries are getting translated into practical biomedical results through bioinformatics applications. Work on proteomics and genomics will continue using highly sophisticated software tools and data networks that can carry multimedia databases. Thus, the research will be in the development of multimedia databases in various areas of life sciences and biotechnology. There will be an urgent need for development of software tools for data mining, analysis and modelling, and downstream processing. It has now been universally recognized that bioinformatics is the key to the new grand data-intensive molecular biology that will lead us in this century.

7 Activities in Bioinformatics

We can split the activities in bioinformatics in two areas:

1. **The organization**: this includes the creation of databases of biological information and the maintenance of the databases. This is very important as we are sequencing tens of millions of bases a year and undertaking to sequence whole organism genomes. The growth of the sequence databases is an unbroken exponential.
2. **Analysis of the data**: this includes the following:
   - Development of methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences.
   - Clustering protein sequences into families of related sequences and the development of protein models.

- Aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships
- The development of new algorithms and statistics with which to assess relationships among members of large data sets.
- The development and implementation of tools that enable efficient access and management of different types of information and
- The analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures in studying microbial diversity.

8 The Need for Bioinformatics

- Whole genome analysis and sequences
- Experimental analysis involving thousands of genes simultaneously
- DNA Chips and Array Analyses – expression arrays, Comparative analysis between species and strains
- Proteomics: 'Proteome' of an organism.
- Medical applications: Genetic Disease – Pharmaceutical and Biotech Industry
- Forensic applications
- Agricultural applications

9 Databases

Computational analysis and comparative microbial genomic studies are taking shape at a faster rate leading to the development of different types of function prediction concepts, most important of them being the gene context and gene content analysis. Gene content analysis is a comparison of gene repertoires across different genomes (Shah et al. 2005; Luscombe et al. 2001). The postgenomic problems like protein structural determination and issues of gene function identification become more promising (Gomez et al. 2008) with the rapidly increasing number of completely sequenced genomes. Predicting the structures of proteins encoded by genes of interest provides subtle clues regarding the functions of these proteins (Idekar et al. 2001).

Various databases have been established for storing genomic data, and the internet makes it possible for these data to be accessed and shared by the public. Since there are different types of genomic data, it is impossible to build one database containing all data. Currently there are two types of genomic databases. Primary databases contain sequences and structures (for example, NCBI GenBank) and related annotations, bibliographies, and cross-references to other databases and provide the basis for biological studies; secondary databases contain biological knowledge obtained by analyzing genomic sequences and structure data. The database of

Clusters of Orthologous Groups of proteins (COGs, http://www.ncbi.nlm.nih.gov/COG), for example, contains information for phylogenetic analysis (Tatusov et al. 1997, 2003). The Ribosomal Database Project (RDP) provides ribosome related data and annotated bacterial and archaeal small-subunit 16S rRNA sequences (Cole et al. 2005, 2009; Larsen et al. 1993). Knowledge from these databases can help to process biological data efficiently. For example, the Gene Ontology database has been used to process microarray datasets (Barrell et al. 2009; Harris et al. 2004). Nucleic acid sequence analysis has proven to be a valuable asset for organism identification in a number of applications. Some of the most interesting applications of this technology are for the identifications of variant strains of known species, the identification of un-cultivatable organisms in clinical samples and the recognition of new species.

10 Web-Based Resources for Microbial Genomics

MicrobesOnline: MicrobesOnline is a website for browsing and comparing prokaryotic genomes. MicrobesOnline is a product of the Virtual Institute for Microbial Stress and Survival, which is sponsored by the US Department of Energy Genomic Science Program.

Integrated Microbial Genomes (IMG): The Integrated Microbial Genomes (IMG) system serves as a community resource for comparative analysis and annotation of all publicly available genomes from three domains of life, in a uniquely integrated context.

CAMERA (Community Cyber infrastructure for Advanced Marine Microbial Ecology Research and Analysis): The aim of CAMERA is to serve the needs of the microbial ecology research community by creating a rich, distinctive data repository and a bioinformatics tools resource that will address many of the unique challenges of metagenomic analysis.

DOE JGI Microbial Genomics Database: From this site we can get details about JGI projects, or go directly to the individual microbial sites. All of the individual sites include direct access to download sequence file(s), BLAST, and view annotations.

GOLD™ Genomes OnLine Database: GOLD is a World Wide Web resource for comprehensive access to information regarding complete and ongoing genome projects, as well as metagenomes and metadata, around the world.

JCVI Comprehensive Microbial Resource (formerly The Institute for Genomic Research): The Comprehensive Microbial Resource (CMR) is a free website used to display information on all of the publicly available, complete prokaryotic genomes.

Sanger Centre Bacterial Genomes: The Sanger Institute bacterial sequencing effort is concentrated on pathogens and model organisms. The site provides a list of projects funded, underway or completed; all data from these projects are immediately and freely available.

Microbial Genomes from Genome Channel: Genome Channel is a computer-annotated listing of genomes maintained by the Computational Biology group at Oak Ridge National Laboratory.

Protein Data Bank: The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

KEGG (Kyoto Encyclopaedia of Genes and Genomes): A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviours from genomic and molecular information. Towards this end a bioinformatics resource named KEGG has been developed as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Centre of Kyoto University and the Human Genome Centre of the University of Tokyo.

11 Data Retrieval Methods and Online Resources for Microbial Diversity

In order to use the information available in databases, an efficient information retrieval method should be used to obtain all related information quickly. Such methods are different, depending on the type of data to be retrieved. FASTA and BLAST are the two most widely used methods for retrieving sequence data. FASTA was the first fast sequence searching algorithm used for comparing a query sequence against a database (Plewniak 2008; Pearson 1990). The FASTA algorithm performs a rapid and approximate search for matched sequence segments followed by application of the Smith-Waterman alignment algorithm (Plewniak 2008; Pearson 1991) to these segments. Depending upon the application there are several softwares available online for free to retrieve the microbial data. Some of them are briefly described below:

11.1 Pairwise Alignment

A number of computational methods have been developed and used in genomic studies. Of these methods, genetic sequence alignment is the foundation for many other methods and widely used in comparative genomics. A good alignment method should give biologically meaningful results and at the same time be computationally efficient. There are two types of alignment methods, local alignments and global alignments. The former methods try to identify similar segments between two sequences while the latter try to align the entire length of two sequences. Methods for aligning two sequences are called pairwise alignment methods. BLAST and

FASTA are two widely used pairwise alignment methods. BLAST (Basic Local Alignment Search Tool) is a rapid sequence database search tool which is more efficient than FASTA. The output of BLAST is a list of high-scoring segment pairs (HSPs) and an “E value” which is an estimate of the probability of finding an HSP with score S. The E value is often used as a standardized measure for estimating the statistical significance of sequence similarity.

These methods can be extended to multiple sequences; however, multiple sequence alignment (MSA) is more complicated. ClustalW (Larkin et al. 2007; Thompson et al. 2002) is a widely used MSA method which is efficient for aligning protein sequences and short nucleotide sequences. However, it may fail for distantly related sequences (Lin et al. 2011). PSI-BLAST (Lee et al. 2008; Schäffer et al. 2001; Altschul et al. 1997) is a very successful method for detecting weak similarities. Two recently developed algorithms, MLAGAN (Brudno et al. 2003) and MAVID (Dewey 2007; Bray and Pachter 2003, 2004), are designed for global alignment of both evolutionarily close and distant megabase length genomic sequences. However, a phylogenetic tree is assumed to be known for use with MLAGAN. MAVID is a progressive global alignment program that works by recursively aligning the ‘alignments’ at ancestral nodes of the guide phylogenetic tree. MAUVE is used for comparing long genome sequences efficiently and takes into account possible large-scale evolutionary events among sequences (Darling et al. 2004).

11.2 Phylogenetic Analysis

The goal of phylogenetic analysis is to reconstruct the evolutionary history of a set of organisms. In molecular epidemiology, it helps to elucidate mechanisms that lead to microbial outbreaks and epidemics. Phylogenetic analysis usually begins with multiple sequence alignment of the sequences of a set of organisms. After obtaining an MSA, a number of different phylogenetic methods can be used to compute phylogenetic trees. These methods can be broadly classified into maximum parsimony, distance, and maximum likelihood methods (Stark et al. 2010; Takahashi and Nei 2000). The difference between these methods is how they define which tree is best among all possible trees. Maximum parsimony tries to find an evolutionary tree or trees which require a minimum number of changes from the common ancestral sequences. For maximum likelihood methods, given the MSA, the probability of a specific tree occurring is computed, and the one or ones with the highest values are considered to be the evolutionary tree or trees. Distance-based methods construct a tree by hierarchical clustering methods using a distance matrix for all organisms that is computed using MSA. To use MSA for phylogenetic analysis, it is necessary to assume an underlying mutation model. Of the ones that have been proposed, the Jukes-Cantor (JC) model (Som 2006; Takahashi and Nei 2000) is the simplest one. In the JC model, each base in a DNA sequence has an equal mutation rate and all complementary pairs of the four nucleotides A, T, C and G have equal substitution rates. These assumptions are not realistic in practice, so many complex models have been proposed and tried. Successful phylogenetic analysis requires a suitable model. Phylogenetic analysis of microbial strains is problematic due to its dynamic nature (Wilmes et al. 2009). Different genes among strains may contain contradictory information about their evolution. Consensus trees have been suggested as a solution. An alternative is the introduction of networks that represent the evolutionary relationships between microbial strains.

11.3 AGeS: A Software System for Microbial Genome Sequence Annotation

AGeS is genome sequence annotation software which is a fully integrated with high performance software system to analyze DNA sequences and predict the protein-coding regions for completed and draft bacterial genomes. It predicts genomic features using a number of bioinformatics methods and provides visualization based on the familiar genome browser.

11.4 SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data

Ribosomal RNA sequence data analyzing tool SILVA is available online at http://www.arb-silva.de/. Sequencing ribosomal RNA (rRNA) genes is currently the method of choice for phylogenetic reconstruction, nucleic acid based detection and quantification of microbial diversity (Pruesse et al. 2007). To cope with the flood of data, the SILVA system was implemented to provide a central, comprehensive web resource for up to date, quality controlled databases of aligned small and large subunit rRNA sequences from the bacteria and archaea domains. This programme is designed as a central comprehensive resource by integrating multiple taxonomic classifications and the latest validly described nomenclature as well as additional information, such as if a sequence was derived from a cultivated organism, a type strain, or belongs to a genome project.

11.5 16S and 23S Ribosomal RNA Mutation Database

Access to the expanded versions of the 16S and 23S Ribosomal RNA Mutation Databases has been improved to permit searches of the lists of alterations for all the data from (1) one specific organism, (2) one specific nucleotide position, (3) one specific phenotype. The URL for the searchable version of the Databases is: http://ribosome.fandm.edu.

11.6 5S Ribosomal RNA Database

5S Ribosomal RNA Database provides information on nucleotide sequences of 5S rRNAs and their genes. The sequences for particular organisms can be retrieved as single files using a taxonomic browser or in multiple sequence structural alignments. This programme is freely available at http://biobases.ibch.poznan.pl/5SData/.

11.7 Greengenes

This is an online full-length small-subunit (SSU) rRNA gene database called greengenes available at <http://greengenes.lbl.gov/> that keeps pace with public submissions of both archaeal and bacterial 16S rDNA sequences has been established (DeSantis et al. 2003). It addresses a number of limitations currently associated with SSU rRNA records in the public databases by providing automated chimera-screening, taxonomic placement of unclassified environmental sequences using multiple published taxonomies for each record, multiple standard alignments and uniform sequence-associated information curated from GenBank records. Greengenes also provides a suite of utensils for manipulation of sequences including an alignment tool and has been streamlined to interface with the widely used ARB program.

11.8 Ribosomal Database Project

The Ribosomal Database Project – II (RDP-II) (Maidak et al. 2001) available at <http://rdp.cme.msu.edu/> provides data, tools and services related to ribosomal RNA sequences to the research community. It offers aligned and annotated rRNA sequence data, analysis services, and phylogenetic inferences derived from these data. Currently available on the RDP-II website as a beta release, 9.0 provides over 50,000 annotated (eu) bacterial sequences aligned with a secondary-structure based alignment algorithm (Brown 2000). Data subsets are available for sequences of length 1,200 or greater and for sequences from type material. Annotation goals include up-to-date name, strain and culture deposit information, sequence length and quality information. In order to provide a phylogenetic context for the data, RDP-II makes available over 100 trees that span the phylogenetic breadth of life. Web based research tools are provided for comparing user submitted sequences to the RDP-II database (Sequence Match), aligning user sequences against the nearest RDP sequence (Sequence Aligner), examining probe and primer specificity (Probe Match), testing for chimeric sequences (Chimera Check), generating a distance matrix (Similarity Matrix), analyzing T-RFLP data (T-RFLP and TAP-TRFLP), a java-based phylogenetic tree browser (Sub Trees), a sequence search and selection tool (Hierarchy Browser) and a phylogenetic tree building and visualization tool (Phylip Interface). The latter tool has been enhanced to allow a choice of either the

Phylip neighbor-joining (Felsenstein 1993) or Weighbor weighted neighbor-joining (Bruno et al. 2000) programs for tree construction.

11.9 RISSC – Ribosomal Internal Spacer Sequence Collection

This is a database of ribosomal 16S-23S spacer sequences intended mainly for molecular biology studies in typing, phylogeny and population genetics. It compiles more than 2,500 entries of edited DNA sequence data from the 16S-23S ribosomal spacers present in most prokaryotes and organelles. Ribosomal spacers have proven to be extremely useful tools for typing and identifying closely related prokaryotes due to their high variability in size and/or sequence, much more so than the flanking 16S and 23S rRNA genes. These genes are commonly used to establish molecular relationships among microbes at a taxonomic level of species or higher (e.g. genus, domain). However their internal transcribed spacers (ITS) are much more useful to discriminate at the species or even strain level (Iwen et al. 2002). RISSC available at <http://ulises.umh.es/RISSC> provides the scientific community with a comprehensive set of ribosomal spacer sequences, fully edited and characterized with a key feature as is the presence/absence of tRNA genes within them, ready to be used and compared with their own ITS sequences.

11.10 probeBase

probeBase is a curated database of annotated rRNA-targeted oligonucleotide probes and supporting information (Loy et al. 2003, 2007). Rapid access to probe, microarray and reference data is achieved by powerful search tools and via different lists that are based on selected categories such as functional or taxonomic properties of the target organism(s), or the hybridization system in which the probes were applied. Additional information on probe coverage and specificity is available through direct submissions of probe sequences from probeBase to RDP-II and Greengenes, two major rRNA sequence databases.

ProbeBase available at <http://www.microbial-ecology.net/probebase> entries increased from 700 to more than 1,200 during the past 3 years. Several options for submission of single probes or entire probe sets, even prior to publication of newly developed probes, should further contribute to keeping probeBase an up-to-date and useful resource.

11.11 RRNDB

The Ribosomal RNA Operon Copy Number Database (RRNDB) available at <http://rrndb.cme.msu.edu/> contains annotated information on rRNA operon copy number among prokaryotes. Gene redundancy is uncommon in prokaryotic genomes, however rRNA genes can vary from one to as many as 15 copies. Despite the widespread use of 16S rRNA gene sequences for identification of prokaryotes, information on the number and sequence of individual rRNA genes on a genome is not readily accessible. Each entry in RRNDB contains detailed information linked directly to external websites including the Ribosomal Database Project, GenBank, PubMed, and several culture collections.

12 Identification of New Species or Variant Strains of Known Species

Bioinformatics has facilitated researchers to study microbial biodiversity because of its direct interventions in molecular identification, data storage and retrieval system that were the objects and the worrisome of systematic research. The bioinformatics driven approaches enabled people to work efficiently on microbial diversity, identification, characterization, molecular taxonomy and community analysis patterns of both culturable and unculturable organisms. Description of new species, genera and even molecular taxa emerged dramatically in the literature after 1990s and these efforts are largely driven by advances in sequencing technologies. The utilization of phenotypic identification methods classically requires a probability-based analysis to determine identity. In cases where identification probabilities are 98% with known species, the identification is generally considered acceptable. The lower the probability percentage however, the less accurate the identification becomes, frequently resulting in supplemental testing to resolve discrepancies among test results. It is not unusual for the laboratory to be unable to identify variant strains of known species using phenotypic methods. DNA sequencing now allows the laboratory a means to resolve those instances where phenotypic testing cannot differentiate among closely related organisms.

The recognition of a species that does not match known schemes for phenotypic identification may represent a previously unrecognized species (Relman 2002). Sequencing of areas within the rDNA complex may be useful to suggest a new species when there is a <98% of the sequence similarity with known species. The ability to separate a new species from an atypical strain of a known species is however, difficult. The first approach to recognition of a new species is to determine the phylogenetic position of the suspect new species compared to closely related known species. Phylogenetic trees using the 16S gene for bacteria and the 18S gene for fungi are commonly used for this type of analysis. The 16S rRNA approach is rooted in the concept of point mutation due to their slow mutation rate. Before microbial genomes were sequenced, using 16SrRNA database was considered and bacteria, archaea, and eukaryotes were identified.

A high degree of phenotypic consistency and rDNA sequence similarity as well as, a significant degree of DNA-DNA hybridization, is suggestive of a new species.

13 Bioinformatics Challenges

Many bioinformatics tools have been borrowed from the fields of artificial intelligence, data mining, and statistical methods. However, the characteristics of biological data may differ significantly from those of the original data for which the methods were developed. Though many computational methods have been introduced for genomic data analysis based on these methods, several challenges still exist. Though public databases such as GenBank are useful, the lack of quality sequences and the absence of sequence information on a large number of species as well as the availability of computational tools to reliably analyze the results are drawbacks to this technology. A typical DNA microarray might have thousands of features (probes) for, at most, one hundred samples. Feature reduction is typically required before these sorts of analyses can be performed (Al-Khaldi et al. 2012; Bier et al. 2008; Yauk and Berndt 2007). Another challenge is integrating data from different sources. These datasets might show a high degree of heterogeneity and might also vary in quality. They might be generated using different experimental platforms or based on different molecular methods. Using these data together efficiently requires developing suitable bioinformatics methods. Of these methods, the simplest one is to put several datasets together to build a larger dataset and then analyze this larger dataset. However, this method will not work if the formats of the original datasets differ. Furthermore, the best processing methods for different datasets are not the same. For example, Dice coefficents work well for some PFGE data but does not work well for some VNTR data. Thus, it might be an impossible task to choose an optimal method for a combined dataset. An alternate method is to process different datasets separately and then combine the results to obtain the final result. The difficulties with this kind of method, however, are determining the extent to which the different sources of data should contribute and explaining the combined results.

14 Conclusion

The development of computational methods based on the organized algorithms, interpretational skills and high storage capacities facilitated comparison of entire genomes and thus permit biologists to study more complex evolutionary trends like gene duplication, horizontal gene transfer and prediction of factors important in speciation (Nakashima et al. 2005). Bioinformatics researchers have compared extensively multiple genomes to correlate and classify the genomes into various families and to study evolution. It has been established by many researchers that overall evolution is a combination of point based mutation giving rise to restructuring of genomes based upon gene duplications, gene insertion, gene deletion, horizontal gene transfer etc. The ultimate aim of such studies lies in deciphering the evolutionary lineages among the group of organisms in a quest to determine the tree of life and the last universal common ancestor. The progress in bioinformatics and wet-lab techniques has to remain interdependent and focused complementing each other for their own progress and for the progress of biotechnology in future.

15 Some More Web Addresses for Bioinformatics Tools

Name of tool/database	Web address
ASD	http://www.ebi.ac.uk/asd
AUGUSTUS	http://augustus.gobics.de/bin/npsa_automat.pl?page=npsa_gor4.html
BLAST	http://www.ncbi.nlm.nih.gov/blast
CFSSP	http://www.biogem.org/tool/chou-fasman/
Clustal W	http://www.ebi.ac.uk/Tools/clustalw2/index.html
ComputpI/Mw	http://web.expasy.org/compute_pi/
CpG Island Searcher	http://www.uscnorris.com/cpgislands2/cpg.aspx
CpGPlot	http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html
DDBJ BLAST	http://blast.ddbj.nig.ac.jp
DNA tools	http://biology.semo.edu/cgi-bin/dnatools.pl
Entrez Gene	http://www.ncbi.nlm.nih.gov/sites/entrez
ESLPred2	http://www.imtech.res.in/raghava/eslpred2/
ExPaSy	http://expasy.org/tools/
FEX	http://www.softberry.ru/berry.phtml
FGENESH	http://www.softberry.ru/berry.phtml
GeneMark.hmm	http://www.itba.mi.cnr.it/webgene/
GOR	http://npsa-pbil.ibcp.fr/cgi-
HMMgene	http://www.cbs.dtu.dk/services/HMMgene/
MGI	http://www.informatics.jax.org/
MultiLoc2	http://abi.inf.uni-tuebingen.de/Services/MultiLoc2
Myristoylator	http://web.expasy.org/myristoylator/
NetAcet	http://www.cbs.dtu.dk/services/NetAcet/
NetOGlyc	http://www.cbs.dtu.dk/services/NetOGlyc
NetPhos	http://www.cbs.dtu.dk/services/NetPhos/
NetPhosK	http://www.cbs.dtu.dk/services/NetPhosK/
NetSurfP	http://www.cbs.dtu.dk/services/NetSurfP/
NMT	http://mendel.imp.ac.at/myristate/SUPLpredictor.htm
OligoCalc	http://www.basic.northwestern.edu/biotools/oligocalc.html
PSIPRED v3.0	http://bioinf.cs.ucl.ac.uk/psipred/
SherLoc2	http://abi.inf.uni-tuebingen.de/Services/SherLoc2
SIGSCAN	http://www-bimas.cit.nih.gov/molbio/signal/
SMS	http://www.bioinformatics.org/sms/
TermiNator	http://www.isv.cnrs-gif.fr/terminator3/index.html
TFBIND	http://tfbind.hgc.jp/
TFSEARCH	http://www.cbrc.jp/research/db/TFSEARCH.html

Glossary

Homology Searches: BLAST & FASTA Background Information: The three BLAST programs that one will commonly use are BLASTN, BLASTP and BLASTX. BLASTN will compare your DNA sequence with all the DNA sequences in the nonredundant database (nr). BLASTP will compare your protein sequence with all the protein sequences in nr. In BLASTX nucleotide sequence of interest will be translated in all six reading frames and the products compared with the nr protein database. A tutorial is also available at NCBI.

BLAST Homepage – (NCBI) Found at http://blast.ncbi.nlm.nih.gov/Blast.cgi. It is widely used for homology searches. BLAST stands for Basic Local Alignment Search Tool and it displays a number of organism and query specific blast.

Nucleotide BLAST (BLASTN) N.B. the default database is "human genomic and transcript" not "nucleotide collection (nt/nr)"

Protein BLAST (BLASTP) This program is also coupled with a motif search. If you suspect that your protein may only show weak sequence similarity to other proteins, I would suggest clicking on the PSI-BLAST (Position-Specific Iterated BLAST) feature. NCBI also provides a PSI-BLAST tutorial. PSI-BLAST searches to yield better delineation of true and false positives.

Translated BLAST (BLASTX) TBLASTX searches translated nucleotide databases using a translated nucleotide query; while TBLASTN searches translated nucleotide databases using a protein query. These are useful resources if you are interested in homologs in unfinished genomes. Undeter "Databases" select "genomic survey sequences", "High throughput genomic sequences" or "whole-genome shotgun reads"

Blast with Microbial Genomes (BLASTN, TBLASTN, TBLASTX etc.). It permits us to compare a nucleic acid or protein sequence against finished archaeal and bacterial genomes. Depending upon the time of day your results may appear almost immediately or the search may be delayed or not accepted at all. For PSI-BLAST, and other searches it is recommended to frequently enter information in the "Entrez Query" section e.g. *Escherichia coli* [organism] or Viruses [organism] to see "hits" specifically to E. coli or viruses/bacteriophages. It is advisable to always select "Show results in a new window"

EMB BLAST-(European Molecular Biology network). Very convenient since it permits one to specifically search databases such as prokaryote, bacteriophage, fungal, & 16S rRNA using BLASTN, and specific bacterial genomes or SwissProt using BLASTX or BLASTN.

ParAlign It employs a heuristic method for sequence alignment. In essence, ParAlign is about as sensitive as Smith-Waterman but runs at the speed of BLAST.

GTOP Sequence Homology Search (Laboratory for Gene-Product Informatics, National Institute of Genetics, Japan) – offers BLASTP search capability against individual archaea, bacteria, eukaryota, and viruses.

T4-like Phage NCBI MegaBLAST (Tulane Univ., New Orleans, U.S.A. & CNRS, Toulouse, France) This includes a growing list of T4-like completed phage sequences as well as those in the draft and contig stages of completion.

WU-BLAST (Washington University BLAST) The emphasis of this tool is to find regions of sequence similarity quickly, with minimum loss of sensitivity. This will yield functional and evolutionary clues about the structure and function of the novel sequence.

Batch BLAST (Greengene web server; University of Massachusetts, Lowell, U.S.A.) was developed by Michael V. Graves for DNA or protein BLAST sequence analysis against the NCBI databases. It allows one to submit a file that contains multiple sequences and then will organize the results by each individual sequence contained in the file.

HHPred Homology detection & structure prediction is a method for database searching and structure prediction that is as easy to use as BLAST but is much more sensitive in finding remote homologs. HHpred is the first server that is based on the pairwise comparison of profile hidden Markov models (HMMs). Whereas most conventional sequence search methods search sequence databases such as UniProt or the NR, HHpred searches alignment databases, like Pfam or SMART. This greatly simplifies the list of hits to a number of sequence families instead of a clutter of single sequences. HHpred accepts a single query sequence or a multiple alignment as input.

PSI-BLAST or PHI-BLAST search Position-Specific Iterative BLAST creates a profile after the initial search.

BLAST 2 BLAST two sequences against one another. This utilizes BLASTN, P, X as well as TBLASTN and TBLASTX.

Gene Context Tool It is an incredible tool for visualizing the genome context of a gene or group of genes.

TC-BLAST It scans the transport protein database (TC-DB) producing alignments and phylogenetic trees. The TC-DB details a comprehensive classification system for membrane transport proteins known as the Transport Commission (TC) system.

MEROPS BLAST This permits one to screen protein sequences against an extensive database of characterized peptidases.

SEARCHGTr It is web-based software for the analysis of glycosyltransferases involved in the biosynthesis of a variety of pharmaceutically important compounds like adriamycin, erythromycin, vancomycin etc. This software has been developed based on a comprehensive analysis of sequence/structural features of 102 GTrs of known specificity from 52 natural product biosynthetic gene clusters.

PipeAlign It offers an integrated approach to protein family analysis through a cascade of different sequence analysis programs (BALLAST, DbClustal multiple alignment program, Rascal alignment analysis) removal of any sequences that do not belong to the protein family are performed by the NorMD, and clustered into potential functional subfamilies using Secator or DPC.

MPsrch (EMBL-EBI) This sequence sequence comparison tool implements the true Smith and Waterman algorithm identifying hits in cases where Blast and Fasta fail and also reports fewer false-positives. This software provides information on Match %; % Query Match (% of the query sequence matched); Conservative changes; Mismatches; Indels; and Gaps.

GOAnno This web tool automatically annotates proteins according to the Gene Ontology using hierarchised multiple alignments. Positioning the query protein in its aligned functional subfamily represents a key step to obtain highly reliable predicted GO annotation based on the GOAnno algorithm.

COMPASS This is a profile-based method for the detection of remote sequence similarity and the prediction of protein structure. The server features three major developments: (i) improved statistical accuracy; (ii) increased speed from parallel implementation; and (iii) new functional features facilitating structure prediction. These features include visualization tools that allow the user to quickly and effectively analyze specific local structural region predictions suggested by COMPASS alignments.

MineBlast It performs BLASTP searches in UniProt to identify names and synonyms based on homologous proteins and subsequently queries PubMed, using combined search terms in order to find and present relevant literature.

Comparison of homology between two small genomes: SCAN2 (Softberry.com) It provides one with a colour-coded graphical alignment of genome length DNAs in Java. In the top panel regions of high sequence identity are presented in red. By highlighting the grey yellow, green, black boxes one can select specific regions for examination of the sequence alignment.

Advanced PipMaker It aligns two DNA sequences and returns a percent identity plot of that alignment, together with a traditional textual form of the alignment. We may need to download it for viewing and manipulating the output from pairwise alignment programs such as PipMaker representations of the alignments.

JDotter: A Java Dot Plot Viewer (Viral Bioinformatics Resource Center, University of Victoria, Canada) – a dot matrix plotter for Java. It produces similar diagrams to the above mentioned programs, but with better control on output.

multi-zPicture: multiple sequence alignment tool provides nice dotplot graphs and dynamic visualizations. If simple gene locations are provided in the form (e.g. >2,000–5,000 RNA_polymerase; indicates that the RNA polymerase gene is found on the plus strand between bases 2,000 and 5,000) this data will be added to the dynamic visualization. zPicture alignments can be automatically submitted to rVista to identify conserved transcription factor binding sites.

GeneOrder 3.0 This is ideal for comparing small GenBank genomes (up to 2 Mb). Each gene from the Query sequence is compared to all of the genes from the Reference sequence using BLASTP. There are two display formats: graphical and tabular.

CoreGenes This programme is designed to analyze two to five genomes simultaneously, generating a table of related genes – orthologs and putative orthologs. These entries are linked to their GenBank data. It has a limit of 0.35 Mb, while the newer version CoreGenes 2.0 extends the limit to approx. 2.0 Mb. If data is not present in GenBank, using this site will be very helpful.

CoreGenes 3.0 This is the latest member in the CoreGenes family of tools. It determines unique genes contained in a pair of proteomes. This has proved extremely useful in determining unique genes in comparisons between large Myoviridae.

References

Al-Khaldi SF, Mossoba MM, Allard MM, Lienau EK, Brown ED (2012) Bacterial identification and subtyping using DNA microarray and DNA sequencing. Methods Mol Biol 881:73–95

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R (2009) The GOA database in 2009–an integrated gene ontology annotation resource. Nucleic Acids Res 37:D396–D403

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2002) GenBank. Nucleic Acids Res 28:15–18

Bier FF, von Nickisch-Rosenegk M, Ehrentreich-Förster E, Reiss E, Henkel J, Strehlow R, Andresen D (2008) DNA microarrays. Adv Biochem Eng Biotechnol 109:433–453

Bilitewski U (2009) DNA microarrays: an introduction to the technology. Methods Mol Biol 509:1–14

Boerlin P, Reid-Smith RJ (2008) Antimicrobial resistance: its emergence and transmission. Anim Health Res Rev 2:115–126

Borucki MK, Kim SH, Call DR, Smole SC, Pagotto F (2004) Selective discrimination of Listeria monocytogenes epidemic strains by a mixed-genome DNA microarray compared to discrimination by pulsed-field gel electrophoresis, ribotyping, and multilocus sequence typing. J Clin Microbiol 42:5270–5276

Bray N, Pachter L (2003) MAVID multiple alignment server. Nucleic Acids Res 31:3525–3526

Bray N, Pachter L (2004) MAVID: constrained ancestral alignment of multiple sequences. Genome Res 14:693–699

Brown MP (2000) Small subunit ribosomal RNA modeling using stochastic context free grammar. Proc Int Conf Intell Syst Mol Biol 8:57–66

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S, NISC Comparative Sequencing Program (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731

Bruno WJ, Socci ND, Halpern AL (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol Biol Evol 17:189–197

Call DR, Borucki MK, Besser TE (2003) Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of Listeria monocytogenes. J Clin Microbiol 41:632–639

Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM (2005) The ribosomal database project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res 33(Database issue)

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM (2009) The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141–D145

Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403

DeSantis TZ, Dubosarskiy I, Murray SR, Andersen GL (2003) Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. Bioinformatics 19:1461–1468

Dewey CN (2007) Aligning multiple whole genomes with Mercator and MAVID. Methods Mol Biol 395:221–236

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2128

Feinbaum R (2001) Introduction to plasmid biology. Curr Protoc Mol Biol Chapter 1:Unit 1.5

Felsenstein J (1993) PHYLIP (phylogeny inference package) version 3.5c. Department of Genetics, University of Washington, Seattle

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Foster JA, Moore JH, Gilbert JA, Bunge J (2012) Microbiome studies: analytical tools and techniques. Pac Symp Biocomput 17:200–202

Foxman B, Zhang L, Koopman JS, Manning SD, Marrs CF (2005) Choosing an appropriate bacterial typing technique for epidemiologic studies. Epidemiol Perspect Innov 25:2–10

Gomez SM, Choi K, Wu Y (2008) Prediction of protein-protein interaction networks. Curr Protoc Bioinform Chapter 8:Unit 8.2

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium (2004) The gene ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261

Idekar T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–372

Iwen PC, Hinrichs SH, Rupp ME (2002) Utilization of the internal transcribed spacer regions as molecular targets to detect and identify human fungal pathogens. Med Mycol 40:87–109

Johnston JW (2010). Laboratory growth and maintenance of *Haemophilus influenzae*. Curr Protoc Microbiol Chapter 6:Unit 6D.1

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

Larsen N, Olsen GJ, Maidak BL, McCaughey MJ, Overbeek R, Macke TJ, Marsh TL, Woese CR (1993) The ribosomal database project. Nucleic Acids Res 21:3021–3023

Lee MM, Chan MK, Bundschuh R (2008) Simple is beautiful: a straightforward approach to improve the delineation of true and false positives in PSI-BLAST searches. Bioinformatics 24:1339–1343

Lin HN, Notredame C, Chang JM, Sung TY, Hsu WL (2011) Improving the alignment quality of consistency based aligners with an evaluation function using synonymous protein words. PLoS One 6:e27872

Loy A, Horn M, Wagner M (2003) probeBase: an online resource for rRNA-targeted oligonucleotide probes. Nucleic Acids Res 31:514–516

Loy A, Maixner F, Wagner M, Horn M (2007) probeBase – an online resource for rRNA-targeted oligonucleotide probes: new features 2007. Nucleic Acids Res 35(Database issue):D800–D804

Luscombe NM, Greenbaum D, Gerstein M (2001) What is bioinformatics? A proposed definition and overview of the field. Methods Inform Med 40:346–358

Maidak BL, Cole JR, Lilburn TG, Parker CT, Saxman PR, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM, Tiedje JM (2001) The RDP-II (Ribosomal Database Project). Nucleic Acids Res 29:173–174

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 4:63–72

Mølbak L, Tett A, Ussery DW, Wall K, Turner S, Bailey M, Field D (2003) The plasmid genome database. Microbiology 149:3043–3045

Nakashima N, Mitani Y, Tamura T (2005) Actinomycetes as host cells for production of recombinant proteins. Microb Cell Fact 4:7

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98

Pearson WR (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 11:635–650

Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529–533

Plewniak F (2008) Database similarity searches. Methods Mol Biol 484:361–378

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188–7196

Relman DA (2002) New technologies, human-microbe interactions, and the search for previously unrecognized pathogens. J Infect Dis 186(Suppl 2):254–258

Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005

Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF (2005) Atlas – a data warehouse for integrative bioinformatics. BMC Bioinformatics 21:6–34

Som A (2006) Theoretical foundation to estimate the relative efficiencies of the Jukes-Cantor + gamma model and the Jukes-Cantor model in obtaining the correct phylogenetic tree. Gene 385:103–110

Stark M, Berger SA, Stamatakis A, von Mering C (2010) MLTreeMap – accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. BMC Genomics 5:11–461

Summers DK (1996) The biology of plasmids. Blackwell Science, Oxford

Takahashi K, Nei M (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol 17:1251–1258

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41

Tenover FC, Arbeit RD, Goering RV (1997) How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists. Infect Control Hosp Epidemiol 18:426–439

Thomas CM (2000) The horizontal gene pool: bacterial plasmids and gene spread. Harwood Academic, Amsterdam

Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinform. Chapter 2:Unit 2.3

Wan Y, Broschat SL, Call DR (2007) Validation of mixed-genome microarrays as a method for genetic discrimination. Appl Environ Microbiol 73:1425–1432

Wilmes P, Simmons SL, Denef VJ, Banfield JF (2009) The dynamic genetic repertoire of microbial communities. FEMS Microbiol Rev 33:109–132

Woese CR (1987) Bacterial evolution. Microbiol Mol Biol Rev 51:221–271

Yauk CL, Berndt ML (2007) Review of the literature examining the correlation among DNA microarray technologies. Environ Mol Mutagen 48:380–394

Zhou J (2003) Microarrays for bacterial detection and microbial community analysis. Curr Opin Microbiol 6:288–294