Genome Sequence Databases: Types of Data and Bioinformatic Tools

A G-Preciado, M Peimbert, and E Merino, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

Glossary

curation Curation is the process of examining, testing and selecting information before its incorporation into a database.

exhaustive algorithm Exhaustive algorithm iteratively produces the entire solution space for a given problem, checks to see if the problem is solved, and continues until a correct solution is generated, at which point the optimal solution is returned.

heuristic algorithm Heuristic algorithm is an algorithm that makes educated guesses to solve a problem. This results in a faster solution without necessarily an optimal solution.

HMM A Markov model is a statistical model in which the probability of an event depends on the immediately previous event. In a HMM (for hidden Markov model) the parameters of the process are 'hidden' and the challenge is to determine them from the observable data.

OTU OTU stands for Operational Taxonomic Unit.

protein domain Protein domain is a protein evolution unit. It can be the entire protein or a compact part of protein structure.

sequence motif Sequence motif is a characteristic nucleotide or amino acid sequence that is conserved in a group of sequences. In most cases, it has a biological function, such as the catalytic site of a protein, or a DNA-binding site.

Abbreviations

BBH	bidirectional best hit
BLAST	Basic Local Alignment and Search Tool
CAMERA	Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis
CDD	conserved domain database
CGH	comparative genomic hybridization
CIBEX	Centre for Information Biology Gene Expression
CMR	Comprehensive Microbial Resource
COG	Cluster of Orthologous Groups of Proteins
DBTBS	Database of Transcriptional regulation in *B. subtilis*
DDBJ	DNA Data Bank of Japan
EMBL-EBI	European Molecular Biology Laboratory-European Bioinformatics Institute
GEO	Gene Expression Omnibus
GO	Gene Ontology

GOLD	Genomes Online Database
HGT	Horizontal gene transfer
HMM	hidden Markov model
IMG/M	Integrated Microbial Genomes/ Metagenomes
INFERNAL	inference of RNA Alignment
iTOL	Interactive Tree Of Life
KEGG	The Kyoto Encyclopaedia of Genes and Genomes
LSU	large subunit sequence
MAMMOTH	Matching molecular models obtained from theory
MAST	Motif Alignment and Search Tool
MBDBs	Molecular Biology Databases
MCMC	Markov chain Monte Carlo
MEME	Multiple EM for Motif Elicitation
MeSH	Medical Subject Heading
ML	Maximum Likelihood
MP	Maximum Parsimony
MSA	Multiple Sequence Alignment
MStA	Multiple Structure Alignments
MUSCLE	multiple Sequence Comparison by Log-Expectation
NCBI	National Center for Biotechnology Information
ncRNAs	noncoding RNAs
NJ	Neighbor Joining
OMSSA	Open Mass Spectrometry Search Algorithm
OPD	Open Proteomics Database
ORF	Open Reading Frame
OTU	Operational Taxonomic Unit
PAML	Phylogenetic analysis using maximum likelihood
PDB	Protein Data Bank
PHYLIP	phylogeny Inference Package
PIR	Protein Information Resource
PRF	Protein Research Foundation
PRIDE	Proteomics Identification Database
PSI-BLAST	Position-Specific Iterated BLAST
PSSM-based	Position Specific Scoring Matrix-based
RFLPs	restriction fragment length polymorphisms
rRNA	ribosomal RNA
RSA	Regulatory Sequence Analysis
SCFGs	stochastic context-free grammars
SCOP	Structure Classification of Proteins
SIDD	stress-induced duplex destabilization
SMART	Simple Modular Architecture Research Tool
SMD	Stanford Microarray Database
SSU	small subunit sequence
TF	transcription factor
TRs	transcriptional regulator
UniProt	Universal Protein Resource
UPGMA	Unweighted Pair Group Method with Arithmetic mean
WGS	Whole Genome Shotgun

Defining Statement

The vast and diverse data generated from the recent large-scale genomic projects has no precedent. For optimal use, it has been compiled and organized in different kinds of Molecular Biology Databases. This article reviews some of the most important databases and the software developed for their analyses.

Introduction

In the past decade, large scale projects have generated a vast amount of molecular biological data that has been deposited and organized into different Molecular Biology Databases (MBDBs). These MBDBs include information on genomics, proteomics, transcriptomics, interactomics, and metabolomics among many others. More than 1000 different MBDBs are publicly or commercially available and have become an essential element of every day scientific activity, making it possible to relate data to a specific biological problem and to assist the scientist to guide their research. MBDBs users can easily find answers to questions such as: which gene codes for an enzyme that performs a particular reaction in a specific organism? How and to what extent is such a gene regulated? What is the three dimensional structure of its corresponding enzyme? What other enzymes participate in the same metabolic pathway? Which scientific articles are related to this gene, protein, or pathway? (**Figure 1**). Furthermore, since most of the homologous proteins – proteins that have evolved from a common ancestor – are structurally and commonly functionally related, MBDBs users can easily identify, by simple sequence comparisons, other organisms carrying homologous genes and, in general, extend the aforementioned questions to these new set of organisms to find common properties and

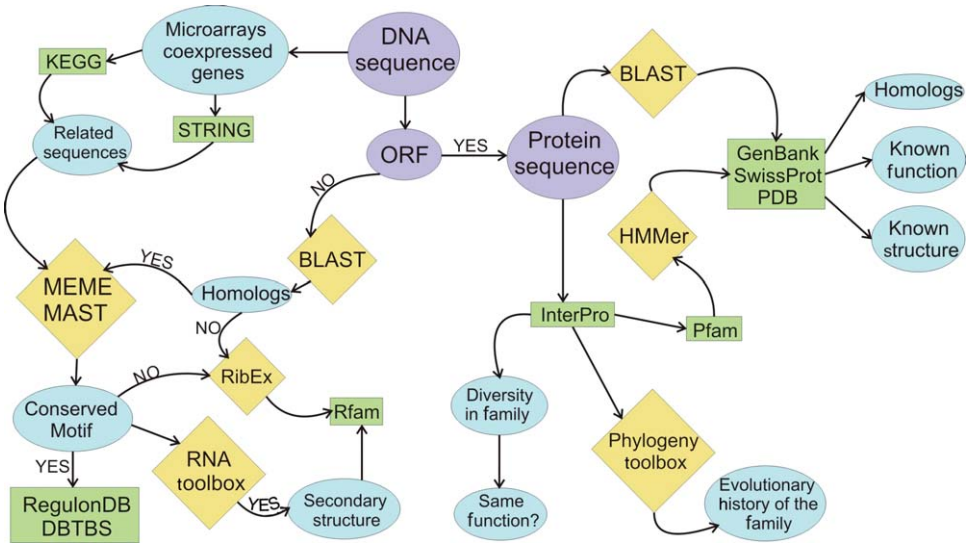

Figure 1 General flow pathway for the use of molecular biology databases.

generate general and properly supported conclusions. In fact, due to the relevance of MBDBs to the scientific community, one of the most important journals, Nucleic Acids Research devotes every year a freely available issue to biological databases and another one to papers describing web-based software resources. In this article, we review the main biological databases and the public domain software that has been developed to analyze them.

Literature Databases

The NCBI's PubMed database includes citations from life science journals for biomedical articles, most of them with abstracts and many with links to the full-text articles. It is heavily linked to other core Entrez databases, such as Nucleotide, Protein, Gene, Structure, and PubChem where it provides a crucial bridge between the data of molecular biology and the scientific literature. PubMed records are also linked to one another within Entrez as 'related articles' on the basis of computationally detected similarities using the Medical Subject Heading (MeSH) terms and the text of titles and abstracts. Also, it includes digitized back content of many journals, going back in some cases to the 1800s or early 1900s. ISI Web of Knowledge^{SM} platform covers literature databases of sciences, social sciences, arts, and humanities; ISI Web of Knowledge^{SM} also includes abstracts and links to the full-text articles. This platform can be very useful especially when literature outside life science field is required.

Gateways to Databases and Bioinformatics Tools

NCBI

The National Center for Biotechnology Information (NCBI) maintains 31 databases. The internet address of some of these and other important databases, and web servers are listed in **Table 1**. Entrez is an integrated database retrieval system that enables text searching using simple Boolean queries. Entrez searches rapidly across all NCBI databases and returns the counts of matching records in each database, including DNA and protein sequences (GenBank and Proteins, respectively), NCBI taxonomy, genomes, population sets, gene expression data, gene-oriented sequence clusters in UniGene, protein structures, alignment-based protein domains and the biomedical literature via PubMed, and online books. Results can be saved in a local file, shown in the browser as plain text. Results may be also sent to the Entrez clipboard where they may be recalled later using My NCBI. In addition, PubMed results and those from other databases may be emailed directly from Entrez or exported. Entrez's My NCBI allows users to store personal configuration options, such as search filters and document delivery providers. My NCBI also saves searches and can automatically email updated search results.

EMBL-EBI

The European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) server houses more than 50 databases and 50 bioinformatics tools. EMBL-EBI databases include DNA and protein sequences (EMBL and UniProt, respectively), protein structures, gene

Table 1 Table of molecular biology databases and software for their analysis

Literature databases	
Nucleic Acid Research Journal	http://nar.oxfordjournals.org/
PubMed	http://www.ncbi.nlm.nih.gov
ISI Web of Knowledge[SM]	http://www.isiknowledge.com/
Gateways to databases and bioinformatic tools	
NCBI	http://www.ncbi.nlm.nih.gov/
EMBL-EBI	http://www.ebi.ac.uk/
Nucleotide sequence databases	
Entrez Nucleotide	http://www.ncbi.nlm.nih.gov
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk
DNA Data Bank of Japan	http://www.ddbj.nig.ac.jp/
Protein sequence databases	
Entrez Protein	http://www.ncbi.nlm.nih.gov
UniProt	http://beta.uniprot.org
SwissProt	http://ca.expasy.org
Searching for sequence similarity	
BLAST	http://www.ncbi.nlm.nih.gov
MEME	http://meme.sdsc.edu
MAST	http://meme.sdsc.edu
Oligo-analysis	http://rsat.ulb.ac.be
ClustalW	http://www.ebi.ac.uk
T-Coffee	http://www.ch.embnet.org
MUSCLE	http://phylogenomics.berkeley.edu
HMMER	http://hmmer.janelia.org/
Protein databases	
COGs	http://www.ncbi.nlm.nih.gov
PROSITE	http://ca.expasy.org
PRINTS	http://www.bioinf.manchester.ac.uk
ProDom	http://prodom.prabi.fr/
Pfam	http://www.sanger.ac.uk
SMART	http://smart.embl.de/
TIGRFAMs	http://www.tigr.org
PIRSF	http://pir.georgetown.edu
InterPro	http://www.ebi.ac.uk
CDD	http://www.ncbi.nlm.nih.gov
Protein structure	
PDB	http://www.rcsb.org
SCOP	http://scop.mrc-lmb.cam.ac.uk
CATH	http://www.cathdb.info/
Protein structure visualization	
RasMol	http://www.umass.edu/microbio/rasmol/
DeepView	http://ca.expasy.org
Protein structure alignments	
MAMMOTH	http://ub.cbm.uam.es
Dali	http://ekhidna.biocenter.helsinki.fi
Protein structure prediction	
Swiss-Model	http://swissmodel.expasy.org/
MODELLER	http://www.salilab.org/modeller/
Fold recognition	
Phyre	http://www.sbg.bio.ic.ac.uk
PSIPRED	http://bioinf.cs.ucl.ac.uk
Gene3D	http://gene3d.biochem.ucl.ac.uk
Protein classification based on ontology	
The Gene Ontology	http://www.geneontology.org/
Resources for genome-scale analysis	
Entrez genome	http://www.ncbi.nlm.nih.gov
TaxPlot	http://www.ncbi.nlm.nih.gov
CMR	http://www.tigr.org
GOLD	http://www.genomesonline.org
Genome Reviews	http://www.ebi.ac.uk
Microbes Online	http://www.microbesonline.org/
Integr8	http://www.ebi.ac.uk
UCSC Genome Browser	http://genome.ucsc.edu/

(*Continued*)

Table 1 (Continued)

Metabolomics	
KEGG	http://www.genome.jp
EcoCyc	http://ecocyc.org/
Metagenomics	
IMG/M	http://img.jgi.doe.gov/m
CAMERA	http://camera.calit2.net/
Taxonomy databases	
UniProt Taxonomy	http://beta.uniprot.org
NCBI Taxonomy	http://www.ncbi.nlm.nih.gov
Phylogenetic analysis algorithms	
MrBayes	http://mrbayes.csit.fsu.edu/
PAUP*	http://paup.csit.fsu.edu/
PHYLIP	http://evolution.genetics.washington.edu
PAML	http://abacus.gene.ucl.ac.uk
TreeView	http://taxonomy.zoology.gla.ac.uk
Universal tree of life	
iTOL	http://itol.embl.de/
The ARB Project	http://www.arb-home.de/
Silva	http://www.arb-silva.de/
European Ribosomal RNA database	http://bioinformatics.psb.ugent.be/webtools/rRNA/
Ribosomal Database Project II	http://rdp.cme.msu.edu/
Greengenes	http://greengenes.lbl.gov/
Gene expression	
Stanford Microarray Data base	http://genome-www5.stanford.edu/
GEO	http://www.ncbi.nlm.nih.gov
E. coli GenExpDB	http://chase.ou.edu
ArrayExpress	http://www.ebi.ac.uk
CIBEX	http://cibex.nig.ac.jp
Resources for molecular structure and proteomics	
Swiss-2DPAGE	http://ca.expasy.org
OMSSA	http://pubchem.ncbi.nlm.nih.gov
Mascot	http://www.matrixscience.com/
PRIDE	http://www.ebi.ac.uk
OPD	http://bioinformatics.icmb.utexas.edu
Resources for gene regulation analysis	
RegulonDB	http://regulondb.ccg.unam.mx
DBTBS	http://dbtbs.hgc.jp/
Neural Network Promoter Prediction	http://www.fruitfly.org
WebSIDD	http://www.genomecenter.ucdavis.edu/benham/sidd/
Regulatory Sequence Analysis Tools	http://rsat.scmbb.ulb.ac.be
Predicted Attenuators in Bacteria	http://cmgm.stanford.edu/~merino
RibEx	http://ribex.ibt.unam.mx
Gene Context Tool	http://gecont.ibt.unam.mx
MBDBs and analysis protrams of RNA regulatory elements	
Rfam	http://www.sanger.ac.uk/Software/Rfam
INFERNAL	http://infernal.janelia.org/
MFOLD	http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html
Vienna RNA Package	http://www.tbi.univie.ac.at
HotKnots	http://www.cs.ubc.ca/labs/beta/Software/HotKnots/
RNAMST	http://bioinfo.csie.ncu.edu.tw
SCARNA	http://www.scarna.org
STRAL	http://www.biophys.uni-duesseldorf.de/stral
MARNA	http://www.bioinf.uni-freiburg.de/Software/MARNA/
FOLDALIGN	http://foldalign.ku.dk/
STEMLOC	http://biowiki.org
Dynalign	http://rna.urmc.rochester.edu
PSTAG	http://pstag.dna.bio.keio.ac.jp/
Mifold	http://www.lcb.uu.se/~evaf/Mlfold/
RNAshapes	http://bibiserv.techfak.uni-bielefeld.de
KnetFold	http://knetfold.abcc.ncifcrf.gov/
COVE	http://selab.wustl.edu/software/cove
RNAmine	http://rnamine.ncrna.org

(*Continued*)

Table 1 (Continued)

Dedicated integration systems for molecular biology databases	
DBGET	http://www.genome.jp
SRS	http://srs.ebi.ac.uk
STRING	http://string.embl.de/

expression data, molecular interactions, several kinds of alignments, literature, and so on. The server has different browsers; EB-eye performs searches in all the databases. Once you know the proper database entries, data retrieval can be performed easily by EBI Dbfecht; up to 200 entries can be retrieved. EBI-tools comprise mainly sequence, structure, and expression analysis tools.

Sequence Databases

Nucleotide Sequence Databases

NCBI GenBank, EMBL Nucleotide Sequence Database (EMBL), and the DNA Data Bank of Japan (DDBJ) are comprehensive databases that contain publicly available nucleotide sequences. The three organizations synchronize their data on a daily basis to ensure worldwide coverage. GenBank/EMBL/DDBJ data are submitted by the scientific community, primarily from large-scale sequencing projects. Each sequence entry includes a concise description of the sequence, the scientific name of the source organism and bibliographic references. Records cannot be updated, corrected, or amended without the permission of the original submitter. GenBank/EMBL/DDBJ records include individual genes, Whole Genome Shotgun (WGS), RNA, high-throughput cDNA, synthetic sequences, and environmental sequencing (for which the source organism is unknown). Due to its completeness and standing as a primary data provider, GenBank/EMBL/DDBJ is the initial source for many MBDBs.

Protein Sequence Databases

NCBI GenPept and TrEMBL used to be the comprehensive protein sequence databases; they were produced by translating GenBank and EMBL, respectively. Now, the protein sequences can be found at NCBI Protein and Universal Protein Resource (UniProt). NCBI Protein is compiled from a variety of sources, including Swiss-Prot, PIR (Protein Information Resource), PRF (Protein Research Foundation), PDB (Protein Data Bank), and translations from annotated coding regions in GenBank/EMBL/DDBJ. UniProt fused TrEMBL, Swiss-Prot, and PIR.

Swiss-Prot (more properly known as UniProtKB/Swiss-Prot) is a manually annotated protein sequence database with information extracted from literature and curator-evaluated computational analysis. Although Swiss-Prot is at least ten times smaller than UniProt or Proteins, it is a high quality database.

Searching for Sequence Similarity

The nucleotide and protein sequences collected in the sequence databases come from very heterogeneous organisms which might have diverged hundreds of millions of years ago. Regardless of this enormous period of time, these organisms share remarkably important similarities, since some of their genes and proteins were present in their last common ancestor. Such genes/proteins are said to be homologous. Homologous genes/proteins can be commonly identified in sequence databases by the comparisons of their nucleotide or amino acid sequences. There are different approaches to perform a sequence search for homologous proteins: the election of the most convenient depends on many factors, such as the size of the database; the expected similarity of the query with their potential targets; the available computer resources; and the speed of the search and the required accuracy of the results among others. Here, we summarize the main search approaches and their corresponding publicly available software.

Genomic Searches Using Pairwise Comparison

The simplest approach to perform a database search considers the comparison of a pair of sequences, commonly known as pairwise comparison. This is based on the alignment of two sequences; it can denote that the two sequences are similar either globally or locally. If the similarity is global, similarity may reflect that they are closely related and, hence, they may have the same function. If the similarity is local, this may indicate evolutionary constraints restricted to these particular regions. In order to perform extremely fast and accurate sequence similarity searches of a particular sequence (commonly known as the query sequence) against nucleotide or amino acid databases, the NCBI has developed the BLAST (Basic Local Alignment and Search Tool) set of programs that can be used locally or via their web server. BLAST programs look for small sets of continuous characters or 'words' in the sequences of the database that corresponds to fragments of the sequence query. The length of the words can be specified by the user and depends on the kind of database; commonly, 3 for amino

acid databases and 11 for nucleotide databases. BLAST assumes that the significant alignments contain highly similar pairs of aligned words. The similarity between any pair of words is scored using substitution matrices, such as BLOSUM and PAM, which express the probabilities that one amino acid can be replaced by another in a set of homologous proteins. If the similarity between two words has a score greater than a specific predetermined value, the aligned word is called a 'hit' and is further considered in the analysis. After the identification of all hits, BLAST tries to extend each hit in both directions to connect neighbor hits in a bigger alignment. Insertions and deletions are not considered during this stage of analysis. The resulting extended aligned regions are further considered only if the corresponding alignment scores are greater than a predetermined cut-off value. Finally, BLAST performs a new alignment between the query sequence and the database sequence allowing gaps to extend the regions of the sequence similarity. Each alignment of the search is scored and assigned to a measure of statistical significance called the expectation value (E-value, i.e., the number of alignments with at least the same score that would have been expected to occur in the database by chance) which is used to sort and limit the alignments reported to the user. BLAST takes into account the amino acid composition of the query sequence in the estimation of statistical significance. This composition-based statistical treatment, used in conventional protein BLAST searches, tends to reduce the number of false-positive database hits.

The aforementioned heuristic approach taken by BLAST importantly speeds the process of searching for similar sequences as much as 50 times, in comparison to exhaustive algorithms, that search for the best alignment solution. This speed characteristic of BLAST is particularly important considering the enormous sizes of the commonly used databases such as GenBank or Swiss-Prot.

Considering the type of query sequence and database, the following alternatives in a BLAST search are available: (1) the query and the databases are nucleotide sequences (blastn); (2) the query and the databases are amino acid sequences (blastp); (3) the query is a nucleotide sequence that is translated into its six reading frames to be compared with an amino acid database (blastx); and (4) the query is an amino acid sequence and is compared with a nucleotide database dynamically translated into their six reading frames (tblastn). The Web BLAST output service offers a new Tree View option to create a dendrogram that clusters sequences according to their distances from the query sequence. This display is helpful for organizing the presence of aberrant or unusual sequences or natural groupings of related sequences such as members of a gene family or homologues from other species in the BLAST output. In addition to the aforementioned alternatives, a comparison of two DNA or protein sequences that produces a dot-plot representation of the alignments can be obtained using BLAST2Sequences.

For genomic searches, the MegaBLAST program can be used. This program was designed to find nearly exact matches and operates up to ten times faster than the standard nucleotide BLAST. MegaBLAST is the default search program for NCBI's Genomic BLAST pages. Genomic BLAST may be used to search the genomic sequence of an organism. Its searches can be displayed within their genomic context using the Map Viewer to show the location of neighboring genes and nearby genomic landmarks.

For finding distant relatives of a protein, a much more sensitive BLAST version called PSI-BLAST (Position-Specific Iterated BLAST) can be used. This program initially performs a standard BLAST search to identify closely related proteins which are then used to elaborate a consensus profile of the protein family that reflects the specific tendency of certain residues to be present in particular positions of the sequence. The profile is used as a new query to perform a new search against the database. As a result of this new search, new members of the family could be identified and considered in addition to the original set of proteins to construct a new and more representative profile. This process can be repeated iteratively until the user considers that all the members in the database have been identified.

Genomic Searches Using Profiles

Structurally relevant residues as well as catalytic active sites of enzymes have a tendency to be conserved in a family of homologous proteins. These conserved regions or sequence patterns occur repeatedly in a group of related protein or DNA sequences and are known as motifs. Motifs can be represented by a profile matrix where the frequencies for each amino acid at each position are evaluated from the conserved region's residue distribution rather than from a more general distribution. Motifs can be used to search, in a database, for other proteins of the family. This is particularly relevant to identify distantly related proteins that might not present evident similarity across their entire sequences, but conserve the essential residues of the group. Motifs are not exclusive to protein sequences; they can also be used to represent relevant residues in a family of RNA (e.g., tRNA and catalytic RNA) or DNA (e.g., regulatory protein-binding sites) sequences. One of the advantages of the sequence searches based on motifs over the pairwise methods is their ability to use the characteristics of the whole family of sequences during the database search. One of the most common programs used to identify motifs in a set of nucleotide or amino acid related

sequences is MEME (Multiple EM for Motif Elicitation). This program finds, in a set of unaligned sequences, all the motifs that are statistically over-represented. The most worth-mentioning options of MEME allow the user to select: (1) if the motif should be found as one or more occurrences per sequence; (2) the minimum and maximum size of the motif; (3) the minimum statistical significance of the motif; and (4) the maximum number of motifs to find. MEME represents each motif as a scoring matrix that expresses the probability of each possible residue at each position in the pattern. This matrix can be used to search databases for homologue or related sequences using the MAST (Motif Alignment and Search Tool) program. MAST calculates the statistical significance of the matches of a group of motifs characteristic of a protein or a DNA sequence family in a target sequence. For each motif, MAST finds the position in the sequence that best matches, and it represents the statistical significance of the match as a p-value (i.e., the probability of observing a match with a score at least as good when the motif is compared to a random sequence). In order to refine a motif, the resulting sequences obtained by MAST can be selected in conjunction with the seed sequences to perform a new MEME–MAST cycle. This cycle can be repeated until no more new sequences are obtained, resulting in a refined motif. Individual MEME motifs do not contain gaps; instead patterns with gaps are represented by MEME by two or more separate motifs. These individual motifs can be integrated into a more complete model using the program Meta-MEME that combines the set of motifs into a single one using hidden Markov models (HMMs) (see below).

Oligo-analysis is a second program that can be used to perform sequence searches based on motifs. It was originally created to uncover binding sites for transcription factors in *Saccharomyces cerevisiae*, but it works very nicely on any organism. In contrast with heuristic methods, oligo-analysis is an exhaustive algorithm. Its range of detection is however limited to relatively simple patterns: short motifs with a highly conserved core. These features seem to be shared by a good number of regulatory sites in yeast. The oligo-analysis program is contained in the RSA (Regulatory Sequence Analysis) Tools website (see below). Nicely, oligo-analysis has an option to create random sequences; hence, it provides a negative control to assure that the patterns obtained are statistically significant. Analogous to the previously described MAST program, it possesses a tool named pattern matching, which searches for the motif in the genome of interest, and it will draw a new feature-map with the newly predicted sites. Hence, a cyclic process can also be used as the one described for MEME and MAST. In contrast to the MEME and MAST programs, oligo-analysis works better if all the analyzed sequences belong to a single organism, because it adjusts the frequencies of the nucleotide composition of the genome to the search of over-represented motifs. MEME and MAST also posses this option, but they work better than oligo-analysis especially with large motifs and with a set of orthologous genes.

A comparison of these and other sequence search methods was reviewed by Tompa and colleagues in 2005.

Genomic Searches Using Multiple Sequence Alignment and Hidden Markov Models

Conserved motifs can also be identified by the comparison of more than two sequences at a time in a process called Multiple Sequence Alignment (MSA). Since the computational process of MSA is much more complex than the simple pairwise alignments, most of the MSA programs use heuristic approaches. One of the most common heuristic MSA protocols is ClustalW, which considers the progressive pairwise alignments on successively less closely related sequences. It can be used via web servers or locally. A second alternative of a progressive alignment program is T-Coffee. This program calculates pairwise alignments by combining the direct alignment of the pair with indirect alignments that aligns each sequence of the pair to a third sequence. T-Coffee is slower than ClustalW, but usually generates more accurate alignments, especially if the sequences to be aligned are distantly related. Finally, MUSCLE (Multiple Sequence Comparison by Log-Expectation) is a third commonly used MSA program. MUSCLE is claimed to perform better and faster MSA than ClustalW or T-Coffee, since the distance measure that it uses to assess the relatedness of two sequences is updated between iteration stages.

MSA are considered in molecular biology as an important primary source of information for different studies, such as phylogenetic analysis (see below) or homology database searches. In the former case, sequence motifs can be identified directly from the conserved regions of the MSA and used to construct models that represent the essential regions of the nucleotide or protein group of sequences. One efficient protocol to simultaneously consider the different motifs from a MSA is by HMMs. A Markov model is a statistical model in which the probability of an event depends on the immediately previous event. In a HMM, the parameters of the process are 'hidden' and the challenge is to determine them from the observable data. In the case of biological sequences, the observed events are represented by the presence of certain types of residues (i.e., nucleotide or amino acid) in a specific column of the alignment and the 'hidden' events are the biological properties of the process, for example, the secondary structure of a protein or the probability of a given DNA sequence to be part of a gene. One of the most popular analysis packages that uses HMMs to perform database searches is HMMER which includes, among others, programs to build the HMM from a multiple

sequence alignment (hmmbuild), to calibrate the model and determine the parameters for more sensitive searches (hmmcalibrate) and to search, in a sequence database, for sequences that match an HMM (hmmsearch).

There are several databases of relevant precompiled HMM of conserved protein families or conserved domains such as Pfam, SMART, TIGRFAMs, and PIRSF which are considered below.

Protein Databases

The Cluster of Orthologous Groups of Proteins Database

The NCBI's Clusters of Orthologous Groups of proteins (COGs) database has been designed as an attempt to classify proteins from completely sequenced genomes on the basis of the orthology concept. Orthologues are direct evolutionary counterparts related by vertical descent as opposed to paralogues which are genes within the same genome related by duplication. Typically, orthologous proteins have the same domain architecture and the same function.

The COGs reflect one-to-one relationships. COGs have been identified on the basis of all-against-all sequence comparison of the proteins encoded in complete genomes using the gapped BLAST program. The construction procedure is based on the simple notion that any group of at least three proteins from distant genomes that are more similar to each other than they are to any other proteins from the same genomes are most likely to belong to an orthologous family. This prediction holds even if the absolute level of sequence similarity between the proteins in question is relatively low, and thus the COG approach accommodates both slowly-evolving and fast-evolving genes. Moreover, this allows COGs to accommodate the possibility that a single (or multiple) protein(s) from one genome may be related to several paralogues in a second genome (one-to-many or many-to-many relationships).

The most straightforward application of the COGs is the prediction of functions of individual proteins or protein sets. This is done by fitting proteins into a COG using the COGNITOR program. The COG website offers automatic means to isolate all COGs with a particular phylogenetic pattern (i.e., a pattern of species that are represented or not represented in a given COG), for example, those that are found only in pathogenic bacteria. More generally, the COG system is a convenient platform for a variety of evolution-oriented analyses of protein families.

The COG website contains the following principal types of data: (1) list of all COGs organized by the (predicted) functional category and hyperlinked to (2) individual COG pages; each COG page shows the respective phylogenetic pattern and is hyperlinked to: (a) pictorial representations of BLAST search outputs for each member of the COG, (b) a multiple alignment of the COG members produced automatically using the ClustalW program (see above), and (c) a cluster dendrogram generated using the BLAST scores (see above) as the measure of similarity between proteins; (3) the COGNITOR page, where a protein sequence can be pasted, searched against the database of proteins from complete genomes, and assigned to a COG; (4) a phylogenetic pattern search tool; and (5) a matrix of co-occurrence of genomes in COGs.

Moreover, a web page that contains additional structural and functional information on the COG as a whole and individual members is now associated with each COG. These pages include systematic classification of the COG members under the current classification systems for enzyme or transporters (if applicable); indications of which COG members (if any) have been genetically and biochemically characterized; information on the domain architecture of the proteins comprising the COG and the three dimensional structure of the domains if known or predictable; a succinct summary of the common structural and functional features of the COG members; and peculiarities of individual COG members.

Protein Functional Classification and Protein Signatures

Databases with signatures diagnostic for protein families, domains, or functional sites are important tools for the computational functional classification of newly determined sequences that lack biochemical characterization. Computational annotation of protein function is generally obtained via sequence similarity: once a close neighbor with known function has been identified, its annotation is copied to the sequence with known function. This strategy works very well in functionally homogeneous families.

In some cases the sequence of an unknown protein is too distantly related to any protein to detect its resemblance by pairwise sequence alignment (see above). However, relationships can be revealed by the occurrence of a particular motif in its sequence. These motifs, typically around 10–20 amino acids in length, arise because specific residues and regions thought or proved to be important to the biological function of a group of proteins are conserved in both sequence and structure during evolution. These biologically significant regions or residues are generally: enzyme catalytic sites; prosthetic group attachment sites; amino acids involved in binding a metal ion; cysteines involved in disulphide bonds; and regions involved in binding a molecule or another protein. Some families are defined not just by one motif but by the co-occurrence of two or more motifs of low specificity.

The increasing amount of genomic sequences that need to be annotated has lead to proliferation of protein domain families and protein domain signature databases.

Protein domains are units of molecular evolution, usually associated with particular aspects of molecular function such as catalysis or binding. In general, they represent discrete units of three dimensional (3D) structures. Most proteins are built up in a modular fashion from two or more domains fused together. The identification of functionally characterized domains in protein sequences may give the first clues as to their molecular and cellular function.

PROSITE

PROSITE is an annotated collection of motif descriptors dedicated to the identification of protein families and domains. The motif descriptors used in PROSITE are either patterns or profiles, which are derived from multiple alignments of homologous sequences. PROSITE patterns are short sequence motifs, while PROSITE profiles are position specific score matrices. Profiles characterize protein domains over their entire length, and they are more sensitive than patterns. Profiles and patterns have complementary qualities. Patterns, confined to small regions with high sequence similarity, are often powerful predictors of protein functions such as enzymatic activities. Profiles covering complete domains are more suitable for predicting protein structural properties.

PRINTS

PRINTS database houses a collection of protein fingerprints. Fingerprints are groups of conserved sequence motifs that together provide diagnostic signatures for protein families. The tools available for searching PRINTS are (1) a BLAST server, for searches against sequences matched in PRINTS database and (2) the FingerPRINTS, which can suit for searches against fingerprints in the database – this affords greater specificity than the BLAST implementation. PRINTS has a hierarchical structure which allows associations to be traced from subfamily to superfamily relations. This is relevant to putative distantly related clan members that share no significant sequence similarity.

ProDom

ProDom is a comprehensive database of protein domain families generated from the global comparison of all available sequences. ProDom families are built by an automated process based on a recursive use of PSI-BLAST homology searches. The ProDom website allows querying of ProDom in a variety of ways, such as accession number, ProDom families, related databases, and keywords. The output is either information on a given domain family or cartoons displaying the domain arrangements of all proteins matching the query. One can also compare a sequence of interest via BLAST to the ProDom database. ProDom will suggest a possible domain arrangement for any query protein. When 3D structures are available for target domains, the output is directly linked to both SWISS-MODEL and Geno3D servers.

Pfam

Pfam is a database of protein domains and families. It contains curated multiple sequence alignments for each family and corresponding profile HMMs (see above). Pfam families are divided into two categories, Pfam-A and Pfam-B. Each Pfam-A family consists of a curated seed alignment containing a small set of representative members of the family, a profile HMM built from the seed alignment and an automatically generated full alignment which contains all detectable protein sequences belonging to the family. Pfam-B entries are automatically generated from the ProDom database and are represented by a single alignment. The use of representative seed alignments for Pfam-A families allow efficient and sustainable manual curation of alignments and annotation, while the automatic generation of full alignments and Pfam-B clusters ensures that Pfam is a comprehensive classification of protein families that scales effectively with the growth of the sequence database.

Within Pfam, several metagenomic datasets have been included. These new datasets contain many novel protein sequences, which are currently unannotated. This section, within Pfam, enables the community to assess our current understanding of the domain composition found in such environmental datasets. Moreover, this will provide a potential source of new Pfam families and/or allow verification of families where there are few representatives.

SMART

The Simple Modular Architecture Research Tool (SMART) is an online resource used for protein domain identification and the analysis of protein domain architectures. The basic data of SMART are high-quality manually derived alignments of protein domain families. As HMMs (see above), SMART alignments allow the identification of protein domain in sequence databases. Protein sequences can be scanned for the presence of important catalytic amino acids. Absence of one of these amino acids very likely results in loss of catalytic activity. The data provide a framework for understanding the evolution and function of genes and proteins throughout the living world. Its genomic perspective allows further cross-referencing with protein–protein interaction maps, making SMART an invaluable tool for systems biologists to interpret pathways and networks.

TIGRFAMs

TIGRFAMs is a collection of manually curated protein families consisting of HMMs (see above), multiple sequence alignments, commentaries, Gene Ontology (GO) assignments, literature references, and pointers

related to TIGRFAMs, Pfam, and InterPro models. TIGRFAMs contains models of full-length proteins and shorter regions at the level of superfamilies, subfamilies, and equivalogues, where equivalogues are sets of homologous proteins conserved with respect to function since their last common ancestor. The models in the TIGRFAMs database have been built specifically to aid in automated annotation of microbial genes, particularly by focusing on the creation of equivalogue family models. TIGRFAMs uses the term equivalogue to describe the relationship of proteins conserved in function since their last common ancestor, where both orthology and horizontal gene transfer may be part of the evolutionary history.

PIRSF

PIRSF is a network classification system that accommodates a flexible number of levels from superfamily to subfamily to reflect varying degrees of sequence conservation. Members of a PIRSF homeomorphic family share full-length sequence similarity with a common domain architecture (homeomorphic) and have a common evolutionary origin (monophyletic). PIRSF HMMs are designed to cover the full length of a protein sequence, and thus to include all domains within the sequence. In this way, PIRSF homeomorphic families tend to encompass one or more of the existing InterPro domain entries and show the domain composition of UniProt sequences. Classification based on full-length protein allows annotation of both generic biochemical and specific biological functions, identification of domain and family relationships, and classification of multidomain proteins.

Integrated Databases

InterPro: A Database of Protein Families

The EMBL InterPro incorporates the major protein signature databases into a single resource. These include PROSITE, which uses regular expressions and profiles; PRINTS, which uses Position Specific Scoring Matrix-based (PSSM-based) fingerprints; ProDom, which uses automatic sequence clustering; and Pfam, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, and Gene3D, all of which use HMMs (see above). Protein signatures from these databases that describe the same family or domain, in terms of sequence positions and protein coverage are integrated into single InterPro entries, to which are added annotation and cross-references. Proteins with 3D structures modeled by MODBASE and SWISS-MODEL have links to the structure predictions from the matched graphical views. These links complement the experimentally determined structures in the PDB.

InterPro unites all these databases capitalizing on their individual strengths, producing a single entity that is far greater than the sum of its parts. A primary application of InterPro is the annotation and functional classification of uncharacterized sequences. The EBI is using InterPro for enhancing the automated annotation of TrEMBL. InterPro has also proven its usefulness for whole proteome analysis with comparative genome analysis in several organisms. InterPro has provided a useful tool for protein sequence analysis and characterization.

Conserved Domain Database

The NCBI conserved domain database (CDD) contains PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the SMART, Pfam, and domain alignments derived from COGs. CDD attempts to collate the set of protein domains characterized so far and to organize related domain models in a hierarchical fashion, meant to reflect major ancient gene duplication events and subsequent functional diversification.

Whenever possible CDD hits are linked to structures which, coupled with a multiple sequence alignment of representatives of the domain hit, are equipped with advanced alignment-building tools that use the PSI-BLAST and threading algorithms. In alignment curation, information from 3D structure and structure superposition is considered when possible to define structurally conserved cores.

The Conserved Domain Architecture Retrieval Tool allows searches of protein databases on the basis of a conserved domain and returns the domain architectures of database proteins containing the query domain.

Protein Structure

Protein Data Bank

PDB includes all the public 3D structures for proteins, nucleic acids, and carbohydrates. PDB records contain atomic (x, y, z) coordinates of macromolecules, a brief macromolecular description, the authors of the structure, the biological source, the protein or nucleotide sequence, literature references, and some relevant experimental data. Most 3D structure data are obtained from X-ray crystallography and NMR spectroscopy; they provide a wealth of information on the biological function, on mechanisms linked to the function, and on the evolutionary history of macromolecules and relationships between them. PDB interface provides search and retrieval interfaces and cross references to other structural databases.

There are some domain structure classification databases. Structure Classification of Proteins (SCOP) contains a structural and evolutionary classification of

the proteins in the PDB. In SCOP, a multidomain protein is split up into its constituent domains, which are then considered separately. SCOP is a hierarchical classification system, which utilizes four levels: class, fold, superfamily, and family. class and fold lack evolutionary relationship evidence. CATH is another structure domain hierarchical classification system. The main difference with SCOP is that CATH is a semiautomatic classification while SCOP is manually done.

Protein Structure Visualization

PDB files are long text files; atom coordinates are not comprehensible by reading these files. Moreover, PDB files do not include connectivity data. For simplification, protein structures can be represented in many ways depending on the information to be conveyed (e.g., wire models for comparisons, ribbon models to highlight secondary structures, ball and stick models to detail, and surface models for electrostatic potentials). There are several PDB visualization tools that transform the coordinates into virtual 3D structures. One of the most popular free software available is RasMol, whose drawback is that one must master its command-line language. An amateur-friendly, free visualization program is DeepView (Swiss Pdb-Viewer), which has links to many bioinformatics resources.

Protein Structure Alignments

Long distance evolutionary relationships can be detected by protein structure similarities. This is useful when no detectable protein sequence homologues are available for the gene of interest, but instead, the structure of the gene of interest is known. If this is the case, structural homologues can be found. MAMMOTH (Matching molecular models obtained from theory) is a pairwise comparison method. MAMMOTH takes a PDB file as query and searches in a PDB files database, like SCOP. Using a heuristic algorithm, it calculates a structural similarity score based on the likelihood of obtaining an alignment between two proteins or between two conformations of the same protein by chance.

Moreover, Multiple Structure Alignments (MStA) can also be performed using Dali and MAMMOTH-mult server. Dali searches the PDB for those structures similar to the query. For this server, the query can be either a coordinate file or a PDB ID. Dali can also perform a two-structure alignment. On the contrary, MAMMOTH-mult server can be used in two ways: it can either multiple align a target protein against a given SCOP superfamily or align among them a set of input proteins. Both servers will return an MStA via e-mail.

Protein Structure Prediction

In 1962, C.B. Anfinsen demonstrated for ribonuclease that protein structure is encoded in the protein sequence. Since then, many efforts have been made for protein structure prediction based on the amino acid sequence. Nowadays, some programs based on homologous structures are good predictors. Some programs for structure determination, based on physicochemical and statistical properties, have been developed successfully, but these programs still need a user with protein structure and bioinformatics expertise. Homology based programs work much better.

Swiss-Model is a fully automated protein structure homology server. The entry is a protein sequence and the output is an atom coordinate file and the names of the structure templates. This server divides the prediction into five stages. First, it finds protein homologues with known structure by sequence comparison. Second, it selects those templates with more than 25% identity in tracks of more than 20 residues; some sequences do not have homologues with known structure, so for these cases Swiss-Model cannot give an answer. Third, it makes some input files. Fourth, it generates the model by threading the new amino acid sequence into the known backbone. Fifth, it refines the model by energy minimization to avoid clashes and holes.

MODELLER is another program for protein structure prediction based on homology; it is more powerful than Swiss-Model, since many parameters can be changed. Nevertheless, MODELLER is not fully automated and it must be installed in your own computer.

Fold Recognition Tools

Another approach to know the 3D structure of a sequence is fold recognition. Fold recognition is used when the query protein is distantly related to protein(s) with known structure; many of these programs are based on homology detection by HMM. The structures generated by fold recognition used to be less accurate than those from prediction programs, and should be used carefully. PSIPRED and PHYRE are protein recognition programs that run via web pages.

Fold Recognition Databases

The Gene3D database is focused on providing structural annotation for protein sequences without structural representatives, including the complete proteome sets. The structural annotation is generated using HMMs based on the CATH domain families. Gene3D maps CATH domain families to protein sequences. This is a similar task to that carried out by SUPERFAMILY for SCOP. The Gene3D website includes a BLAST search

facility that will identify the likely family that the query sequence belongs to.

Protein Classification Based on Ontology

The GO project has defined specific terms and vocabulary to describe common properties of genes and proteins. GO is a structured network consisting of defined terms and relationships between them that describe three attributes of gene products: their molecular function, biological process, and cellular component. There are three separate aspects to this effort: first, the development and maintenance of the ontologies themselves; second, the annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases; and third, development of tools that facilitate the creation, maintenance, and use of ontologies.

Resources for Genome-Scale Analysis

The NCBI Entrez Genome provides access to over 370 complete microbial genomic sequences. Specialized viewers and BLAST pages are also available for viruses. Genomes are chosen from alphabetical listing or a phylogenetic tree and can be examined at increasing levels of details ranging from a graphical overview of an entire genome to the level of a single gene. At the level of a genome or chromosome, a coding region display gives the locations of coding regions, the lengths, names, and GenBank identification numbers of the protein products. An RNA genes view lists the locations and names for ribosomal and transfer RNA genes. A summary of COG functional groups is also presented. At the level of a single gene, links are provided to sequence neighbors for the implied protein with links to the COGs database.

For complete microbial genomes, precomputed BLAST neighbors for protein sequences, including their taxonomic distribution and links to 3D structures, are given in TaxTables and PDBTables, respectively. The Entrez Genome Project database is supported by providing an overview of the status of complete and in-progress large-scale sequencing, assembly, annotation, and mapping projects. For bacterial organisms, Genome Project indexes a number of characteristics of interest to biologists, such as organism morphology and motility; environmental requirements, such as salinity, temperature, and pH range; oxygen requirements; and pathogenicity. The database allows genome sequence centers to register their project early in the sequencing process so that project data can be linked to other NCBI-hosted data at the earliest opportunity.

NCBI's TaxPlot plots similarities in the proteomes of two organisms to that of a reference organism for more than 580 bacterial and archaeal genomes.

Comprehensive Microbial Resource

Comprehensive Microbial Resource (CMR) contains robust annotation of all complete microbial genomes and allows a wide variety of data retrieval. Retrievals can be based on protein properties such as molecular weight or hydrophobicity, GC-content, functional role assignments and taxonomy. The CMR also has special web-based tools divided into Genome Tools and Comparative Tools. Genome Tools include a list of the available genomes; a list of all genes in a selected genome; a list of genes ordered by categories; detailed information on each of the DNA molecules found in the organism (chromosomes, plasmids) including the topology (linear or circular), length, A, T, G, C percentage and number of genes; information on the characteristics of organisms derived from genomic data and literature sources; KEGG pathway displays; lists of all the intergenic regions for a selected DNA molecule as well as lists all the interRNA regions for a selected DNA molecule. The Genome Tools also possess different graphical displays to view the genome, the genetic context of selected genes, and a circular image of a selected DNA molecule, including representation of all genes on the molecule as well as all tRNAs and rRNAs. One can also retrieve sequence or a list of genes between a pair of coordinates for the selected DNA molecule, search all proteins in a genome for a given motif, display restriction digest information for a genome, display a computer model of a 2-dimensional gel for a selected organism, display the %GC for a set 'window' of nucleotides across the entire DNA molecule, show codon usage within a genome, and use the computer program Primer3 to find primers for a selected gene or organism.

Moreover, the Comparative Tools include Protein Homology Tools, which show the number of proteins in a reference genome that have hits up to 15 selected comparison genomes, the display of orthologue information across genomes, and the number of protein hits a reference genome has in common with all of the genomes in the CMR based on blast searches. It can also compare the number of protein hits between a reference and comparison organism in a scatter plot and compare the %GC content between a reference and comparison organism. It can align genomes using MUMmer to align any two DNA molecules in the database based on exact DNA sequence matches. It uses NUCmer to compare two closely related DNA molecules and PROmer to compare the protein sequences between two DNA molecules in the database.

The Genomes Online Database

The 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. According to GOLD, nearly 700 microbial genomes have been published at the time of writing with over 3000 other projects ongoing and in the process of being launched.

Genome Reviews

The Genome Reviews database provides an up-to-date, standardized and comprehensively annotated view of the genomic sequence of organisms with completely deciphered genomes. Currently, Genome Reviews contains the genomes of archaea, bacteria, bacteriophage, and selected eukaryotes. Genome Reviews is available as a MySQL relational database or a flat file format derived from the EMBL Nucleotide Sequence Database.

Microbes Online

Microbes Online is a publicly available suite of web-based comparative genomic tools, which include operon and regulon predictions, a multispecies genome browser, a multispecies GO browser, a comparative KEGG metabolic pathway viewer, a Bioinformatics Workbench for in-depth sequence analysis, and Gene Carts that allow users to save genes of interest for further study while they browse. An additional interface for genome annotation is provided.

Integr8

The Integr8 web portal provides easy access to integrated information about deciphered genomes and their corresponding proteomes. Available data includes DNA sequences (from databases including the EMBL Nucleotide Sequence Database, Genome Reviews, and Ensembl); protein sequences (from databases including the UniProt Knowledgebase and IPI); statistical genome and proteome analysis (performed using InterPro, CluSTr, and GOA); and information about orthology, paralogy, and synteny.

UCSC Genome Browser

The UCSC Genome Browser provides support for genome-centric exploration of archaeal genomes enriched with data from computational analysis and experimental studies.

Microbial genomes can be explored using a variety of analysis tools provided by resources that often further enrich the data in archival or curated public resources, some of which are described in this article.

Metabolomics

The Kyoto Encyclopedia of Genes and Genomes

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a comprehensive set of metabolic pathways. KEGG is displayed as graphic maps, than can be viewed as a network of genes (enzymes) or as a network of compounds. KEGG is constructed from four databases: PATHWAY contains 354 reference pathways manually drawn, GENES is a collection of the genes and proteins from complete genomes, LIGAND is a compilation of chemical compounds including drugs approved in United States and Japan, and BRITE is a collection of hierarchies and binary relations of the other databases. KEGG includes genes of 574 bacteria and 49 archaea (KEGG version 45; 8 January 2008). KEGG graphic interface is nicely presented. Reference pathway maps show general organization metabolism in various species; these maps are boxes (proteins) and circles (compounds) connected by arrows. Color representations indicate presence of those proteins in a particular organism. Atlas tool displays global maps that can be colored by the user to highlight any genes or compounds. KEGG can be browsed directly over the pathway graphs, or by NCBI, UniProt, or KEGG identifiers. Lists of orthologous genes are provided as well as several cross-links to other databases. KEGG can be used for modeling and simulation, search, and retrieval. It also includes sections of genetic information processing and environmental transduction (e.g., two-component systems).

EcoCyc and BioCyc: An Encyclopedia of Genes and Metabolism

EcoCyc is a comprehensive database resource for *Escherichia coli*. It contains curated information of genome, transcription regulation, membrane transporters, and metabolism. Chromosome maps can be browsed by gene name, or nucleotide number, to see graphical representation of that particular region. Pathways can also be browsed to obtain maps. Fully bibliographic information is obtained. BioCyc is the extended version for other genomes and metagenomes. Genomes can be selected for comparative analysis. Since these databases rely on available literature, EcoCyc is the more interesting of these databases.

Metagenomics

'Metagenomics' describes the functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample. Current microbiological culturing techniques are inadequate for studying the vast majority of microorganisms. Consequently, many organisms

remain underrepresented in the main sequence databases. Recently, with the advent of better sequencing technologies, large samples from environments such as the sea have been sequenced directly, thereby avoiding the need for culturing. Sequencing using this approach gives rise to many sequences from a diverse set of organisms, albeit at low read coverage and with no knowledge of the source organisms. For example, these advances have enabled the adaptation of shotgun sequencing to metagenomic samples. A 2004 metagenomic study of the Sargasso Sea found DNA from nearly 2000 different species including 148 types of bacteria never seen before, obtaining over 1 million kilobase of nonredundant sequence. Metagenomics is reviewed in 'Metagenomics'.

Structure and Function of Microbial Communities

Metagenomics concerns the extraction, cloning, and analysis of the entire genetic complement of a habitat. Metagenome analysis is expected to provide a comprehensive picture of the gene functions and metabolic capacity of microbial communities. Several statistical tools for describing and comparing microbial communities have been developed by Patrick Schloss and Jo Handelsman. Some of these tools include DOTUR, which calculates an estimate of the richness and diversity in a community; SONS, which defines the structure and memberships of two communities; LIBSHUFF, which is a statistical test to compare community structures and determines whether two samples are drawn from the same population or whether one is a subset of the other; and TreeClimber, which describes gene flow from a given phylogeny, that is, determines whether the differences between two communities arose due to random variation or whether lineages from one community had become more dominant through negative or positive selection pressures.

Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) is a web resource for metagenomic research. CAMERA's debut coincides with the publication of the Venter Global Ocean Sampling expedition's extensive dataset cataloguing of over 6 million new genes from uncultured marine microbes.

CAMERA's aim is to create a rich, distinctive data repository and bioinformatics tools resource that will address many of the unique challenges of metagenomics and enable researchers to unravel the biology of environmental microorganisms. CAMERA's database includes environmental metagenomic and genomic sequence data, associated environmental parameters ('metadata'), precompiled search results, and software tools to support powerful cross-analysis of environmental samples.

Integrated Microbial Genomes/Metagenomes

Integrated Microbial Genomes/Metagenomes (IMG/M) is an experimental metagenome data management and analysis system, which provides tools and viewers for analyzing both metagenomes and isolated genomes individually or in a comparative context. Comparative analysis of the metagenomes in the context of available reference isolated genomes could potentially reveal large-scale patterns of biochemical interactions and habitat-specific correlations in the host environment that might otherwise be missed.

Functional roles of genes can be characterized in the context of pathways, whereby pathways are associated with genes via gene products that can function as enzymes catalyzing individual reactions of metabolic pathways. Similar to isolated microbes, the metabolic capacity of a whole microbiome can be characterized by analyzing the metabolic maps inferred from the gene content and distribution of its composite genome. Comparative data analysis plays an important role in understanding the biology of isolated microbial genomes. Similar to isolated genomes, the analysis of metagenomes in the comparative context of other genomes is substantially more efficient than analyzing each metagenome in isolation. Microbiome samples can be compared in terms of presence and abundance of certain gene families or of certain metabolic pathways. These analyses help to infer the metabolic capabilities of the component organisms in the community, and thus identifying the key members of the microbiome that perform community-essential tasks and pinpoint the metabolic interactions within the microbiome and between the microbiome and its host environment.

IMG integrates bacterial, archaeal, and selected eukaryotic genomic data collected from multiple data sources. In addition to the isolated genomes, IMG/M includes metagenome sequences generated from an acid mine drainage biofilm, an agricultural soil sample, three isolated deep sea 'whale fall' carcasses, and two enhanced biological phosphorus removing sludge samples.

The data model for the IMG/M data warehouse allows integrating primary genomic sequence information, computationally predicted and curated gene models, precomputed sequence similarity relationships, and functional annotations and pathway information in a coherent biological context. Isolated organisms are identified via their taxonomic lineage (domain, phylum, class, order, family, genus, species, and strain). For each genome, the primary DNA sequence and its organization in scaffolds and/or contigs are recorded. Genomic features,

such as predicted coding sequences and some functional RNAs are also recorded. Protein coding genes are further characterized in terms of molecular function and participation in pathways. Proteins are grouped into families based on sequence similarity. Pathways, reactions, and compounds are included from KEGG and LIGAND. Additional functional annotations according to GO terms are provided by EBI Genome Reviews, while COG provides clusters of orthologous groups of genes (see above). Orthologue and paralogue gene relationships for isolated microbial organisms are computed based on bidirectional best hit (BBH) with clusters formed using a Markov Clustering method. Isolate organisms are characterized in terms of phenotypes (e.g., morphology and geochemistry), ecotype (including geographical coordinates), and disease.

A Phylogenetic Profiler tool allows comparing the gene content between isolate genomes or metagenomes, by defining a profile for the genes in terms of presence or absence of homologues in other entities. Similar to isolate genomes, differences in gene content between metagenomes can be correlated with a specific phenotype or environment, while comparison of the gene content within the metagenome helps inferring the metabolic capabilities of the component populations and identifying the organisms that may be responsible for community-essential tasks.

The Occurrence Profile tools allow examining profiles of genes and functions across metagenomes and isolate organisms. This might give insights of the evolutionary history of the selected gene and may potentially be functionally linked, or co-regulated in a pathway. The Functional Occurrence Profile tools, such as COG Profile, Pfam Profile, and Enzyme Profile, show the occurrence profiles for functional characterizations such as COGs, Pfam families, or enzymes involved in pathways metagenomes and genomes. This tool is especially useful for analysis of datasets obtained from the communities with high species diversity, where little or no sequence assembly can be achieved; for such datasets identification of predominant families allows users to infer habitat-specific biological traits.

IMG/M provides support for the exploration and comparative analysis of metagenomes and their component population in the context of other metagenomes and isolate genomes.

Taxonomy and Phylogeny

UniProt Taxonomy Database

The UniProt taxonomy database integrates taxonomy data compiled in the NCBI database and data specific to the UniProt Knowledgebase. Organisms are classified in a hierarchical tree structure. This taxonomy database contains every node (taxon) of the tree. UniProtKB taxonomy data are manually curated; next to manually verified organism names, it provides a selection of external links, organism strains, and viral host information.

NCBI Taxonomy Database

The NCBI taxonomy database indexes named organisms that are represented in the databases with at least one nucleotide or protein sequence. The Taxonomy Browser can be used to view the taxonomic position or retrieve data from any of the principal Entrez databases for a particular organism or group. Entrez Taxonomy displays include custom taxonomic trees representing user-specified subsets of the full NCBI taxonomy.

Phylogeny Prediction

Phylogeny and molecular evolution have provided a huge toolbox to study the evolutionary history of organisms, allowing inferences of phylogenetic relations (ancestor–descendent) of protein domains, genes, and organisms. The resulting phylogenetic hypotheses are crucial to make phylogeny predictions or inferences. It also allows estimation of evolutionary forces (such as selection, genetic drift, migration, and recombination) in protein domains, genes, genomes, and populations. Phylogeny can make important hypotheses such as the universal tree of life and distinction between orthologues and paralogues as well as the important inferences of their function, for example, the likely change of function between paralogue proteins.

Phylogeny's objective is to trace an ancestor–descendant relation of organisms through different taxonomic levels. The markers used to build a phylogeny are contained in the (DNA or protein) sequences, and these include restriction fragment length polymorphisms (RFLPs), genomic fingerprints, among others. The sequences that contain this information must be aligned with the previously described bioinformatic programs like ClustalW, T-Coffee, and MUSCLE.

In general, DNA sequences give a finer resolution of the evolutionary history of an organism since a great variability exists in the substitution rate within DNA sequences, for example, comparing coding regions and intergenic regions, catalytic residues versus noncatalytic residues, structural domains versus nonstructural domains, third positions versus first and second positions of codons in coding sequences, and stems versus loops of rRNAs and tRNAs. Moreover, different genes evolve at different rates; viral genes evolve very fast in contrast to the slow evolutionary rate of 16S rRNAs.

Horizontal gene transfer (HGT) and homoplasy represent a problem and a limitation of phylogenies. Homoplasy occurs when characters are similar, but are

not derived from a common ancestor. There are several types of homoplasy: parallel evolution, which is the independent evolution to reach the same final state, from the same ancestral state; convergent evolution, which is the independent evolution to reach the same final state, from a different ancestral state; and secondary loss, which is a reversion to the ancestral state.

A phylogenetic tree is a mathematical structure used to represent the evolutionary history among a group of sequences or organisms. Phylogenetic inference requires a precise selection of the method to use from all the available ones, given a set of sequences. The aim of phylogenetic inference is to obtain the best estimate of an evolutionary history based on the incomplete and noisy information contained in the sequences.

One of the most commonly used methods to construct a phylogenetic tree is based on distances between sequences coming from a multiple sequence alignment. The distance values are arranged as a distance matrix whose values depend on the evolutionary model selected and could be used to calculate the tree by the Unweighted Pair Group Method with Arithmetic mean (UPGMA) and Neighbor Joining (NJ) methods. The clustering methods, UPGMA and NJ, reconstruct the tree from a distance matrix. These are very fast methods, but very sensitive to certain parameters such as the order in which Operational Taxonomic Unit (OTUs) are added to the tree. This is because the distance matrix is built pairwise, that is, a distance measure is chosen to quantify the differences between a pair of items. NJ and UPGMA are good only to have a quick idea of how your tree looks, but the resulting tree will not be robust.

Alternatively to distance methods, trees can be constructed using discrete methods such as Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian. These methods consider each site (column) in the alignment directly. The MP and ML methods follow a different optimization criterion, which allows a better selection of the resulting topology from millions of topologies that are to be analyzed. The big limitation of these optimization criteria is that they are computationally costly.

MP assumes an implicit evolutionary model that prefers the resulting phylogeny with the minimum substitutions needed. Parsimony informative sites are those that partition sequences into at least two groups each with at least two members. MP uses the branch and bound optimization algorithm. Exhaustive search and branch and bound methods guarantee finding the best tree. However, exhaustive search methods do not work for anything more than ten sequences on existing single-processor computers. It is important to consider that MP often gives multiple equally parsimonious trees making it difficult to choose among them; it underestimates branch lengths and it does not take account of multiple substitutions at a given site.

ML tree reconstruction is an explicit statistical technique based on the likelihood framework. ML makes several assumptions of substitution models. The most typical are (1) the probability of a change is independent of the prior history of the site (a Markov Model; see above), (2) substitution probabilities do not change with time or over the tree (a homogeneous Markov process), and (3) change is time reversible. All sites are informative because a site that has the same base for two sequences tells us something about the time separating the two molecules.

There are several ML advantages: it is mathematically rigorous and performs well in computer simulations, it takes into account multiple/hidden substitutions, and there is a large support of statistical theory for likelihood estimation and inference and extensions to Bayesian analysis. However, there are disadvantages: ML may be inconsistent if the model of evolution is miss-specified, it is computationally tedious and intensive, and it is not immediately intuitive.

Rooted and Unrooted Trees

A rooted phylogenetic tree is a tree that has a unique node that corresponds to the most recent common ancestor of all the elements in the tree. The strategy that is normally used to root a tree is the inclusion of an outgroup. This outgroup should be close enough to the rest of the sequences in order to infer phylogenetic relationships, but far enough to be a clear outgroup. On the other contrary, an unrooted tree is commonly used to show the relatedness of the leaf nodes without making assumptions about common ancestry. Obtaining a tree is seldom the end of an analysis. It is usually the beginning. There are several statistical tests that one may want to perform concerning the quality of the phylogenetic tree and the data about the process of evolution and about patterns of evolution. Several methods have been proposed that attach numerical values to nodes in trees that are intended to provide some measure of the strength of support for that node. The most popular of these is the bootstrap. Bootstrapping is a modern statistical technique that uses computer-intensive random resampling of data to determine sampling error or confidence intervals for some estimated parameter. In bootstrap phylogenies, characters are resampled with replacement to create many bootstrap replicate datasets. Each bootstrap replicate dataset is analyzed. The agreement among the resulting trees is summarized with a majority-rule consensus tree. The frequencies of occurrence of groups, bootstrap proportions, are a measure of support for those groups. High BPs (e.g., >80%) are indicative of strong support for a particular clade. Usually, 1000 or more bootstrap pseudoreplicates are generated. The bootstrap is totally reliant on the accuracy of the tree-building method. One can get

good bootstrap support for the wrong group if the tree-building method is inappropriate.

Phylogenetic Analysis Algorithms

MrBayes

MrBayes is a program for the Bayesian inference of phylogenies from nucleic acid sequences, protein sequences, and morphological characters. Bayesian inference of phylogeny is based on a quantity called the posterior probability distribution of trees, which is the probability of a tree conditioned on the observations. The conditioning is accomplished using Bayes's theorem. The posterior probability distribution of trees is impossible to calculate analytically; instead, MrBayes uses a simulation technique called Markov chain Monte Carlo (MCMC) to approximate the posterior probabilities of trees.

The output is several files with the parameters that were sampled by the MCMC algorithm. MrBayes can summarize the information in these files for the user. It can also use a hierarchical Bayesian framework to infer sites that are under natural selection. It allows for rate variation among sites and a variety of models of sequence evolution. Testing the resulting tree of MrBayes differs from the other programs. The numbers in the branches are not bootstrap values, but probabilities *a posteriori* for each clade. Since these probabilities are always higher than the bootstrap probabilities, it cannot be considered equally. Moreover, we have to take into account the fact that MCMC starts from a random position, so it will take some time before it reaches the general vicinity of values from the target distribution. This period is called 'burn-in' period, and any nonrepresentative samples taken during this period should be discarded. The easiest way to determine how long to allow for a burn-in is to plot a parameter of interest to determine if it has plateaued. This information is contained in the parameter output file of MrBayes.

*PAUP**

PAUP* originally meant Phylogenetic Analysis Using Parsimony. PAUP* is a major analytical tool in phylogenetic analysis. It makes available a very wide variety of analytical methods in a single environment and can be operated via window/mouse, command-line, or scripts. It includes parsimony, distance matrix, invariants, maximum likelihood methods, and many indices and statistical tests. Unfortunately PAUP* is a commercial program (available from Sinauer Associates), although it is quite a good value.

PHYLIP

PHYLIP (the Phylogeny Inference Package) is one of the most important packages of programs for inferring phylogenies. It is available free over the Internet, and written to work on as many different kinds of computer systems as possible. Methods that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences (protein, DNA, and RNA sequences), gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

Phylogenetic analysis using maximum likelihood

Phylogenetic analysis using maximum likelihood (PAML) is a package of programs for phylogenetic analyses of DNA or protein sequences using Maximum Likelihood. The programs can estimate branch lengths in a phylogenetic tree and parameters in the evolutionary model such as the transition/transversion rate ratio, the gamma parameter for variable substitution rates among sites, rate parameters for different genes, and synonymous and nonsynonymous substitution rates. PAML can also test evolutionary models, calculate substitution rates at particular sites, reconstruct ancestral nucleotide or amino acid sequences, simulate DNA and protein sequence evolution, compute distances based on the synonymous and nonsynonymous changes, and of course do phylogenetic tree reconstruction by ML and Bayesian Markov Chain Monte Carlo methods.

TreeView

TreeView is a program for displaying and manipulating trees. It can draw rooted and unrooted trees, display bootstrap values, and edit trees by moving branches, collapsing them, and rerooting. It provides a simple way to view the contents of a NEXUS, PHYLIP, Hennig86, ClustalW, or other format tree file, and allows the user to create publication quality trees.

Universal Tree of Life

The amount and diversity of species with at least partial sequence information is rapidly increasing and the tree of life is constantly being redrawn. Phylogenetic trees represent a backbone of various other biological studies and it is therefore essential to have the state-of-art tools for their display, customization, and interpretation. In the following section we will describe some of these.

iTOL

Interactive Tree of Life (iTOL) is a web based tool for the display, manipulation, and annotation of phylogenetic trees. Branches can be pruned or collapsed, and any node can be used to reroot the tree. Various types of data, such as genome size or protein domain repertoires, can be mapped onto the tree. iTOL can automatically determine taxonomic classes of all internal nodes and assign proper scientific names to leafs. iTOL is the first visualization tool that supports the display of HGTs. Export to several bitmap and vector graphics formats are supported.

ARB

The ribosomal RNA (rRNA) molecule has been considered the 'gold-standard' for the investigation of the phylogeny and ecology of mircoorganisms. The rapidly increasing number of available rRNA sequences led to the development of ARB. ARB is an integrated package of cooperating software tools for data handling and analysis that fulfils the necessity of rRNA-based identification systems. A central database of processed (aligned) sequences and any type of additional data linked to the respective sequence entries is structured according to phylogeny or other user-defined criteria.

It provides tools for building up databases of RNA sequences, aligning them, and searching, editing, modifying, profiling, and constructing trees. ARB uses its own RNA sequence database, which is a manually curated and quality checked dataset for ribosomal RNA genes. These datasets are maintained in collaboration with the 'arb-silva' project and can be obtained from the ARB Silva database site. For phylogenies, it uses programs from PHYLIP and fastDNAml, as well as its own ARB Neighbor-Joining program. ARB also incorporates a variety of other sequence analysis software. It can handle large numbers of sequences and has sophisticated tree drawing and manipulation. ARB is distributed as executables for a variety of versions of UNIX.

The SILVA system

This is a system implemented to provide a central comprehensive web resource to update quality-controlled databases of aligned rRNA sequences from the Bacteria, Archaea and Eukarya domains. SILVA serves as the main data source for ARB. In addition to the ARB approach, there are currently three projects offering access to a set of curated rRNA sequence and alignment databases: the European rRNA Databank, the Ribosomal Database Project, and the greengenes project. All four projects offer at least one 16S rRNA dataset, but vary in the amount of sequences, quality checks, alignments, and update procedures. However, the ARB project is the only platform that actively incorporates homologous small (SSU) as well as large (LSU) subunit sequences from all three domains of life, the Bacteria, Archaea (16S/23S), and Eukarya (18S/28S).

Resources for the Analysis of Gene Expression

When a microarray is used to assay gene expression, the resulting data must be analyzed to tell which genes are up-regulated, down-regulated or do not present a change in the given experiment. These changes, derived from the fluorescence intensity of each probe in the array, must be translated into numerical values. These measured values are sometimes irreproducible. Moreover, these values must be normalized in order to be compared with the other conditions in the experiment, or with other arrays.

There are three widely used techniques that can be used to normalize gene-expression data from single array hybridization: (1) total intensity normalization, (2) normalization using regression techniques, and (3) normalization using ratio statistics. All of these techniques assume that all (or most) of the genes in the array should have an average expression ratio equal to one. The normalization factor used to adjust the data to compensate for experimental variability and to normalize the fluorescence signals from the two samples being compared.

Once significant signals are obtained, depending on the experiment's objective, the next step is to cluster the signals from genes with similar expression. Various clustering techniques can be applied to the identification of patterns in gene-expression data. Most of the cluster analysis techniques are hierarchical. These differ in the manner in which distances are calculated between the growing clusters and the remaining members of the dataset, including other clusters. Clustering algorithms include, but are not limited to, the following. (1) Single-linkage clustering, which tends to produce clusters that are 'loose' because clusters can be joined if any two members are close together. (2) Complete-linkage clustering, which tends to produce very compact clusters of elements and the clusters are often very similar in size. (3) Average-linkage clustering. There are, in fact, various methods for calculating averages; the most common is the UPGMA. In this method, the two clusters with the lowest average distance are joined together to form a new cluster (see above) called (4) weighted pair-group average, which is identical to UPGMA, except that in the computations, the size of the respective clusters is used as a weight. Hence, this method (rather than UPGMA) should be used when the cluster sizes are suspected to be greatly uneven: (5) within-groups clustering, which is similar to UPGMA except that clusters are merged and a cluster average is used for further calculations rather than the individual cluster elements. This tends to produce tighter clusters than UPGMA: (6) Ward's method, which produces the smallest possible increase in the sum of square errors.

When the microarray datasets are ready to be uploaded in several publicly available databases, it is desirable that they meet certain criteria such as MIAME and preferably a MAGE-TAB format. MIAME describes the Minimum Information About a Microarray Experiment that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. MIAME does not specify a particular format; however, obviously the data are more useful if they are encoded in a way that the essential information specified by MIAME can be

accessed easily. Most of the databases described below support the use of the MAGE-TAB format, which is based on spreadsheets or MAGE-ML. The public repositories ArrayExpress at the EBI (UK), GEO at NCBI (US), Stanford Microarray Database (SMD) at Stanford (US), and CIBEX at DDBJ (Japan) are designed to accept, hold, and distribute MIAME compliant microarray data.

The information contained in these gene expression arrays can serve different purposes in understanding the biology of the selected organism; in particular, they are commonly used for the prediction of regulatory motifs of co-regulated genes. The regulatory regions of these genes are usually used as input to programs such as MEME, oligo-analysis, and RibEx, which were described above. In general, the results obtained from these programs may give insights regarding the discovery of new regulatory sites.

The Stanford Microarray Database

The SMD is a research tool and archive that allows hundreds of researchers worldwide to store, annotate, analyze, and share data generated by microarray technology. SMD supports most major microarray platforms and is MIAME-supportive. The primary mission of SMD is to be a research tool that supports researchers from the point of data generation to data publication and dissemination, but it also provides unrestricted access to analysis tools and public data from 300 publications.

SMD stores gene expression as well as array CGH (comparative genomic hybridization) and chIP-chip (chromatin immunoprecipitation on array) experiments. SMD supports multiple microarray platforms (spotted cDNA or oligonucleotide analysis, Affymetrix, Agilent, Combimatrix, and Nimblegen arrays). It has a data pipeline that can communicate published data directly to ArrayExpress and GEO.

Defining a set of microarrays of interest is the first step for an analysis process. SMD provides two different search forms as effective tools to locate microarrays of interest. Once the microarrays are selected, users decide whether they intend to do an analysis that is applicable to one array at a time or to a group of arrays. Since microarray data are known to be sensitive to several experimental factors, it is important to be able to asses the quality of the data collected from a microarray and to normalize the data appropriately. SMD has several tools that can be used for the assessment of microarray quality. Background correction and normalization methods are often used to correct experimental biases in microarray data, and SMD accordingly provides access to several normalization and background correction methods using the marray and limma packages, respectively, from BioConductor.

Gene Expression Omnibus

The NCBI's Gene Expression Omnibus (GEO) is a data repository and retrieval system for microarray and other forms of high-throughput molecular abundance data generated by the scientific community. In addition to gene expression data, GEO accepts array CGH data, ChIP-chip data, SNP array data, and some proteomic data types. The GEO repository accepts MIAME-compliant data submissions.

E. coli GenExpDB

The University of Oklahoma Bioinformatics Core hosts the *E. coli* Gene Expression Database. *E. coli* GenExpDB was designed from the biologist's perspective, with a tool box-style, integrated and interactive display interface and simplified warehouse architecture; it has a wide variety of microarrays for different conditions.

Expression-Ring displays gene expression ratio data in a blue/yellow heat map of concentric rings, one for each experiment, with all displayed Transcription Factor (TF) genes labeled on outside ring(s) and with hyperbolas connecting to all target genes regulated by each TF.

ArrayExpress

ArrayExpress is a public database for high-throughput functional genomics data. ArrayExpress consists of two parts: the ArrayExpress Repository, which is a MIAME supportive public archive of microarray data, and the ArrayExpress Data Warehouse, which is a database of gene expression profiles selected from the repository and consistently reannotated. Archived experiments can be queried by experiment attributes, such as keywords, species, array platform, authors, journals, or accession numbers. Gene expression profiles can be queried by gene names and properties, such as GO terms, and gene expression profiles can be visualized.

Centre for Information Biology Gene Expression Database

The gene expression database Centre for Information Biology Gene Expression Database (CIBEX), with a data retrieval system in compliance with MIAME, a standard that the MGED Society has developed for comparing data produced in microarray experiments in different laboratories worldwide. CIBEX serves as a public repository for a wide range of high-throughput experimental data in gene expression research, including microarray-based experiments measuring mRNA, serial analysis of gene expression (SAGE tags), and mass spectrometry proteomic data.

Resources for Proteomics

Swiss-2DPAGE

Swiss-2DPAGE database stores data from two-dimensional polyacrylamide gel electrophoresis including reference maps, images, mapping procedures, and cross-references. Data can be searched by protein name, protein accession number, by experimental pI/MW range, and by clicking on the spot.

Open Mass Spectrometry Search Algorithm

The NCBI's Open Mass Spectrometry Search Algorithm (OMSSA) is an efficient search engine for identifying tandem MS (MS/MS) peptide spectra by searching libraries of known protein sequences. It allows up to 2000 spectra to be analyzed in a single session using either BLAST 'nr' or refseq_protein sequence libraries for comparison. Standalone versions of OMSSA for several popular computer platforms that accept larger batches of spectra and allows searches of custom sequence libraries are available.

Mascot: A Search Engine that Uses Mass Spectrometry Data to Identify Proteins from Primary Sequence Databases

Additionally to OMSSA, there is another search engine called Mascot. This tool uses mass spectrometry data to identify proteins from primary sequence databases. Mascot is unique in that it integrates all of the proven methods of searching: (1) peptide mass fingerprint, in which the only experimental data are peptide mass values; (2) sequence query, in which peptide mass data are combined with amino acid sequence and composition information; (3) MS/MS ion search, which uses uninterpreted MS/MS data from one or more peptides.

The Proteomics Identification Database

Proteomics Identification Database (PRIDE) is a centralized, standards compliant, public data repository for proteomics data. This database contains information from experiments, identified proteins, identified peptides, unique peptides, and spectra. PRIDE offers a web-based query interface, a user-friendly data upload facility, and a documented application programming interface for direct computational access. It supports identification from both MS-based and gel-based techniques. Processed peak list arising from MS, MS/MS, and higher MS levels are supported. PRIDE retrieves the complete set of protein identifications for a publication, along with the supporting peptide identifications and hyperlinks to further information. PRIDE also finds all relevant datasets for a particular protein of interest.

The Open Proteomics Database

Open Proteomics Database (OPD) stores and disseminates MS-based proteomics data. The data residing in OPD represent diverse proteomics samples – some interpreted, some uninterpreted, some on simple but defined samples to be used for training algorithms, and some on highly complex samples, such as whole-cell lysates from different organisms. In all, proteomics data from *E. coli*, *Mycobacterium smegmatis*, *S. cerevisiae*, and *Homo sapiens* are represented with roughly 400 000 total mass spectra, cataloguing the expression of several thousand proteins overall. All data are freely accessible with the intent that computational groups interested in studying the many computational problems posed by proteomics will have a source of protein mass spectra and expression data.

Resources for Gene Regulation Analysis

As was reviewed in 'Posttranscriptional regulation' and 'Transcriptional regulation', regulation of gene expression can occur at multiple levels, transcription initiation being the most common way of regulation, but also can take place at the posttranscriptional level. In any case, the main regulatory elements are localized upstream of operons; however, regulatory sites are sometimes found inside or at the end of the transcription unit. Our current knowledge of different genes, operons, and regulatory mechanisms is quite variable. For a few model organisms, such as *E. coli* or *Bacillus subtilis*, the regulatory mechanisms are very well characterized for the vast majority of their genes; whereas for the majority of other organisms, their regulatory elements have been poorly characterized or not characterized at all. Therefore, regulatory databases of model organisms are of great value to infer the gene regulation in other phylogenetically related organisms. In any case, these kinds of databases constitute a useful source to guide and design experimental work.

RegulonDB: A Database for Transcriptional Regulation in *E. coli*

RegulonDB is the internationally recognized reference database of *E. coli* K-12 offering curated knowledge of the regulatory network and operon organization. It is currently the largest electronically encoded database of the regulatory network of any free-living organism. It is a model of the complex regulation of transcription initiation or regulatory network of the cell, as well as a model of the organization of the genes in transcription units, operons, and simple and complex regulons. Continuous

curation of the original scientific literature provides the evidence behind every single object and feature. This knowledge is complemented by comprehensive computational predictions across the complete genome. Literature-based and predicted data are clearly distinguished in the database. RegulonDB public releases are synchronized with those of EcoCyc, since RegulonDB's curation supports both databases. The complex biology of regulation is simplified in a navigation scheme based on three major streams: genes, operons, and regulons. Regulatory knowledge is directly available in every navigation step. Displays combine graphic and textual information and are organized allowing different levels of detail and biological context. This knowledge is the backbone of an integrated system for the graphic display of the network, graphic, and tabular microarray comparisons with curated and predicted objects, as well as predictions across bacterial genomes and predicted networks of functionally related gene products.

With RegulonDB, the user can get mechanistic information about the different transcription units and their regulatory elements, such as promoters and their sigma factor types, genes and their ribosome binding sites, terminators, binding site of specific transcriptional regulator (TRs) as well as their organization into regulatory phrases, active and inactive conformations of TRs and regulons simple and complex.

DBTBS: A Database for Transcriptional Regulation in *B. subtilis*

The counterpart of RegulonDB is DBTBS (Database of Transcriptional Regulation in *B. subtilis*), a reference database of transcriptional regulation in the other model organism, *B. subtilis*, summarizing the experimentally characterized transcription factors, their recognition sequences, and the genes they regulate. The goal of this database is to help elucidate its complete gene regulatory network. The construction of the DBTBS aims to compare the results of systematic experiments with the rich source of individual experimental results accumulated so far. The DBTBS database contains a collection of experimentally validated gene regulatory relations and the corresponding transcription factor binding sites upstream of *B. subtilis* genes as well as experimentally validated *B. subtilis* operons and their terminators. Its current version is constructed by surveying the scientific literature and contains the information of binding factors and gene regulatory relations. For each promoter, all of its known *cis*-elements are listed according to their positions, while these *cis*-elements are aligned to illustrate the consensus sequence for each transcription factor. All probable transcription factors coded in the genome are classified using Pfam motifs.

Given the increase in the number of fully sequenced bacterial genomes, DBTBS has extended its usability in comparative regulatory genomics. A new section on the conservation of the upstream regulatory sequences among homologous genes in 40 Gram-positive bacterial species as well as on the presence of overrepresented hexameric motifs that may have regulatory functions was created.

Predictive Web Pages on Gene Regulation

In addition to the two aforementioned regulatory databases that compile data from molecular biology experimental work, there are some other databases and web pages dealing with the *in silico* prediction of regulatory elements. Although the accuracy of the computer predictions is obviously not as solid as the data coming from the experimental analysis, predictive analysis is very important for the design of working hypotheses. Some examples of predictive regulatory databases and web sites are as follows.

Neural Network Promoter Prediction

Computer prediction of promoter elements is one of the most commonly used analyses of experimental scientists. Neural Network Promoter Prediction is a web page that predicts both prokaryote and eukaryote promoters based on neural networks that have been trained to recognize promoter elements using a pruning iterative procedure that deletes those weights in the network that add the lowest predictive value to the overall promoter prediction. This pruned neural network gives clues about the importance of specific positions in the different types of promoter elements by studying their relative weights.

WebSIDD

WebSIDD is a web-based service designed to predict locations and extents of the stress-induced duplex destabilization (SIDD) that occur in a double-stranded DNA sequence, on which a specified level of super-helical stress has been imposed. The algorithm calculates the approximate equilibrium statistical mechanical distribution of a population of identical molecules among the accessible states. Its output is the calculated transition probability and destabilization energy of each base pair in the sequence. The structural and energy parameters used in the calculation are all determined experimentally. This method has illuminated the roles of SIDD properties in the regulation of diverse biological processes, such as transcription initiation (promoter prediction) and termination. The prediction of promoter sequences is much more accurate if it considers the prediction of SIDD and sequence-dependent motifs finder algorithms are taken simultaneously.

Regulatory sequence analysis tools

This site offers a series of tools dedicated to the detection of regulatory signals in noncoding sequences that are grouped into the following main blocks of analysis: (1) sequence retrieval, (2) regulatory pattern discovery (string-based pattern discovery, matrix-based pattern discovery; see above, 'Oligo-analysis'), and (3) regulatory pattern matching that includes the genome-scale pattern matching to scan entire genomes for genes having a particular regulatory motif. Interestingly, this web site offers a computer application to predict regulatory motifs from clusters of co-expressed genes based on microarray data. Each one of the Regulatory Sequence Analysis Tools (RSA Tools) is presented as a form to fill. For each form, a manual page provides detailed information about the parameters. The RSA Tools web site offers a set of clear tutorials to get familiarized with their tools.

Predicted transcription attenuation in bacteria

Gene regulation by transcription termination–antitermination, often called transcription attenuation, is a strategy commonly used by bacteria to sense a specific metabolic signal and enables a response that directs RNA polymerase to either terminate transcription or transcribe the downstream genes of an operon (for a review of these mechanisms see 'Posttranscriptional Regulation').

The decision whether to terminate transcription is often based on the selective arrangement of one of the two mutually exclusive RNA secondary structures in the nascent transcript, the antiterminator and the terminator. Transcription attenuation web page compiles a computer-based predict transcription attenuators for fully sequence genomes. The computer predictions are based on the search of potential alternative RNA-hairpin structures in the leader sequence that precedes a particular gene or operon. The predicted transcription attenuators in this database are clustered by organisms or by COG classification.

RibEx: a web server for the prediction of riboswitches and other conserved regulatory elements

This web tool clusters the intergenic region of orthologous genes by an iterative process; conserved motifs across phylogenetically distant organisms are identified. These motifs correspond to reported riboswitches and other likely regulatory systems that appear to depend on conserved RNA structures. A riboswitch is a part of an mRNA leader sequence capable of binding, with great specificity and affinity, a signal molecule without the intervention of any protein factor. One part of the riboswitch can fold to form either of two alternative hairpin structures, one of which functions as an intrinsic transcriptional terminator or a secondary structure that blocks gene translation. The binding of the riboswitch to the metabolite controls the downstream gene expression by selecting between these two alternative conformations. RibEx allows the visual inspection of these conserved motifs and riboswitches in any sequence given by the user.

GeCont: a web server to analyze the genome context of orthologous genes

In bacteria, the coordinate transcription of functionally related genes that belong to the same pathway or process is commonly accomplished by the operon structure. Based on this property, gene function can sometimes be inferred by the inspection of the function of its neighboring genes. This idea is particularly true if the gene context is conserved among many other genomes. GeCont is a web server designed to show the genomic context of a particular gene and their orthologous counterparts, based on COG classification, in the set of fully sequenced organisms.

Public Available Software for the Analysis of Gene Regulation

Consensus

In addition to the aforementioned MEME/MAST and HMMER programs that identifies and searches for conserved motifs in a set of given sequences, the Consensus program has been used to identify regulatory sequences. The operating principle of Consensus assumes that regulatory motifs can be represented by weight matrices. For this purpose, Consensus uses a greedy algorithm that searches for the matrix with maximum information content. It first finds the pair of sequences that share a motif with greatest information content, then finds a third sequence that can be added to the previously identified motif resulting in the motif with the greatest information content, and so on.

MBDBs and Analysis Programs of RNA Regulatory Elements

During the past few years, new roles of RNA molecules in the control of RNA expression have been elucidated. As a result of the continuous efforts done in this field, important databases have been created. For instance, The Wellcome Trust Sanger Institute in collaboration with Janelia Farm have created the Rfam database, which is a large collection of multiple sequence alignments and covariance models covering many common noncoding RNA families.

Rfam and INFERNAL: A Database for RNA Families and Analysis Software

Rfam aims to facilitate the identification and classification of new members of known sequence families, and

distributes annotation of noncoding RNAs (ncRNAs) in over 200 complete genome sequences. A small number of families are essential in all three kingdoms of life with large numbers of smaller families specific for certain taxa. The Rfam database is, thus, a comprehensive collection of ncRNAs, whose products are components of some of the most important cellular machineries, such as the ribosome, the spliceosome, and the telomerase. The known repertoire of ncRNA cellular functions is expanding rapidly. Ribozymes catalyze a range of reactions, such as self-cleavage of hepatitis delta virus transcripts and 5′ maturation of tRNAs by the ubiquitous RNAse P. Riboswitches are in *cis*-regulatory sequences capable of regulating gene expression by directly sensing a metabolite without the intervention of a protein. Examples of some riboswitches and other RNA regulatory elements are described in Posttranscriptional Regulation.

Like Pfam for protein-coding genes, ncRNA sequences can be grouped into families, and much can be learnt about structure and function from multiple sequence alignments of such families. Unlike proteins, ncRNAs often conserve a base-paired secondary structure with low primary sequence similarity. The combined secondary structure and primary sequence profile of a MSA of ncRNAs can be captured by statistical models, called profile stochastic context-free grammars (SCFGs), analogous to profile HMMs of protein alignments.

This database comprises a covariance model for each RNA family, represented by a MSA and profile SCFGs available through its web page. Each family and its model or profile can be downloaded, and then, in conjunction with the INFERNAL software, it is possible to search for any one of the RNA families in a particular sequence or genome.

INFERNAL (Inference of RNA Alignment) is an implementation of a special case of profile SCFGs called covariance models. A CM is like a sequence profile, but it scores a combination of sequence consensus and RNA secondary structure consensus; so in many cases, it is more capable of identifying RNA homologues that conserve their secondary structure more than their primary sequence.

The INFERNAL software package makes consensus RNA secondary structure profiles and uses them to create new structure-based multiple sequence alignments. To make a profile, you need to have a multiple sequence alignment of an RNA sequence family, and the alignment must be annotated with a consensus RNA secondary structure. The program cmbuild takes an annotated multiple alignment as input, and outputs a profile. A profile of a known model, for instance the T box riboswitch, can be downloaded from the Rfam database. Once a profile is obtained (either built from the user's sequences or downloaded from Rfam), it can be used to search for homologue sequences in a database using the program cmsearch. The profile can also be used by the program cmalign to align a set of unaligned sequences. The previous procedure allows the user to build a hand-curated representative alignment of RNA sequence family, and then use this profile to automatically align any number of sequences to that seed profile. This is the strategy used to maintain the Rfam database of RNA multiple alignments and profiles.

Prediction of Secondary Structure

The development of computational tools for the interconnection of sequence and structural information to annotate and discover ncRNAs faces two main limitations: the requirements in computational resources and our understanding of RNA structural and evolutionary rules. As the secondary structure is the main energetic component of RNA architecture, it produces strong constraints for the tertiary structure and its definition constitutes a first and essential step.

Mfold and RNAfold from the Vienna package predict an RNA structure which guarantees a convergence to the minimal free-energy structure. In contrast, HotKnots does not guarantee a convergence to the minimal free-energy structure, its algorithm is based on the widely used free-energy minimization, and by using a heuristic approximation and an extended free-energy model is able to compute the predicted RNA structure, including pseudoknots, in a reasonable amount of time. Alternatively, the RNAMST web server searches for possible RNA structures, with previously user-defined constrains, against several databases, such as Rfam. cmfinder can also be used to search structural homologues in databases.

Multiple alignments of RNA sequences are rarely available. Several programs can be used to achieve this task. Comparative analysis of a family of homologous sequences enables derivation of covariations within an RNA family and thus the identification of the locations of regular helices. However, the identification of the conserved core of the secondary structure within a MSA is a difficult and iterative task, usually performed by hand using human expertise. Many tools enable realization of this structural alignment automatically. Some of these tools predict secondary structure for each individual sequence and perform an alignment of these derived structures, such as SCARNA, STRAL, and MARNA. The main drawback is that folding of any single RNA sequence might not produce the biologically active RNA structure. Other tools propose to predict the secondary structure and align RNA sequences simultaneously. These are called Sankoff-like methods, and FOLDALIGN, STEMLOC, and Dynalign are examples of this. Most of these programs limit their computational costs by doing a pairwise alignment (SCARNA, STEMLOC, FOLDALIGN, and Dynalign), while others (STRAL) make use of heuristic methods or limit sequence length (MARNA). Another strategy is to align unfolded RNA

sequences according to a reference molecule with a secondary structure that has been well described using HMMs, which is used by the program PSTAG.

Hence, if a multiple alignment of RNA sequences is available, Mifold proposes a Matlab package to exploit mutual information measure in order to identify covering sites. RNAalifold, from the Vienna package and for which a web interface is available, predicts the consensus secondary structure for the RNA alignment by minimizing the overall free energy and uses information on compensatory mutations between the sequences to improve secondary structure predictions. An exception to this is RNAcast, now integrated into the RNAshapes analysis package, which can predict an RNA consensus structure from a set of unaligned sequences. RNAcast uses the concept of RNA families, which share a common structure leading to searching in a reduced space and hence decreasing the computing requirements. Finally, KNetFold is a machine learning approach, which measures the distance between each column from an alignment compared to the Rfam alignments. Hence, this program can estimate which columns are paired. The previously described cmfinder uses the Vienna RNA package, and takes as input a set of unaligned sequences and searches the most conserved secondary structures whose length and number of stem-loops are within a user-defined range. The selected subset can be used to construct an initial multiple alignment, which is improved using the probabilistic framework provided by COVE. RNAmine searches for a frequently appearing pattern of stems. This defined pattern of stems can be searched between multiple RNA families. RNAmine, as well as cmfinder, can find a conserved structure in a subset of input sequences.

As previously described for computational approaches for overrepresented sequence motifs, such as MEME and MAST, cyclic processes also apply for the prediction of secondary structures and for finding structural homologues. Once a structural signature is available for a given RNA family, the next logical step is to make a refinement and expansion of the ncRNA family using a genome-scale homology search. An example of this can be done with cmbuild-cmfinder cycles, where from a set of input sequences an RNA profile can be constructed and then be used to search within a given database. The new sequences found can be used for a new set of input sequences for cmbuild, and so on.

The computational tools developed to discover secondary structure from sequence and to discover structural information from RNA families face two main limitations: the requirements in computational resources and our understanding of the RNA structural and evolutionary rules. Among the most promising initiatives in this field is the creation of a consortium dedicated to the construction of a common, dynamic, and controlled vocabulary (or ontology) that should capture all the RNA concepts and their relations, inspired in the previously described GO.

Dedicated Integration Systems for Molecular Biology Databases

As could be observed in this review, MBDBs are very heterogeneous and their distribution is widespread. Important efforts have been taken for the integration of Biological Databases such as Entrez, DBGET, and more importantly SRS that is an indexing and retrieval tool for flat file data libraries, such as the EMBL, SwissProt, or PROSITE (see above). For this porpoise, SRS has developed a special language called ODD that recognizes the different library formats and organizations and extracts other data needed during retrieval. SRS has a friendly web interface that allows easy inspection of retrieval of the entries.

A second example of a dedicated integrative system is called STRING aimed at predicting direct (physical) and indirect (functional) protein–protein interactions based on the quantitative integration of data coming from (1) genomic context, (2) high-throughput experiments, (3) gene coexpression, and (4) reported knowledge. These predictions are very important to relate the molecular functions of individual proteins in a more general and integrated context.

Future of Biological Databases

The exponential growth of most of the aforementioned databases makes clear that their size in the near future will become an important problem. Specialized software dedicated to the maintenance and efficient updating of the information will be required. In addition, the highly diverse types of data that biologists require, such as microarray expression data, metabolic, and protein–protein interaction networks or protein structure, just to mention some of them, creates a crucial challenge to combine and integrate all of the MBDBs by cross-references, with a similar aim as the aforementioned SRS database. Furthermore, in order to fully exploit these database resources, data mining and new knowledge discovery algorithms will need to be developed.

MBDBs are highly diverse, although at some point, most of them are interconnected. Their use depends on the kind of question that has to be solved, and almost always there is more than one pathway to conduct a study. In any case, a DNA sequence is commonly one of the simplest units of information that a user might have as a starting point. Regarding its nature, the relevance of the sequence is commonly considered as a coding region, Open Reading Frame (ORF), or as a DNA/RNA regulatory element. This is the main branch point for many studies.

In the case of coding sequences, a common objective is the protein function assignment. To this end, the DNA sequence is initially translated and used as a query to perform either pairwise searches using BLAST or motif based searches using HMM algorithms such as HMMER. In any case, the common aim of a search is the identification of homologue counterparts, since they might share a similar biological/biochemical function. In the first case, the BLAST search can be done against large databases (e.g., GenBank) or specialized databases (e.g., Swiss-Prot or PDB). In addition to this kind of search, the user might ask if the sequence has conserved motifs previously identified (e.g., protein families). This is particularly important to identify distantly related proteins that might not present evident similarity across their entire sequences, but in smaller discrete regions. These conserved motifs usually correspond to catalytic or prosthetic sites, binding domains, or structure determinants. The user can identify these motifs using the web services of InterPro that incorporates the major protein signature databases, such as PROSITE, PRINTS, ProDom, Pfam, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, and Gene3D, among others. Interestingly, more than one motif can be considered in a database search using HMM programs such as HMMER.

For the analysis of noncoding sequences, such as regulatory elements, a similar strategy can be followed, since the principles of these database searches are the same. This analysis involves specialized databases, web services, and computer programs that consider the DNA/RNA nature of the sequence, for example, the Watson–Crick kind of interactions between nucleic acid molecules. In the case of regulatory elements, it is important to consider that gene regulation in bacteria mainly takes place at transcription initiation or at posttranscriptional events by RNA elements located in the 5′ region of the transcription units. Therefore, a common approach to identify potential regulatory sites or elements considers the analysis of over-represented motifs in a set of 5′ upstream regions of co-expressed genes, commonly identified by microarray experiments. The resulting motifs can be compared against known regulatory sites in databases of model organisms such as *E. coli* (RegulonDB) or *B. subtilis* (DBTBS) or against previously identified RNA regulatory elements described in databases such as Rfam or RibEx.

Regardless of the kind of analysis selected, an issue that the user should keep in mind is that MBDBs are important sources of information to formulate or verify scientific hypotheses. Searching for related data on previously reported scientific literature in databases such as Entrez or PubMed may give new insights to the working project. Incorporating this knowledge into the user's results produces a more complete dataset that can be used as an input to restart the process. The new dataset may be refined in comparison with the previous one, generating, with each cycle, a better and more comprehensive scientific model.

See also: Metagenomics; Posttranscriptional Regulation; Transcriptional Regulation

Further Reading

Altschul SF, Madden TL, Schaffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402.

Bailey TL and Gribskov M (1998) Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics* 14: 48–54.

Baxevanis AD and Ouellette BFF (2001) *Bioinformatics*. United States of America: Wiley-Interscience.

DeLong EF and Karl DM (2005) Genomic perspectives in microbial oceanography. *Nature* 437: 336–342.

Eddy SR (2004) How do RNA folding algorithms work? *Nature Biotechnology* 22: 1457–1458.

Eddy SR (2004) What is a hidden Markov model? *Nature Biotechnology* 22: 1315–1316.

Eddy SR (2004) What is Bayesian statistics? *Nature Biotechnology* 22: 1177–1178.

Finn RD, Tate J, Mistry J, *et al*. 2008. The Pfam protein families database. *Nucleic Acids Research* 36: D281–D288.

Kopp J and Schwede T (2006) The SWISS-MODEL repository: New features and functionalities. *Nucleic Acids Research* 34: D315–D318.

Kouranov A, Xie L, de la CJ, *et al.* (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Research* 34: D302–D305.

Larkin MA, Blackshields G, Brown NP, *et al*. (2007) Clustal W and Clustal X version 2.0. Bioinformatics.

Lee D, Redfern O, and Orengo C (2007) Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology* 8: 995–1005.

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, and Bork P (2006) SMART 5: Domains in the context of genomes and networks. *Nucleic Acids Research* 34: D257–D260.

Mulder NJ, Apweiler R, Attwood TK, *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Research* 33: D201–D205.

Petsko GA and Ringe D (2004) *Protein Structure and Function*. London, UK: Sinauer Associates, Incorporated.

Tompa M, Li N, Bailey TL, *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology* 23: 137–144.

Wheeler DL, Barrett T, and Benson DA (2007) Database resources of the national center for biotechnology information. *Nucleic Acids Research* 35: D5–D12.

Whitfield EJ, Pruess M, and Apweiler R (2006) Bioinformatics database infrastructure for biotechnology research. *Journal of Biotechnology* 124: 629–639.