CHAPTER 10

Genome Analysis

INTRODUCTION

A MAJOR APPLICATION OF BIOINFORMATICS IS analysis of the full genomes of organisms that have been sequenced starting in the late 1990s, including microbial genomes, the budding yeast *Saccharomyces cerevisiae*, the nematode worm *Caenorhabditis elegans*, the plant *Arabidopsis thaliana*, the fruit fly *Drosophila*, and the human genome. Many additional genome sequencing projects are either being planned or are already under way.

The genome is defined as the sum of the genes and intergenic sequences of the haploid cell (Bernardi 1995).

Traditional genetics and molecular biology have been directed toward understanding the role of a particular gene or protein in an important biological process. A gene is sequenced to predict its function or to manipulate its activity or expression. In contrast, the availability of genome sequences provides the sequences of all the genes of an organism so that important genes influencing metabolism, cellular differentiation and development, and disease processes in animals and plants, can be identified and the relevant genes manipulated.

The challenge is to identify those genes that are predicted to have a particular biological function and then to design experiments to test that prediction. This analysis depends on gene prediction using gene models for each organism followed by sequence comparisons between the predicted proteins with other proteins whose function is known from biological studies. To facilitate such comparisons, the genomes of a number of model organisms about which a great deal of biological information is available have been sequenced. Many years of genetic and biochemical research of these model organisms—the bacterium *Escherichia coli*, *S. cerevisiae*, *C. elegans*, *A. thaliana*, and *D. melanogaster*—have led to the accumulation of a large amount of information on gene organization and function. The mouse *Mus musculus* is a genetic model for humans because the two species are so closely related through evolution. A newly identified gene in another organism can be compared to the existing database of information to find whether it has a similar function. Genes involved in human disease, for example, are sometimes found to be similar to a fruit fly gene at the protein sequence level (for an example of how significant this kind of analysis can be, see Rubin et al. 2000). The genetic effects of mutations in the fruit fly's gene will then provide a biochemical, cellular, or developmental model for the human disease. Interestingly, it has not been possible to identify the function of all the genes in model organisms. As a result, a similar gene or family of genes may be found in several organisms, including a model organism, but the function is not known because the gene functions have not yet been analyzed. Hence, continued biological analysis of model organisms in those areas that are not tractable by the tools of bioinformatics has many important applications.

Tracing the phylogenetic history of such uncharacterized genes, characterized genes, and gene domains and gene linkages in diverse organisms is one of the most interesting and challenging aspects of genome analysis. In addition, even though a gene that specifies an important biological function has not been identified, the gene can be traced in individuals using sequence variations that occur among individuals in a population, called sequence polymorphisms. In humans, for example, single nucleotide polymorphisms (SNPs) can be found throughout the genome, including some that are positioned adjacent to an important disease gene. If a particular G → A polymorphism is right next to a defective tumor suppressor gene, for example, that polymorphism serves as a genetic marker for the presence of the defective gene. The applicable genetic principle, genetic linkage, is that closely linked genes seldom become separated by genetic recombination from one generation to the next. Another example of such linked polymorphisms is in crop plants. Features such as plant height and amount of seed produced are influenced by variations in sets of genes, called quantitative trait loci (QTL). Inheritance of QTLs can be traced from one

generation to the next using sequence polymorphisms that are linked to the favored genetic variation without having to wait to observe the effects on plant growth.

The availability of genome sequences greatly facilitates the discovery and utilization of these sequence polymorphisms. It is recognized that some types of genetic variation, including specific human diseases, are best understood at the genome-wide level. The duplication of genes, gene segments, and gene clusters provides opportunities for recombination events that can cause changes in gene copy number or loss of gene function (Lupski 1998).

In summary, the availability of genome sequences provides an unprecedented opportunity to explore genetic variability both between organisms and within the individual organism. We now turn to a comparison of the main features of the genomes that have been sequenced. One major task is to identify the genes that encode proteins and to identify the function of as many of these proteins as possible by database similarity searches.

The entire set of proteins of an organism, including those known from biological studies and those predicted by bioinformatics, is the proteome of the organism.

The proteome may be compared to itself to identify paralogs, families of proteins that have arisen by gene duplication. One proteome may also be compared to another proteome to discover orthologous genes that have kept the same function, genes that have become fused to make a larger protein (or split into two to make two separate proteins), new arrangements of protein domains, and amplification of protein families to perform a new type of biological function (e.g., cell-to-cell communication during development of a multicellular organism). A representative collection of the large number of Web resource pages and references is shown in Table 10.1. This table is divided into six parts, A–F, dealing with resources for prokaryotic genomes (A) which have been the subject of intense sequence analysis, all model organisms (B), human genome and the related mouse genome (C), genome relationships (D), proteome and gene expression analysis (E), and functional characterization of genes (F). Since these sites are constantly being revised, this table will be periodically updated on the book Web site.

GENOME ANATOMY

Early biologists examining a particular plant, animal, or yeast cell using a microscope observed a nucleus (in a eukaryotic cell) with a specific number of chromosomes of variable length and morphology that could be seen at certain stages of cell division. The chromosomes comprised linear DNA molecules in a tightly compact form that was wrapped around protein complexes, called the nucleosome. Nuclei and chromosomes were not observed in bacteria (a prokaryotic cell), but when bacterial DNA was eventually detected, the molecule was usually circular and was also in a compacted form. The following sections outline the structure and composition of prokaryotic and eukaryotic genomes.

Prokaryotic Genomes

The first bacterial genome to be sequenced was that of *Hemophilus influenzae*, a mild human pathogen (Fleischmann et al. 1995). This project was carried out at the Institute of Genomics Research (TIGR, http://www.tigr.org) in part to prove a new genome sequencing method—the shotgun method. A large number of random overlapping fragments were sequenced and then a consensus sequence of the entire 1.8×16^6-bp chromosome of *Hemophilus* was assembled by computer, excepting several regions that had to be assembled manually. Once available, open reading frames were identified, and these were compared to the existing proteins by a database similarity search (see Chapter 7). Approximately 58% of the 1743 predicted genes matched genes of another species, the bacterial

Table 10.1. *Web resources and references for genome information and analysis*

A. Prokaryotic genomes[a]	
MAGPIE: Multipurpose Automated Genome Project Investigation Environment (Gaasterland and Sensen 1998)	http://genomes.rockefeller.edu/magpie
Microbial genome databases	http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/micr.html http://www.techfak.uni-bielefeld.de/techfak/persons/chrisb/ResTools/biotools/biotools10.html http://www-nbrf.georgetown.edu/pir/genome.html#PROK
Comparative genome analysis in P. Bork laboratory (see Web site)	http://www.bork.embl-heidelberg.de/Genome/
TIGR: The Comprehensive Microbial Resource Home Page—the omniome	http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl
U.S. Dept. of Energy Joint Genome Initiative	http://www.jgi.doe.gov/

[a] Also see the COG and PEDANT sites in part D.

B. Genomic databases of model organisms and other genome databases	
Arabidopsis thaliana genome displayer	http://www.kazusa.or.jp/kaos
A. thaliana information resource TAIR	http://www.arabidopsis.org/
Caenorhabditis elegans (worm) database	http://www.wormbase.org/
C. elegans chromosomes	ftp://ftp.sanger.ac.uk/pub/databases/C.elegans_sequences/CHROMOSOMES/
C. elegans genome project	http://www.sanger.ac.uk/Projects/C_elegans/
C. elegans proteome database	http://www.sanger.ac.uk/Projects/C_elegans/wormpep/ http://www.proteome.com/YPDhome.html
Dictyostelium discoideum genome information	http://www.biology.ucsd.edu/others/dsmith/dictydb.html
Drosophila melanogaster Berkeley Drosophila genome project	http://www.fruitfly.org/
D. melanogaster chromosomes	http://flybase.bio.indiana.edu/maps/fbgrmap.html
D. melanogaster: Flybase, a genomic database	http://flybase.bio.indiana.edu/
E. coli genome project	http://www.genetics.wisc.edu/
E. coli genome and proteome database GenProtEC	http://genprotec.mbl.edu/
E. coli index	http://web.bham.ac.uk/bcm4ght6/res.html
Genome databases at NCBI[a]	http://www.ncbi.nlm.nih.gov/Genomes/index.html http://www.ncbi.nlm.nih.gov/Entrez/Genome/main_genomes.html http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/org.html
Genome databases other than NCBI[a]	http://www.techfak.uni-bielefeld.de/techfak/persons/chrisb/ResTools/biotools/biotools10.html http://www-nbrf.georgetown.edu/pir/genome.html
Genome list at NIH	http://molbio.info.nih.gov/molbio/db.html
Mitochondrial DNA Database MitBASE	http://www3.ebi.ac.uk/Research/Mitbase/mitbase.pl
Mouse (*Mus musculus*) genome informatics	http://www.informatics.jax.org/
Plant genome projects supported by the plant genome initiative of the U.S. National Science Foundation	http://www.nsf.gov/bio/dbi/pgrsites.htm
Organelle genome sequences	http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/organelles.html http://www-nbrf.georgetown.edu/pir/genome.html
Parasite genome databases and genome research resources	http://www.ebi.ac.uk/parasites/parasite-genome.html
Retroviral genotyping and analysis site	http://www.ncbi.nlm.nih.gov/retroviruses/
Rice (*Oryza sativa*) genome project	http://rgp.dna.affrc.go.jp/
Saccharomyces cerevisiae: View of 16 chromosomes	http://genome-www.stanford.edu/Saccharomyces/MAP/GENOMICVIEW/GenomicView.html
S. cerevisiae, YPD Yeast Proteome database, a commercial database	http://www.proteome.com/YPDhome.html
S. cerevisiae (budding yeast) database SGD	http://genome-www.stanford.edu/Saccharomyces/

[a] The National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland

Continued.

Table 10.1. *Continued*

C. Human and mouse genome comparisons	
Celera Genomics: The company that assembles genome sequences by automated fragment assembly	http://www.celera.com/
Comparison of human (*Homo sapiens*) and mouse (*M. musculus*) chromosomes	http://www.bioscience.org/urllists/chromos.htm, http://www.ncbi.nlm.nih.gov/Homology/ http://infosrv1.ctd.ornl.gov/TechResources/Human_Genome/publicat/97pr/05g_mous.html http://srs.ebi.ac.uk/, databanks link, MOUSE2HUMAN
Cooperative Human Linkage Center: mouse-clickable map of chromosomes	http://lpg.nci.nih.gov/html-chlc/ChlcIntegratedMaps.html
Draft Human Genome Browser	http://genome.ucsc.edu/goldenPath/hgTracks.html
Human sequence polymorphisms, mutations, and mapping	http://srs.ebi.ac.uk/, databanks link
Human EST project	http://genome.wustl.edu/est/esthmpg.html
Human genome resources at NCBI	http://www.ncbi.nlm.nih.gov/genome/guide/
Human genome research sites provided by Oak Ridge National Labs	http://www.ornl.gov/hgmis/centers.html
Mouse (*M. musculus*) chromosomes: mouse-clickable map	http://brise.ujf-grenoble.fr/~mongelar/clickclientsideV2bis.html
On-line inheritance in man: Johns Hopkins University and NCBI	http://www3.ncbi.nlm.nih.gov/Omim/
Whitehead Institute for Biomedical Research	http://www.ornl.gov/hgmis/research/centers.html
D. Gene and genome relationships and proteome[a] analysis	
Alfresco: Visualization tool for genome comparison	http://www.sanger.ac.uk/Software/Alfresco/
allgenes.org: A comprehensive gene index (catalog) derived from ESTs and predicted genes	http://www.allgenes.org/
CGAP: Cancer genome anatomy project	http://www.ncbi.nlm.nih.gov/CGAP
COG (cluster of orthologous groups): A gene classification system (Tatusov et al. 1997, 2000)	http://www.ncbi.nlm.nih.gov/COG/
Comparative DNA analysis across genomes (genome signatures by nucleotide compositional analysis)[b]	Karlin et al. (1998)
DOGS: Database of genome sizes	http://www.cbs.dtu.dk/databases/DOGS/index.html
E-CELL: A modeling and simulation environment for biochemical and genetic processes (Tomita et al. 1999)	http://www.e-cell.org
FAST_PAN for automatic searches of online EST databases to identify new family members (paralogs) (Retief et al. 1999)	http://www.uvasoftware.org/
GeneCensus Genome Comparisons by encoded protein structures	http://bioinfo.mbb.yale.edu/genome/
GeneQuiz: An integrated system for large-scale biological sequence analysis and data management (Andrade et al. 1999; Hoersch et al. 2000)	http://jura.ebi.ac.uk:8765/ext-genequiz/
Genes and disease: Map location on human chromosomes	http://www.ncbi.nlm.nih.gov/disease/
Genome channel at Oak Ridge National Laboratories	http://compbio.ornl.gov/channel/
GOLD™: Genomes OnLine Database (Kyrpides 1999)	http://wit.integratedgenomics.com/GOLD/

Continued.

Table 10.1. *Continued*

D. Gene and genome relationships and proteome[a] analysis *(continued)*

IMGT ImMunoGeneTics Database specializing in Immunoglobulins, T-cell receptors, and Major Histocompatibility Complex (MHC) of all vertebrate species (Ruiz et al. 2000)	http://www.ebi.ac.uk/imgt/index.html
KEGG: Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto 2000)	http://www.genome.ad.jp/kegg/
MIA Molecular Information Agent: A Web server that searches biological databases for information on a macromolecule	http://mia.sdsc.edu/
Orthologous gene alignments at TIGR	http://www.tigr.org/tdb/toga/orth_tables.html
PEDANT: A protein extraction, description, and analysis tool	http://pedant.mips.biochem.mpg.de/
SEQUEST for identification of proteins following mass spectrometry (Link et al. 1999)	http://thompson.mbt.washington.edu/sequest/
STRING Search Tool for Recurring Instances of Neighboring Genes (see Web page) (Snel et al. 2000b)	http://www.Bork.EMBL-Heidelberg.DE/STRING/
Taxonomy browser at the NCBI arranges genomes taxonomically for sequence retrieval	http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/
UniGene System gene-oriented clusters of GenBank sequences useful for gene identification	http://www.ncbi.nlm.nih.gov/UniGene/
U.S. Dept. of Agriculture, Agricultural Research Service reference site for plant and animal genomes (also see TAIR in model genomes); includes international links	http://genome.cornell.edu/

[a] The full complement of proteins produced by an organism, many following gene prediction.
[b] Whole genomes may be compared at the level of dinucleotide composition, codon usage, strand asymmetry for transcription, and rare oligonucleotides. For example, the dinucleotide TA is underrepresented in most prokaryotic and eukaryotic genomes but not in the genomes of several archaea.

E. Metabolism and regulation,[a] functional genomics

2D gel analysis of proteins: List of organisms	http://www.expasy.ch/ch2d/2d-index.html
AlignAce for promoter analysis of coordinately regulated genes, e.g., microarrays by Gibbs sampling (Roth et al. 1998; Hughes et al. 2000; McGuire et al. 2000)	http://atlas.med.harvard.edu/download/
ArrayExpress database at European Bioinformatics Institute for microarray analysis	http://www.ebi.ac.uk/arrayexpress/
BRITE: Database of protein-protein interactions and cross-reference links (see KEGG)	http://www.genome.ad.jp/brite/brite.html
Ecocyc electronic encyclopedia of genes and metabolism of *E. coli* (Karp et al. 2000)	http://ecocyc.PangeaSystems.com/ecocyc/
EpoDBis: A database of genes that relate to vertebrate red blood cells (Erythropoiesis) (Stoeckert et al. 1999)	http://www.cbil.upenn.edu/EpoDB/index.html
Expression Profiler tools for analysis and clustering of gene expression and sequence data	http://ep.ebi.ac.uk/
Functional genomics sites	http://www.ornl.gov/hgmis/publicat/hgn/hgnarch.html#fg
GeneCensus Genome Comparisons by encoded protein structures	http://bioinfo.mbb.yale.edu/genome/

Continued.

Table 10.1. *Continued*

E. Metabolism and regulation,[a] functional genomics *(continued)*

GENECLUSTER; Tamayo et al. (1999)	http://www.genome.wi.mit.edu/MPR/software.html
GeneRAGE for sequence clustering and domain detection; Enright and Ouzounis (2000)	available from authors
GeneX: A Collaborative Internet Database and Toolset for Gene Expression Data	http://www.ncgr.org/research/genex/
MetaCyc metabolic encyclopedia (see EcoCyc)	http://ecocyc.PangeaSystems.com/ecocyc/
Microarray guide: P. Brown lab	http://cmgm.stanford.edu/pbrown/
Microarray project at NIH	http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/
Microarray software	http://rana.lbl.gov/
microarrays.org: A new public source for microarraying information, tools, and protocols	http://www.microarrays.org/
SMART: For the study of genetically mobile protein domains (Schultz et al. 2000)	http://smart.embl-heidelberg.de/
SWISS-2DPAGE: Two-dimensional polyacrylamide gel electrophoresis database (Hoogland et al. 2000)	http://www.expasy.ch/ch2d/
TIGR: Annotation and gene indexing resources, including analysis of the transcribed sequences represented in the public EST databases.	http://www.tigr.org/tdb/tgi.shtml
WIT (What is there?): Interactive metabolic reconstruction on the Web (Overbeek et al. 2000)	http://wit.mcs.anl.gov/WIT2/
Yeast (*S. cerevisiae*) transcriptome	http://bioinfo.mbb.yale.edu/genome/
Yeast genome (*S. cerevisiae*) on a chip	http://cmgm.stanford.edu/pbrown/yeastchip.html

[a] Identification of regulatory sequences is discussed in Chapter 8, and programs for analysis of eukaryotic promoters are listed in Table 8.6 and on page 371.

F. Gene nomenclature, functional characterization, and genome database development

A. thaliana nomenclature	http://www.arabidopis.org/links/nomenclature.html
Genome Annotation and Information Analysis GAIA (Bailey et al. 1998)	http://www.cbil.upenn.edu/gaia2/gaia
GeneQuiz: An integrated system for large-scale biological sequence analysis and data management (Andrade et al. 1999; Hoersch et al. 2000)	http://jura.ebi.ac.uk:8765/ext-genequiz//genequiz.html
GFF (Gene-Finding Features): Specification for describing genes and other features of genomics	http://www.sanger.ac.uk/Software/GFF/
GO (gene ontology) controlled vocabulary	http://genome-www.stanford.edu/GO/
K2 system for support of distributed heterogeneous database and information resource integration	http://www.cbil.upenn.edu/
Kleisli Project: A tool for broad-scale integration of databanks across the Internet (see Chung and Wong 1999)	http://sdmc.krdl.org.sg/kleisli/
MAGPIE: Multipurpose Automated Genome Project Investigation Environment (Gaasterland and Sensen 1998)	http://www.rockefeller.edu/labheads/gaasterland/gaasterland.html, http://genomes.rockefeller.edu/magpie/index.html, see http://magpie.genome.wisc.edu/tools.html
Mendel Plant Gene Nomenclature Database	http://genome-www.stanford.edu/Mendel/
RefSeq and LocusLink: A curated set of reference sequences with map locations, a foundation for functional annotation of the human genome (Pruitt et al. 2000)	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
TAMBIS: A conceptual model of molecular biology and bioinformatics and methods for querying the model (Baker et al. 1999)	http://img.cs.man.ac.uk/tambis/

Prior to the sequencing of H. influenzae, *the first free-living organism to be sequenced, a large number of viruses had been sequenced. Many of these organisms also serve as model systems for studying replication and gene expression. As an example, the nucleotide sequence of bacteriophage lambda was completed by Sanger et al. (1982). A simple way to retrieve sequences of viral and other extrachromosomal genetic elements such as organelles is through the National Center for Biotechnology Information (NCBI) taxonomy browser at http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html.*

species *E. coli* K-12 that had been the subject of many years of genetic and biochemical research. The identification of these genes allowed the investigators to construct some of the biochemical pathways of the *Hemophilus* cell. The function of the other 42% of the *Hemophilus* genes could not be identified, although some of them were similar to the 38% of *E. coli* genes that were also of unknown function. Other unique sequences that appeared to be associated with the ability of the organism to behave as a human pathogen were also found.

The success of sequencing the *Hemophilus* genome in a relatively short time and with a modest budget heralded the sequencing of a large number of additional prokaryotic organisms (see Table 10.1A; de Bruijn et al. 1998). To date, the genomes of 31 of these species have been sequenced. Organisms were selected for sequencing based on at least three criteria: (1) They had been subjected to a good deal of biological analysis, e.g., *E. coli* and *Bacillus subtilis*, and thus were model prokaryotic organisms; (2) they were an important human pathogen, e.g., *Mycobacterium tuberculosis* (tuberculosis) and *Mycoplasma pneumoniae* (pneumonia); or (3) they were of phylogenetic interest. Analysis of the ribosomal RNA molecules of prokaryotes and eukaryotes had led to the prediction of three main branches in the tree of life represented by Archaea, the Bacteria, and the Eukarya.

For genome sequencing projects, organisms have been sampled from throughout the tree (see Fig. 6.3, p. 243), including some that are in deeper branches of the tree and that have growth properties reminiscent of an ancient environment. A summary of the genome size and composition of a representative list of prokaryotes is given in Table 10.2.

As these genome sequences were collected, they were annotated. Annotation involves identifying open reading frames in the genome sequence using the predicted protein as query sequences in a database similarity search and then adding any significant matches to the genome sequence entry in the sequence database. More sophisticated methods of

Table 10.2. *Features of representative prokaryotic genomes*

Organism (reference)	Phylogenetic group	Genome size (Mbp) (no. protein-encoding genes)	Novel functions
Escherichia coli (Blattner et al. 1997)	Bacteria	4.6 (4288)	model organism
Methanococcus jannaschii (Bult et al. 1996)	Archaea	1.66 (1682)[a]	grows at high temperature and pressure and produces methane
Hemophilus influenzae (Fleischmann et al. 1995)	Bacteria	1.83 (1743)	human pathogen
Mycoplasma pneumoniae (Himmelreich et al. 1996)	Bacteria	0.82 (676)	human pathogen that grows inside cells; metabolically weak
Bacillus subtilis (Kunst et al. 1997)	Bacteria	4.2 (4098)	model organism
Aquifex aeolicus (Deckert et al. 1998)	Bacteria	1.55 (1512)[b]	ancient species, grows at high temperature and can grow in a hydrogen, oxygen, carbon dioxide atmosphere in the presence of only mineral salts
Synechocystis sp. (Kaneko et al. 1996a,b)	Bacteria	3.57 (3168)	ancient organism that produces oxygen by light-harvesting; may have oxygenated atmosphere

The genome in each case is contained on a single circular DNA molecule except where noted. Another bacterial species, *Deinococcus radiodurans*, has two chromosomes of sizes 2.6 and 0.4 Mbp and two additional elements of size 0.17 Mb and 46 Kbp (http://www.tigr.org). Other bacterial species have linear chromosomes (for review, see Volff and Altenbuchner 2000).

[a] *M. jannaschii* has a small and a large extrachromosomal element.

[b] *A. aeolicus* has a single extrachromosomal element.

Prokaryotic organisms are included in the Archaea and Bacteria phylogenetic groups.

searching for protein families described in Chapters 7 and 9 are also used for annotation. In examining the results of such analysis, it is important to look for the method used, the statistical significance of the result, and the overall degree of confidence in the alignments. The analysis should be repeated if necessary. Annotation errors occur when the above criteria are not followed (Kyrpides and Ouzounis 1999). Computational resources listed in Table 10.1 can facilitate the analysis of bacterial genomes. GeneQuiz is an example of such a resource. Also shown in Table 10.1A are Web sites that provide a complete annotation of the prokaryotic genomes that have been sequenced.

Eukaryotic Genomes

In addition to having linear chromosomes within a nucleus, and differing from prokaryotic genomes in this respect, eukaryotic genomes commonly have tandem repeats of sequences and include introns in protein-coding genes.

Sequence Repeats

Centromeres hold newly replicated daughter chromosomes together and serve as a point of attachment for pulling the chromosomes apart during cell division.

Because of the skewed base composition of regions that have repeats, they may be purified by virtue of having different buoyant densities and are known as satellite DNA. The sequences fall into different types, each with a different repeat unit of length 5–200 bp. Most of this repetitive DNA is found near the centromere. Also found in eukaryotic genomes are minisatellites made up of repeat units of up to 25 bp and microsatellites composed of repeat units of 4 bp or less. Microsatellite repeats are found at the ends of eukaryotic chromosomes at the telomeres, which in humans comprise hundreds of copies of a 6-bp repeat TTAGGG.

Telomeres are necessary for chromosomal replication.

In nondividing cells, a mixture of lightly and darkly stained chromosomal regions called heterochromatin and euchromatin, respectively, are observed. The centromeric and telomeric regions are located in the heterochromatin, which is in a compact configuration and is thought not to be transcribed. Genes that are transcribed are located in the less compact euchromatin, to which regulatory proteins have access (for review, see Brown 1999).

Transposable Elements

These elements can comprise a large proportion of the eukaryotic genome as repetitive sequences. Transposable elements (TEs) are thought to play an important role in the evolution of these genomes (Kidwell and Lisch 1997, 2000). TEs are DNA sequences that can move from one chromosomal location to another faster than the chromosome can replicate. Hence, TEs have the potential to increase in number until they comprise a large proportion of the genome sequence, a feature already observed in many plants and animals. They remain detectable in the genome until they blend into the background sequence by mutation. The presence of these elements may be demonstrated using programs for detection of low-complexity regions in sequences (see Chapter 6, p. 308). The percentage of genomes that are composed of TEs is depicted in Figure 10.1. For example, more than one-third of the human genome consists of interspersed repetitive sequences derived from TEs.

Eukaryotic TEs fall into two main classes according to sequence similarity and the mechanism of transposition. Class I elements encode a reverse transcriptase and use RNA-mediated mechanisms of transcription. There are three main subclasses of these TEs—the long terminal repeat (LTR) retrotransposons, retroposons, and retrovirus-like elements with LTRs. The LTR retrotransposons are related by genetic structure to retroviruses. The retroposons include short (80–300 bp long) interspersed nuclear elements (SINES) and

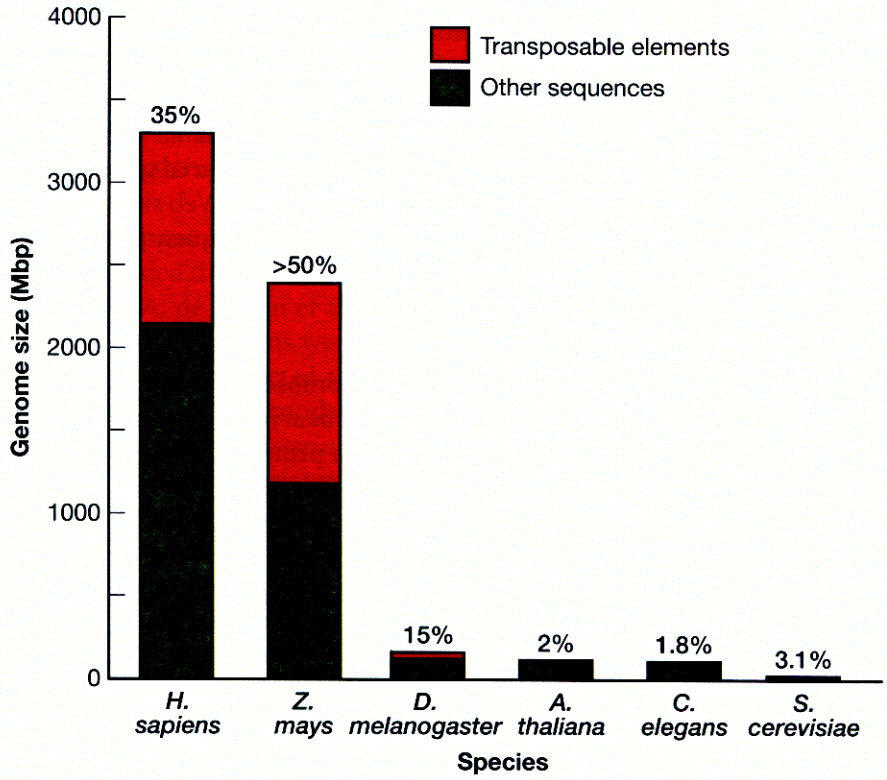


Figure 10.1. Percentages of representative genomes that are made up of transposable elements. The genomes include those of humans, maize, the fruit fly *Drosophila*, the model plant *Arabidopsis*, the nematode *C. elegans*, and budding yeast *S. cerevisiae*, respectively. (Adapted, with permission, from Kidwell and Lisch 2000 [copyright Elsevier Science].)

long (6–8 kbp long) interspersed nuclear elements (LINES). The types of transposable elements that are present in high copy numbers in mammalian genomes are illustrated in Figure 10.2. Ten percent of the human genome comprises one particular family of the SINE element, designated Alu (1.2 million copies) and 14.6% of one particular LINE designated LINE1 (593,000 copies)(Smit 1996).

Vertebrate chromosomes have long (>300 kb) regions of distinct GC richness, repeat content, and gene density, designated isochores in a model of genome organization proposing that genomes are made up of distinct segments of unique composition (Bernardi 1995). Human and mouse chromosomal regions that have a low density of genes are AT-rich and have more Alu or B1/B2 (SINES) than LINE1 elements, whereas the reverse is true for regions that have a high gene density, and those regions are more GC-rich (Henikoff et al. 1997).

The other class of TEs, class II, is made up of elements that employ a DNA-based mechanism of transposition. The human genome contains about 200,000 copies of this class of elements that probably predate human evolution (Smit 1996). Class II elements also include the Activation-Dissociation (Ac-Ds) family in maize and the P element in *Drosophila*.

A third category of TEs has features of both class I and class II TEs. These miniature, inverted repeat TEs (MITES) are 400 bp in length and were discovered in diverse flowering plants where they are frequently associated with regulatory regions of genes. Hence, they could be exerting an influence on regulation of gene expression (Kidwell and Lisch 1997).

The abundance of TEs in the genomes of humans, yeast, maize, and *E. coli* is illustrated in Figure 10.3. The following features are apparent: (1) TEs are present in all of the chromosomes, ranging from bacteria to humans, but their abundance varies; (2) TEs can com-

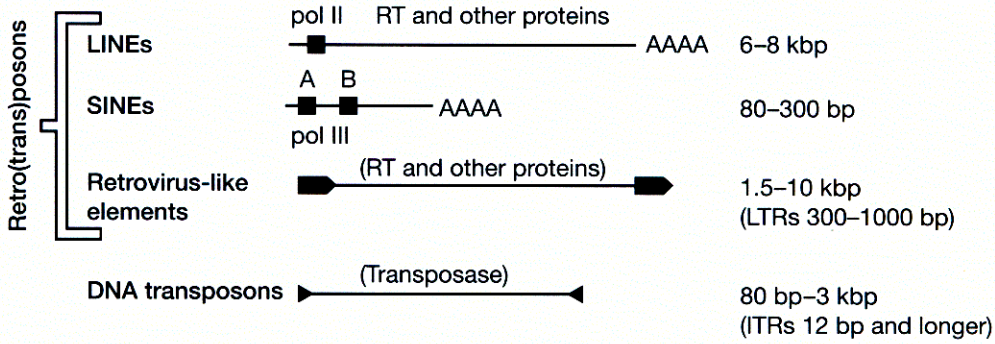


Figure 10.2. Transposable elements that produce high-copy-number interdispersed repeats in mammalian genomes. Shown are class of element, a representation of the structure, size of element plus, in some cases, size of terminal repeats. ■ RNA polymerase II or III promoter; ▬► long terminal repeat (LTR); ►, ◄ inverted terminal repeats; RT reverse transcriptase. Parentheses above elements indicate protein found in autonomous elements. (Redrawn, with permission, from Smit 1996 [copyright Elsevier Science].)

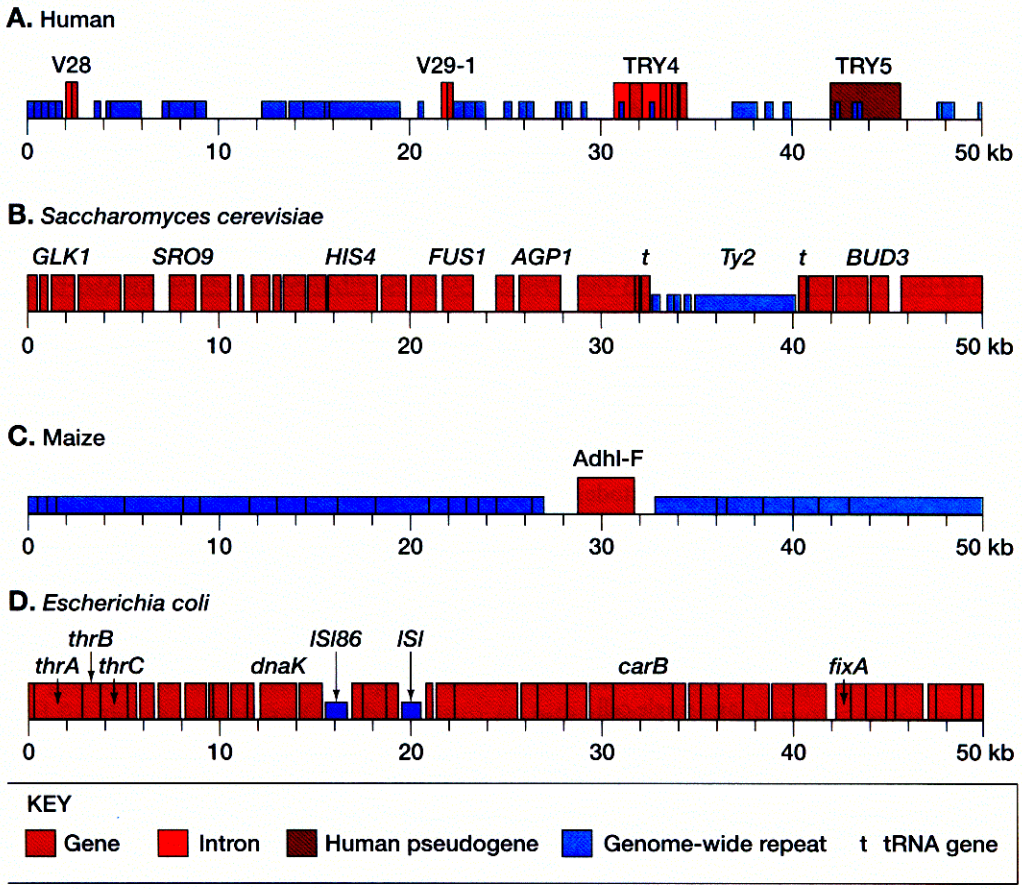


Figure 10.3. Comparison of genome composition in four genomes. (*A*) Human β T-cell receptor locus on chromosome 7. V28 and V29.1 encode parts of the β T-cell receptor proteins that are joined during development of the immune system (Rowen et al. 1996). TRY4, the gene for trypsinogen, and TRY5, a pseudogene related to the trypsinogen family, are not related to the receptor sequence. Why they are located here is not known. (*B*) Segment of yeast chromosome III (Oliver et al. 1992). (*C*, *D*) 50-kb fragments of the maize and *E. coli* chromosomes, respectively (SanMiguel et al. 1996; Blattner et al. 1997). The maize repeats are LTR retrovirus-like elements (Fig. 10.2) that have inserted within the last 3 million years (SanMiguel et al. 1998). (Redrawn, with permission, from Brown 1999 [copyright Wiley-Liss].)

prise a large portion of the genomes of higher eukaryotes, both plants and animals. Thus, only a small fraction of the genome of these organisms carries gene sequences.

Gene Structure Varies in Eukaryotes

Eukaryotic genes that encode proteins are interrupted by introns of varying length and number. In *S. cerevisiae* (budding yeast), only a small fraction of the genes contain introns, and there are a total of 239 introns in the entire genome. In contrast, in individual human genes, introns may be present in numbers exceeding 100 and comprise more than 95% of the gene. Introns can remain at a corresponding position in a eukaryotic gene for long periods of evolutionary time. The origin of introns in eukaryotic genes is not understood but has been accounted for by two models. The "introns-early" view proposes that introns were used to assemble the first genes from sets of ancient conserved exons, whereas the "introns-late" view proposes that introns broke up previously continuous genes by inserting into them (Gilbert et al. 1997).

Complex intron arrangements are often found. RNA of organelles can have introns with introns (Copertino and Hallick 1993), and nuclear genomes can encode genes in which one gene, including introns, is encoded within the introns of a second gene (see, e.g., Cawthon et al. 1990).

The intron structure of genes in a particular eukaryote is used for predicting the location of genes of genome sequences. Other features of eukaryotic genes in a particular organism that are useful for gene prediction include the consensus sequences at exon–intron and intron–exon splice junctions, base composition, codon usage, and preference for neighboring codons. Computational methods described in Chapter 8 incorporate this information into a gene model that may be used to predict the presence of genes in a genome sequence. Although not always correct, these methods provide a useful annotation of a new genome sequence, and in combination with database similarity searches

Table 10.3. *Number of genes predicted to encode proteins in model organisms and humans*

Organism	Biological features	Haploid genome size (Mb)	Predicted number of genes
Arabidopsis thaliana	plant with small genome; genes for metabolism, development by hormones and cell-cell interactions and environmental responses	130	~25,000[a]
Caenorhabditis elegans	worm (nematode) genes for development by a unique cell lineage, nervous system, and reproduction	100	18,424
Drosophila melanogaster	fruit fly; model for developmental processes by hormones and cell-cell interactions	180	13,601
Escherichia coli	bacterium; genes for growth on external sources of energy, transport of molecules through cell membrane, metabolic pathways, and replication as a single cell	4.7	4,288
Homo sapiens (human)	duplicates many gene functions in other model organisms and in addition includes control of higher brain functions	3×10^3	120,000[b]
Saccharomyces cerevisiae	budding yeast; genes for existence as a single-celled organism with the basic structure and organization of the eukaryotic cell	13.5	6,241

Examples of other model organisms that are to be sequenced include the mouse (*Mus musculus*), 3,300 Mb, and rice (*Oryza sativa*), 565 Mb.The mouse genome is a model for the human genome with which it shares a large amount of sequence homology and local gene order. The rice genome is a model for the cereal crops such as wheat (*Triticum aestivum*, genome size 1,700 Mb). The cultivated grasses all share similar genes, and cultivation has resulted in changes in the same genes (Paterson et al. 1995). Plant genomes in general vary in genome size due to the presence of repetitive elements including the number of copies of haploid chromosomes. Wheat, for example, has a hexaploid constitution (for review, see Devos and Gale 2000). The largest plant genomes are members of the Liliaceae family (>87,000 Mb) (see Bennetzen 2000).

[a] Based on the annotation of chromosomes 2 and 4 (Kaneko et al. 1999; Lin et al. 1999).

[b] Based on analysis of 2,000,000 carefully indexed ESTs (Liang et al. 2000). This is higher than previous estimates based on annotation of chromosome 22 (45,000).

described below, provide an indication of the genetic potential of an organism. Numbers of predicted genes estimated from the complete genome sequence of four model eukaryotic organisms are given in Table 10.3. The number of predicted genes in *E. coli* is also given for comparison. Due to the compact gene density in *E. coli* (see Fig. 10.3), there is about one gene per kb of genome sequence. Yeast is about twofold less compact than *E. coli*. Of the remaining genomes, *C. elegans* and *A. thaliana* have approximately the same density of genes (one gene per 6 kb), *Drosophila* being the least dense (one gene per 14 kb). One-sixth of the *Drosophila* sequence is composed of TEs and one-third is heterochromatic regions that do not include genes. Hence, in the euchromatic regions, the gene density in the *Drosophila* genome is one gene per 9 kb. Despite the fact that the lower number of predicted genes in *Drosophila* is smaller than that of the other genomes, the amount of functional diversity, as evidenced by protein family representation, is similar (Adams et al. 2000). Assessment of genome functional diversity is discussed in the following sections.

Pseudogenes

New gene functions are thought to be gained by duplication of an existing gene creating two tandem copies. Functional differentiation then occurs between the copies by mutation and selection. However, because most mutations are deleterious, and because only one gene copy may be needed for function, there is a strong tendency of one copy to accumulate mutations that render the gene nonfunctional. Accordingly, pseudogenes are DNA sequences that were derived from a functional copy of a gene but which have acquired mutations that are deleterious to function (Li 1997). In Figure 10.3A, the pseudogene *TRY5* is similar to the nearby functional gene *TRY4*.

There is also a second type of pseudogene found in eukaryotic genomes called a processed pseudogene. Processed pseudogenes are also derived from a functional gene, but they do not contain introns and lack a promoter; hence, they are not expressed. The origin of these pseudogenes is probably due to reverse transcription of the mRNA of the functional gene and insertion of the cDNA copy into a new chromosomal location by a LINE1 (Fig. 10.2) reverse transcriptase (Weiner 2000).

SEQUENCE ASSEMBLY AND GENE IDENTIFICATION

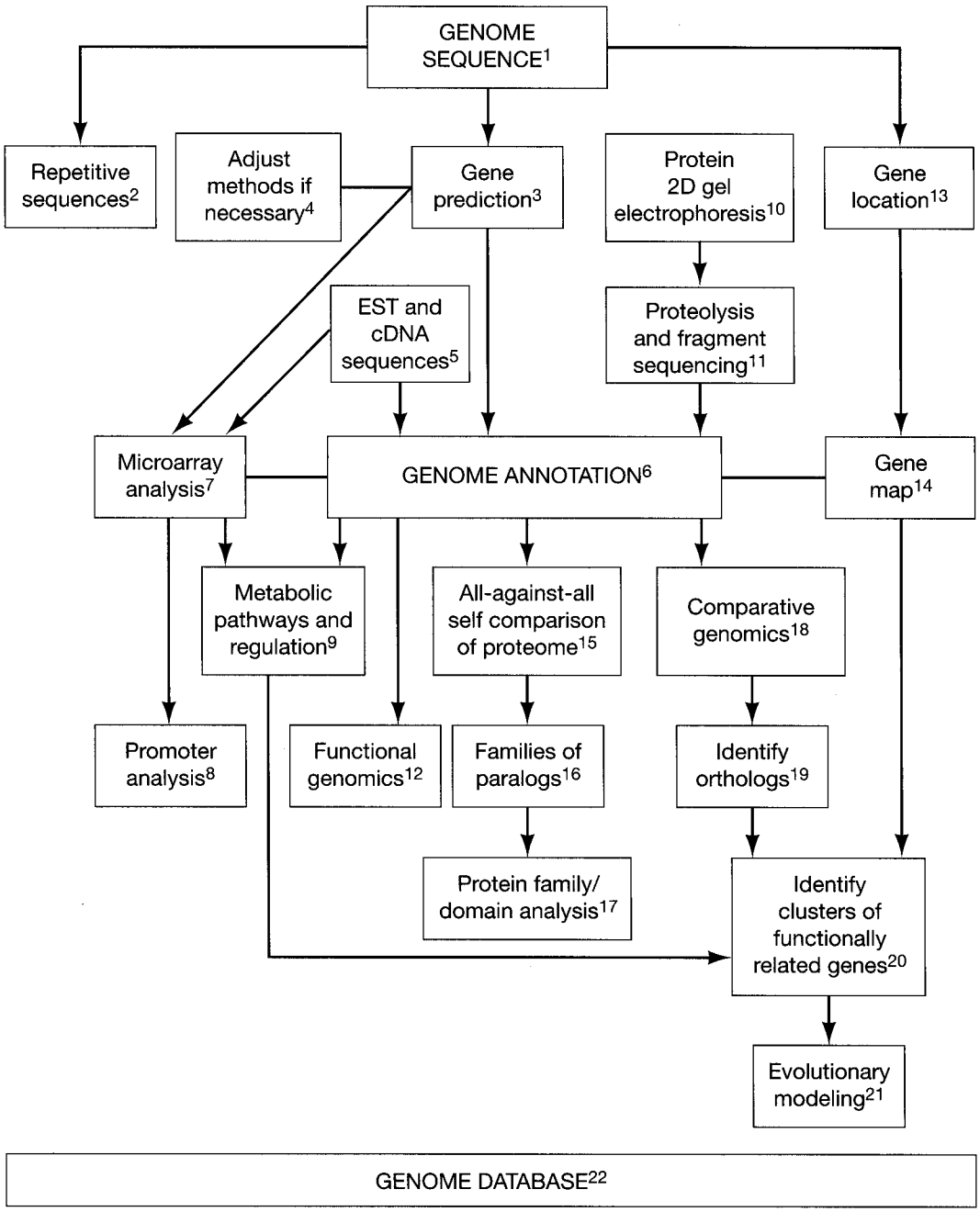
As discussed in Chapter 2, sequencing of genomes depends on the assembly of a large number of DNA reads into a linear, contiguous DNA sequence. The cost and efficiency of this process has been greatly improved by automatic methods of sequence assembly, first used for the sequencing of the bacterium *H. influenzae* (see Prokaryotic genome, p. 481). This same method of assembly was also used, in part, to complete the sequencing of the *Drosophila* (Myers et al. 2000) and human genomes in a timely manner.

As illustrated and explained in the Chapter 10 flowchart (p. 492), each genome sequence is scanned for protein-encoding genes using gene models trained on known gene sequences from the same organism. Methods of gene prediction in eukaryotic genomic DNA are discussed in Chapter 5 (for RNA-encoding genes) and Chapter 8 (for protein-encoding genes).

Identification of the function of protein-encoding genes is discussed in the Chapter 10 flowchart and in Chapter 7. For a new genome, each predicted gene is translated into a protein sequence; the collection of protein sequences encoded by the genome is the proteome of the organism. As illustrated in Figure 10.4, left panel, every protein in the proteome is then used as a query sequence in a database similarity search. Matching database sequences are realigned with the query sequence to evaluate the extent and significance of the alignment, as described in Chapter 2.

Screening the predicted protein sequences against an expressed sequence tag (EST) library confirms the prediction and expression of the gene (see Adams et al. 2000). The collective information on proteome function can then be further analyzed by self-comparison to find duplicated genes (paralogs) and by a proteome-by-proteome comparison to identify orthologs, genes that have maintained the same function through speciation, and other sequence and evolutionary relationships that are important for metabolic, regulatory, and cellular functions. These proteome comparisons are described in the next section.

METHODS



1. Genome sequences are assembled from DNA sequence fragments of approximate length 500 bp obtained using DNA sequencing machines as described in Chapter 2. Chromosomes of a target organism are purified, fragmented, and subcloned in fragments of size hundreds of kbp in bacterial artifical chromosomes (BACs). The BAC fragments are then further subcloned as smaller fragments into plasmid vectors for DNA sequencing (although the ends of BACs may also be sequenced as a way to circumvent problems with sequence repeats; see Myers et al. 2000). Full chromosomal sequences are then assembled from the overlaps in a highly redundant set of fragments by an automatic computational method (Myers et al. 2000) or from the fragment order on a physical map.
2. Eukaryotic genomes comprise classes of repeated elements, including tandem repeats present in centromeres and telomeres, dispersed tandem repeats (minisatellites and macrosatellites), and interdispersed TEs. TEs can comprise one-half or more of the genome sequence. Analysis of sequence repeats is discussed in Chapters 3 and 7. Identification of classes of repeated elements is aided by searchable databases discussed in Chapter 7 (p. 309).
3. Gene identification in prokaryotic organisms is simplified by their lacking introns. Once the sequence patterns that are characteristic of the genes in a particular prokaryotic organism (e.g, codon usage, codon neighbor preference) have been found, gene locations in the genome sequence can be predicted quite accurately. The presence of introns in eukaryotic genomes makes gene prediction more involved because, in addition to the above features, locations of intron–exon and exon–intron splice junctions must also be predicted. Methods of gene prediction in prokaryotes and eukaryotes are discussed in Chapter 8.
4. Gene prediction methods involve training a gene model (e.g., a hidden Markov model or neural network, see Chapter 8) to recognize genes in a particular organism. Due to variations in gene codon preferences and splice junctions (see note 3, Fig. 10.3), a model must usually be trained for each new genome.
5. Since gene prediction methods are only partially accurate (for review, see Bork 1999; see Chapter 8), gene identification is facilitated by high-throughput sequencing of partial cDNA copies of expressed genes (called expressed sequence tags or EST sequences). Presence of ESTs confirms that a predicted gene is transcribed. A more thorough sequencing of full-length cDNA clones may be necessary to confirm the structure of genes chosen for a more detailed analysis.
6. The amino acid sequence of proteins encoded by the predicted genes is used as a query of the protein sequence databases in a database similarity search. A match of a predicted protein sequence to one or more database sequences not only serves to identify the gene function, but also validates the gene prediction. Pseudogenes, gene copies that have lost function, may also be found in this analysis. Only matches with highly significant alignment scores and alignments (see Chapter 3, page 58) should be included. The genome sequence is annotated with the information on gene content and predicted structure, gene location, and functional predictions. The predicted set of proteins for the genome is referred to as the proteome. Accurate annotation is extremely important so that others users of the information are not misinformed. Procedures for searches starting with genome, EST, and cDNA sequence are described in Chapter 8. Usually, not all query proteins will match a database sequence. Hence, it is important to extend the analysis by searching the predicted protein sequence for characteristic domains (conserved amino acid patterns that can be aligned) that serve as a signature of a protein family or of a biochemical or structural feature (see note 17). A further extension is to identify members of protein families or domains that represent a structural fold using the computational tools described in Chapter 9. This additional information also needs to be accurately described and the significance established.
7. Microarray analysis provides a global picture of gene expression for the genome by revealing which genes are expressed at a particular stage of the cell cycle or developmental cycle of an organism, or genes that respond to a given environmental signal to the same extent. This type of information provides an indication as to which genes share a related biological function or may act in the same biochemical pathway and may thereby give clues that will assist in gene identification.
8. Genes that are found to be coregulated either by a microarray analysis or by a protein two-dimensional analysis should share sequence patterns in the promoter region that direct the activity of transcription factors. The types of analyses that are performed are discussed in Chapter 8 (pp. 357–373), and additional tools for analyzing coregulated genes are listed in Table 10.1E.

9. As genes are identified in a new genome sequence, some will be found that are known to act sequentially in a metabolic pathway or to have a known role in gene regulation in other organisms. From this information, the metabolic pathways and metabolic activities of the organism will become apparent. In some cases, the apparent absence of a gene in a well-represented pathway may lead to a more detailed search for the gene. Clustering of genes in the pathway on the genome of a related organism can provide a further hint as to where the gene may be located (see note 20).
10. Individual proteins produced by the genome can be separated to a large extent by this method and specific ones identified by various biochemical and immunological tests. Moreover, changes in levels of proteins in response to an environment signal can be monitored in much the same way as a microarray analysis is performed. Microarrays only detect untranslated mRNAs, whereas a two-dimensional gel protein analysis detects translation products, thus revealing an additional level of regulation. Resources for analysis of regulation by this method are given in Table 10.1D.
11. Protein spots may be excised from a two-dimensional protein gel (see note 10) and subjected to a combination of amino acid sequencing and cleavage analyses using the techniques of mass spectrometry and high-pressure liquid chromatography. Genome regions that encode these sequences can then be identified and the corresponding gene located. A similar method may be used to identify the gene that encodes a particular protein that has been purified and characterized in the laboratory. The computational methods are described in Chapter 7 (p. 295, FASTA tools) and Table 10.1D.
12. Functional genomics involves the preparation of mutant or transgenic organisms with a mutant form of a particular gene usually designed to prevent expression of the gene. The gene function is revealed by any abnormal properties of the mutant organism. This methodology provides a way to test a gene function that is predicted by sequence similarity to be the same as that of a gene of known function in another organism. If the other organism is very different biologically (comparing a predicted plant or animal gene to a known yeast gene), then functional genomics can also shed light on any newly acquired biological role. When two or more members of a gene family are found (see notes 16 and 17), rather than a single match to a known gene, the biological activity of these members may be analyzed by functional genomics to look for diversification of function in the family.

A more detailed analysis of the relative amount of sequence variability in a chromosomal region within populations of closely related species can reveal the presence of genes that are under selection. These regions will not have the expected amount of variability given their linkage: They are in a state of linkage disequilibrium. An example is the BRCA1 *(breast cancer 1) gene of humans and chimpanzees (Huttley et al. 2000).*

13. Since the entire genome sequence is available, as each gene is identified, the relative position of the gene will be known.
14. A map showing the location of each identified gene is made. These relative positions of genes can be compared to similar maps of other organisms to identify rearrangements that have occurred in the genome. Gene order in two related organisms reflects the order that was present in a common ancestor genome. Chromosomal breaks followed by a reassembly of fragments in a different order can produce new gene maps. These types of evolutionary changes in genomes have been modeled by computational methods (p. 512). Gene order is revealed not only by the physical order of genes on the chromosome, but also by genetic analysis. Populations of an organism show sequence variations that are readily detected by DNA sequencing and other analysis methods. The inheritance of genetic diseases in humans and animals (e.g., cancer and heart disease), and of desirable traits in plants, can be traced genetically by pedigree analysis or genetic crosses. Sequence variations (polymorphisms) that are close to (tightly linked) a trait may be used to trace the trait by virtue of the fact that the polymorphism and the trait are seldom separated from one generation to the next. These linked polymorphisms may then be used for mapping and identifying important genes.
15. A comparison is made in which every protein is used as a query in a similarity search against a database composed of the rest of the proteome, and the significant matches are identified by a low expect value ($E<10^{-6}$ was used in a recent analysis by Rubin et al. [2000]). Since many proteins comprise different combinations of a common set of domains, proteins that align along most of their lengths (80% identity is a conservative choice) are chosen to select those that have a conserved domain structure.
16. A set of related proteins identified in step 15 is subjected to a cluster analysis in order to identify the most closely related groups of proteins and to avoid domain-matching. This group of proteins is derived from a gene family of paralogs that have arisen by gene duplication.

17. Each protein in the predicted proteome is again used as a query of a curated protein sequence database such as SwissProt in order to locate similar domains and sequences. The domain composition of each protein is also determined by searching for matches in domain databases such as Interpro, described in Table 9.5. The analysis reveals how many domains and domain combinations are present in the proteome, and reveals any unusual representation that might have biological significance. The number of expressed genes in each family can also be compared to the number in other organisms to determine whether or not there has been an expansion of the family in the genome.
18. Comparative genomics is a comparison of all the proteins in two or more proteomes, the relative locations of related genes in separate genomes, and any local groupings of genes that may be of functional or regulatory significance.
19. Orthologs are genes that are so highly conserved by sequence in different genomes that the proteins they encode are strongly predicted to have the same structure and function and to have arisen from a common ancestor through speciation. To identify orthologs, each protein in the proteome of an organism is used as a query in a similarity search of a database comprising the proteomes of one or more different organisms. The best hit in each proteome is likely to be with an ortholog of the query gene. In comparing two proteomes, a common standard is to require that for each pair of orthologs, the first of the pair is the best hit when the second is used to query the proteome of the first. To find orthologs, very low E value scores ($E<10^{-20}$) for the alignment score and an alignment that includes 60–80% of the query sequence are generally required in order to avoid matches to paralogs. Although these requirements for classification of orthologs are very stringent, a more relaxed set of conditions will lead to many more false-positive predictions. In bacteria, the possibility of horizontal transfer of genes between species also has to be considered (p. 508).
20. In related organisms, both gene content of the genome and gene order on the chromosome are likely to be conserved. As the relationship between the organisms decreases, local groups of genes remain clustered together, but chromosomal rearrangements move the clusters to other locations. In microbial genomes, genes specifying a metabolic pathway may be contiguous on the genome where they are coregulated transcriptionally in an operon by a common promoter. In other organisms, genes that have a related function can also be clustered. Hence, the function of a particular gene can sometimes be predicted, given the known function of a neighboring, closely linked gene. Genomes are also compared at the level of gene content, predicted metabolic functions, regulation as revealed by microarray analysis, and others. These comparisons provide a basis for additional predictions as to which genes are functionally related. Gene fusion events that combine domains found in two proteins in one organism into a composite protein with both domains in a second organism are also found and provide evidence that the proteins physically interact or have a related function.
21. Evolutionary modeling can include a number of types of analyses including (1) the prediction of chromosomal rearrangements that preceded the present arrangement (e.g., a comparison of mouse and human chromosomes), (2) analysis of duplications at the protein domain, gene, chromosomal, and full genome level, and (3) search for horizontal transfer events between separate organisms.
22. Due to the magnitude of the task, the earlier stages of genome analysis including gene prediction and database similarity searches are performed automatically with little human intervention. The genome sequence is then annotated with any information found without involving human judgment. The types of genome analyses in the flowchart also provide many predictions and give rise to many preliminary hypotheses regarding gene function and regulation. As more detailed information is collected by laboratory experiment and by a closer examination of the sequence data, this information needs to be linked to the genome sequence. In addition, the literature, past and present, needs to be scanned for information relevant to the genome. A carefully crafted database that takes into account the entire body of information should then be established. In addition to information on the specific genome of interest, the database should include cross-references to other genomes. To facilitate such intergenome comparisons, common gene vocabularies have been proposed. This slow, expensive, and time-consuming phase of genome analysis is of prime importance if the genome information is to be available in an accurate form for public use.

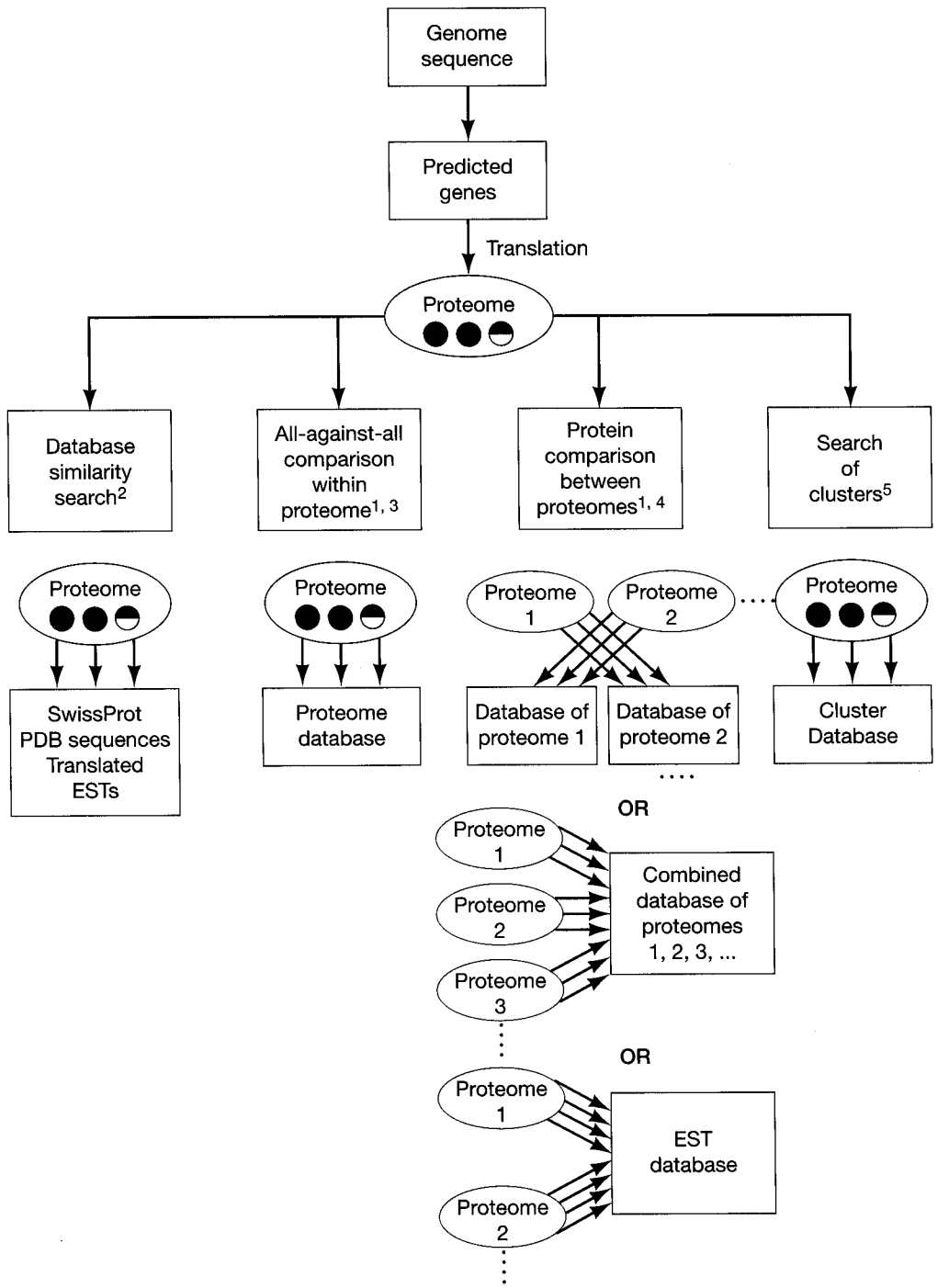
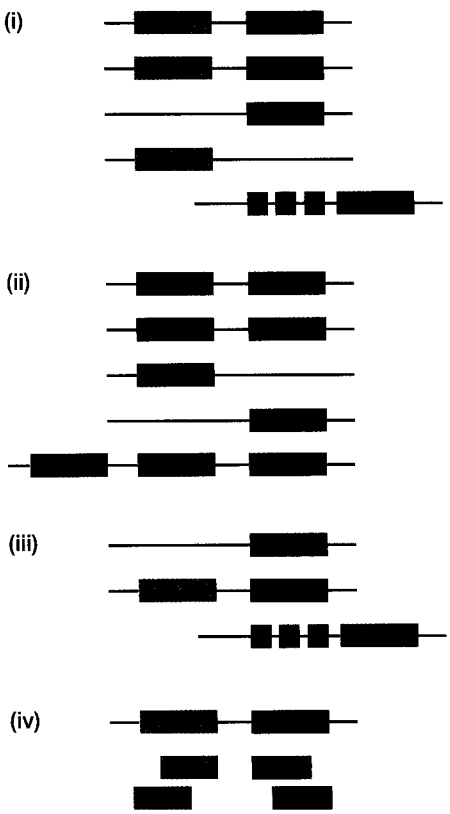


Figure 10.4. Analysis of the proteome encoded by genomes. (*A*) Types of proteome analyses. (*B*) Examples of database hits resulting from domain structure of proteins. (*C*) Cluster analysis of similar sequences. (*D*) Domain identification.

Notes:

1. Due to the large number of comparisons that must be made in these types of analyses (as many as 20,000 by 20,000 sequences) and due to the volume of program output, the procedure must be automated on a local machine using Perl scripts or a similar method and a database system. For BLAST, setting an effective database size appropriate for each search and program is important for obtaining a correct statistical evaluation of alignment scores. The bioperl project provides valuable resources for this purpose (http://www.bioperl.org).

B. Examples of database hits resulting from domain structure of proteins[6]



	Amino acid alignment	Sequence number	Typical range of P/E value[7]
(i)		1 (query)	–
		2	$<10^{-20}$
		3	$10^{-8} - 10^{-20}$
		4	$10^{-8} - 10^{-20}$
		5	$10^{-6} - 10^{-8}$
(ii)		6 (query)	–
		7	$<10^{-20}$
		8	$10^{-8} - 10^{-20}$
		9	$10^{-8} - 10^{-20}$
		10	$<10^{-20}$
(iii)		3 (query)	–
		1,2	$10^{-8} - 10^{-20}$
		5	$10^{-6} - 10^{-8}$
(iv)		1 (query)	–
		EST hits	$<10^{-4}$

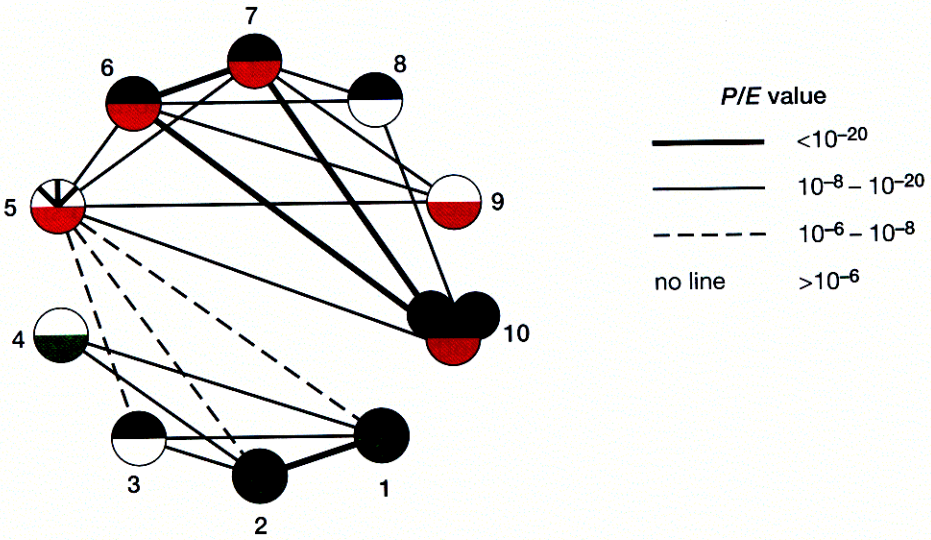
2. Each protein encoded by the genome is used as a query in database similarity searches to identify similar database proteins, some having a known structure or function. Additional searches of EST databases can be used to identify additional relatives of the query sequence. These searches and evaluation of the alignment scores of matching sequences are described in Chapter 7.
3. An all-against-all analysis requires first making a database of the proteome. This database is then sequentially searched by each individual protein sequence of the proteome using a rapid database similarity search tool such as BLAST, WU-BLAST, or FASTA. The scoring systems of these programs vary and are described in Chapter 7. Note also that *P* values of WU-BLAST (Chervitz et al. 1998) are similar to *E* values of NCBI BLAST (Rubin et al. 2000) for values of *P* and $E < 0.05$. This analysis generates a matrix of alignment scores, each with an *E* value and corresponding alignment for each pair of proteins. Recall that the *E* value of an alignment score is the probability that an alignment score as good as the one found would be observed between two random or unrelated sequences in a search of a database of the same size. The lower the *E* value, the more significant the alignment between a pair of matching sequences. In an all-against-all comparison within one proteome, significantly matched pairs of sequences may be paralogs that originated from a gene duplication event in this genome or the genome of an ancestor organism. Unique proteins can be identified through their not matching any other protein. A conservative cutoff *E* value (e.g., 10^{-6}; Rubin et al. 2000) limits the matches to the most significant ones, which are then clustered into families as described below and in the text.
4. To perform a between-proteome analysis, proteome databases are made for the known and predicted genes of two or more genomes. Both single (Chervitz et al. 1998) and combined proteome databases may be made (Rubin et al. 2000). Each protein of one proteome is then selected in turn as a query of the proteome of another organism or the combined proteome of a group of organ-

isms. As in an all-by-all protein comparison within a proteome, a matrix of alignment scores with *E* values is made, and the most closely related sequences in the two organisms are identified. This analysis can predict orthologs, i.e., proteins that have an identical function attributable to descent of the respective genes from a common ancestor. The types of criteria used in bioinformatics to define orthologs include (1) reciprocal database searches with one sequence as query give a best hit of the other sequence (Tatusov et al. 1997); (2) the alignment of the sequences includes at least 80% of each sequence (Chervitz et al. 1998; Rubin et al. 2000); and (3) the sequences are clustered when all matching sequences are subjected to a cluster analysis. The likelihood of orthology is also increased if a set of orthologous pairs are linked together on the respective genomes. The types of analyses are discussed further in the text.

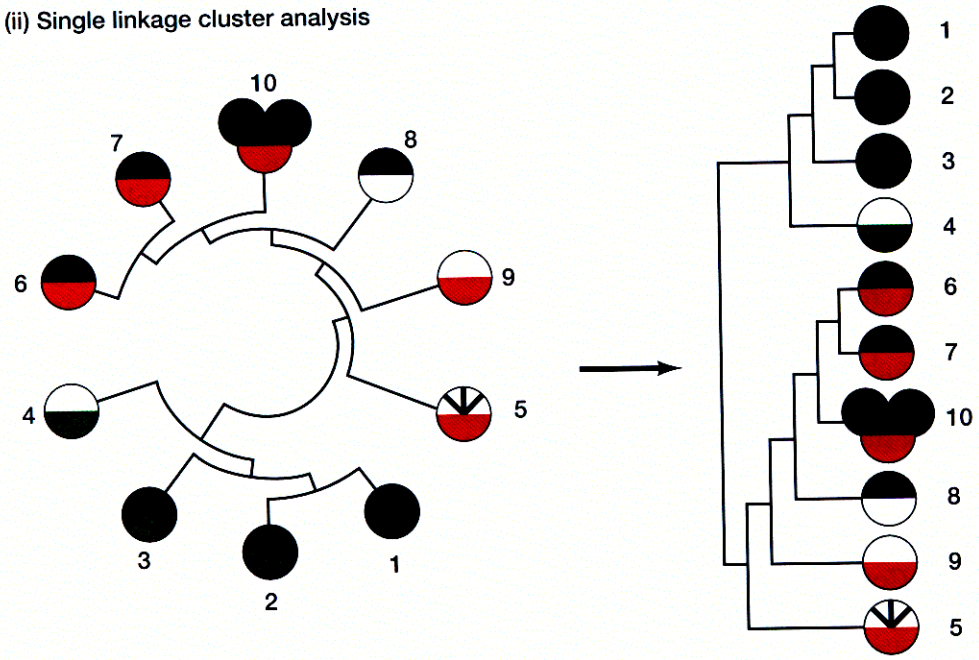
5. The cluster search option is most useful for prokaryotic organisms. Each protein in the proteome is used as a query of a database of protein clusters using the program COGNITOR (Table 10.1, COGs entry). These clusters are composed of orthologous pairs of sequence defined by criterion 1, described in note 4. The database was made by performing an all-by-all genome comparison across a spectrum of prokaryotic organisms and a portion of the yeast proteome (Tatusov et al. 2000). Orthologous pairs of sequence were then merged with clusters or orthologous pairs (COGs) for multiple proteomes as described in the text. COGs have been linked to classes of biochemical function (Tatusov et al. 1997). Hence, matching a query sequence to the COG can potentially identify unique orthologs in another proteome that may have the same function. The COGs database is designed to provide a preliminary indication of orthologous relationships that can be tested by more detailed similarity searches, sequence alignments, and phylogenetic analysis of the matching sequences.
6. Due to the modular nature of proteins, several types of matches may be identified in the all-against-all and between-proteome comparisons. Each colored box represents a hypothetical conserved domain that is matched in the search. The dotted box (sequence 5) represents a less similar domain that will not align as well. Highest-scoring matches corresponding to matching of multiple domains present in the query and in the matched sequence ([i] and [ii], sequence pairs 1 and 2, 6 and 7, etc.). The alignment scores of these pairs should have extremely low *E* values. A multidomain query protein will also match database proteins that have a single domain (as in sequences 1 and 3, 6 and 8). Because only one domain is represented by the alignment, the alignment will in general be shorter and have a poorer (higher) *E* value score than a multidomain alignment. The analysis will also identify matches of a query with a database protein that has two or more copies of query sequence domain (sequence 10). Query sequences with a minimal domain representation (ii) will not score particularly well with any sequence (sequence 3). Duplicate comparisons generated by the method are eliminated. When only an EST library of an organism is available, the proteome may be compared to this library. However, since these databases are generally not complete and any alignments are shorter, it is diffcult to compare these results with the full proteome comparisons. From a biological standpoint, ESTs define expressed genes, whereas proteomes are predicted genes.
7. WU-BLAST produces *P* scores and BLAST (NCBI) *E* scores where $E = -\ln(1 - P)$. For values less than 0.05, $E = P$. The score ranges depicted in this column are hypothetical examples. The choice of a $<10^{-20}$ score is a conservative one for identification of orthologs that should have a similar domain structure, as do the sequences in this example (see Chervitz et al. 1998; Rubin et al. 2000). To define these groups, the distribution of hits below different thresholds should be examined, as in the above references. The higher cutoff score for EST matches is used because the search of an EST database may only produce short alignments.
8. Shown are two representations of the sequence relationships found in part *B*. In (i) the sequences, color coded to represent domain structure, are represented by vertices on graph. In comparing the graphic (i) and single linkage (ii) clusters, note that in (i) each sequence has multiple edges representing links to related sequences, whereas in (ii) the sequences are only connected to one branch on the outermost part of the tree.
9. The sequence alignments found above represent the presence of one or more conserved domains in each cluster or group of clusters. These clusters are next analyzed for the presence of known domains by searches of domain databases as described in Chapter 9. This analysis identifies the number and types of domains that are shared between organisms, or that have been duplicated in proteomes to produce paralogs.

C. Cluster analysis of similar sequences[8]

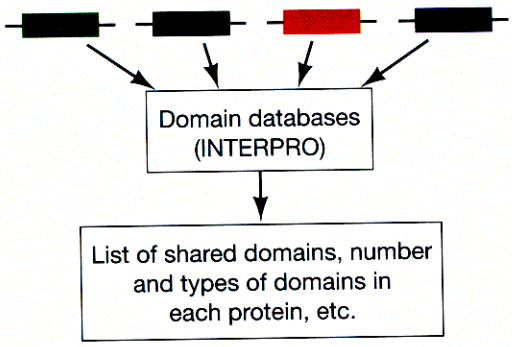
(i) Graphic representation



(ii) Single linkage cluster analysis



D. Domain identification[9]



COMPARATIVE GENOMICS

Comparative genomics includes a comparison of gene number, gene content, and gene location in both prokaryotic and eukaryotic groups of organisms. The availability of complete genome sequences makes possible a comparison of all of the proteins encoded by one genome, the proteome of that organism, with those of another. Because the genome sequence provides both the sequence and the map location of each gene, both the sequence and location can be compared. Sequence comparisons provide information on gene relationships—the number of genes in two organisms that are so similar that they must have the same function and evolutionary history—these genes are orthologs (Fitch 1970). Map locations of orthologous genes may also be compared. If a set of genes is grouped together at a particular chromosomal location, and if a set of similar genes is also grouped together in the genome of another organism, these groups share an evolutionary history.

Proteins may also be clustered into families on the basis of either sequence or structural similarity, as discussed in Chapter 9. Proteins are modular and often comprise separate domains. The number of protein sequences that are available is sufficient to determine that domain shuffling occurs in evolution—domains appear or disappear in particular families, become combined to make new families, or else become separated into two different proteins that are predicted to interact (Snel et al. 2000a). The comparisons of proteomes of different organisms can identify the type of domain changes and also provide an indication as to what biological role they may have in a particular organism.

The assortment and reassortment of protein domains takes place in individual genomes. Proteins with new functions are produced by a gene duplication event in which two tandem copies of a gene are produced (see Fig. 3.3, p. 55). Through mutation and natural selection, one of the copies can develop a new function, leaving the other copy to cover for the original function. However, because most mutations are deleterious to function, often one of the copies becomes a pseudogene. Not all gene duplications are thought to have the above effects. Another scenario is that two duplicated genes both undergo change, but interactions between the proteins stabilize the original function and support the evolution of new ones (Force et al. 1999).

The processes of domain assortment and gene duplication produce families of proteins in organisms. Following speciation, a newly derived genome will inherit the families of ancestor organisms, but will also develop new ones to meet evolutionary challenges. Comparison of each of the proteins encoded by an organism with every protein, an all-against-all comparison, reveals which protein families have been amplified and what rearrangements have occurred as steps in the evolutionary process. When two or more proteins in the proteome share a high degree of similarity because they share the same set of domains (illustrated in Fig. 10.4B), they are likely to be paralogs (Fitch 1970), genes that arose by gene duplication events. Proteins that align over shorter regions share some domains, but also may not share others. Although gene duplication events could have created such variation, other rearrangements may have also occurred, blurring the evolutionary history.

The following sections describe methods to compare prokaryotic and eukaryotic genomes for orthologs and paralogs. It is important to keep in mind the predictive nature of these types of analyses. Decisions about gene relationships depend on careful manual inspection of sequence alignments (Huynen et al. 2000).

Proteome Analysis

All-against-all Self-comparison Reveals Numbers of Gene Families and Duplicated Genes

A comparison of each protein in the proteome with all other proteins distinguishes unique proteins from proteins that have arisen from gene duplication, and also reveals the number of protein families. The domain content of these proteins may also be analyzed. One type of all-against-all proteome comparison is described in Figure 10.4A, second panel. In this analysis, each protein is used as a query in a similarity search against the remaining proteome, and the similar sequences are ranked by the quality and length of the alignments found. The search is conducted in the manner described in Chapter 7, with each alignment score receiving a statistical evaluation (*P* or *E* value). As shown in Figure 10.4B, a match between a query sequence and another proteome sequence with the same domain structure will produce a high-scoring, highly significant alignment. These proteins are designated paralogs because they have almost certainly originated from a gene duplication event. Lower-scoring, less significant alignments may have identified proteins that share domains but not the high degree of sequence similarity that is apparent in the best-scoring alignments. These may also be paralogs, but they may also have a complicated history of domain shuffling that is difficult to reconstruct.

Cluster analysis. To sort out relationships among all of the proteins that are found to be related in a series of searches of the types shown in Figure 10.4B, they are subjected to a clustering analysis shown in Figure 10.4C. Only the relationships revealed by the hypothetical set of searches illustrated in part B are shown. Some of the proteins may have other relationships, which are not depicted in order to simplify the example.

Clustering organizes the proteins into groups by some objective criterion. One criterion for a matching protein pair is the statistical significance of their alignment score (the *P* or *E* value from BLAST searches). The lower this value, the better the alignment. There will be a cutoff *P* or *E* value at which the matches in the BLAST search are no longer considered significant. A value of *P* or $E > 0.01$–0.05 is usually the point at which the alignment score is no longer considered to be significant in order to focus on a more closely related group of proteins. A second criterion for clustering proteins is the distance between each pair of sequences in a multiple sequence alignment. The distance is the number of amino acid changes between the aligned sequences.

Clustering by making subgraphs. Figure 10.4 indicates two ways of clustering related sequences based on the above criteria. Part (i) is a graph in which each sequence is a vertex and each pair of sequences that is matched with a significant alignment score is joined by an edge that is weighted according to the statistical significance of the alignment score. One way to identify the most strongly supported clusters is simply to remove the most weakly supported edges in the graph, in this case the alignments with the highest *P/E* scores (dotted edges). As weaker and weaker links are removed, the remaining combinations of vertices and edges represent most strongly linked sequences. This type of analysis was performed on an initial collection of *E. coli* genes by Labedan and Riley (1995). Their analyses revealed that *E. coli* genes clustered in this manner encode proteins already known to belong to the same broad functional category, EC number, or to have a similar physiological function. For another approach to identify orthologs in microbial genes, see Bansal (1999).

Another method for clustering similar sequences that are likely to be paralogs is described in Rubin et al. (2000). In this method, edges of *E* value $> 1 \times 10^{-6}$ are removed. The remaining graph is then broken down into subgraphs comprising sequences that

share a significant relationship to each other but not to other sequences. The criterion chosen is that the group should mutually share at least two-thirds of all of the edges from this group to all proteins in the proteome. If two proteins A and B share a domain but do not share another domain in A, and if A shares this other domain with a number of other sequences, the algorithm would tend not to cluster A with B (Rubin et al. 2000). Thus, the algorithm favors the selection of proteins with the same domain structure reflecting that these proteins are the most likely ones to be paralogs.

Clustering by single linkage. A second method for clustering related sequences is shown in Figure 10.4C, part (ii). This method is based on the distance criterion for sequence relationships described above. First, a group of related sequences found in the all-against-all proteome comparison is subjected to a multiple sequence alignment usually by CLUSTALW (Chapter 4, p. 154). A distance matrix that shows the number of amino acid changes between each pair of sequences is then made. This matrix is then used to cluster the sequences by a neighbor-joining algorithm. This procedure and the algorithms are the same as those used to make a phylogenetic tree by the distance methods, described in Chapter 8. These methods produce a tree (Fig. 10.4C, part ii, left) or a different representation of the tree called a dendrogram (Fig. 10.4C, part ii, right), that minimizes the number of amino acid changes that would generate the group of sequences. The tree is also defined as a minimum spanning tree (Duran and Odell 1974). The tree and dendrogram cluster the sequences into the most closely related groups. Branches joining the least related sequences may be removed, thus leaving two sub-trees with a small group of sequences. As smaller groups are chosen, the most strongly supported clusters are likely to be made up of paralogs. However, it is not easy to distinguish sequences that are paralogs, i.e., share several domains, from those that share domains but that also share other domains with more distantly related sequences without inspection of the alignments. GeneRage (Table 10.1E) provides an automatic system for classifying protein data sets by means of an iterative refinement approach using local alignments, matrix methods, and single-linkage clustering.

Core proteome. The above types of all-against-all analyses provide an indication as to the number of protein/gene families in an organism. This number represents the core proteome of the organism from which all biological functions have diversified. A representative sample is shown in Table 10.4.

In *Hemophilus*, 1247 of the total number of 1709 proteins do not have paralogs (Rubin et al. 2000). The core proteomes of the worm and fly are similar in size but with a greater number of duplicated genes in the worm. It is quite remarkable that the core proteome of the multicellular organisms (worm and fly) is only twice that of yeast.

Table 10.4. Numbers of gene families and duplicated genes in model organisms (Rubin et al. 2000)

Organism	Total number of genes	Number of gene families[a]	Number of duplicated genes[b]
Hemophilus influenzae (bacteria)	1709	1425[c]	284
Saccharomyces cerevisiae (yeast)	6241	4383	1858
Caenorhabditis elegans (worm)	18,424	9453	8971
Drosophila melanogaster (fly)	13, 600	8065	5536

[a] The number of clustered groups in the all-against-all analysis using the algorithm described in the text. This number represents the core proteome of the organism.

[b] Count of number of duplicated genes within the protein family clusters.

[c] 178 families have paralogs.

Grouping Sequences

The problem of deciding which sequences to include in the same group or cluster and which to separate into different groups or clusters is a recurring one. The conservative approach is to group only very similar sequences together. However, in making a conservative multiple sequence alignment with only very alike sequences, it is not possible to analyze the evolutionary divergence that may have occurred in a family of proteins. Furthermore, if a matrix or profile model is made from this alignment, that model will not be useful for identifying more divergent members of a family. The adventurous approach is to choose a set of marginally alignable sequences to pursue the difficult task of making a multiple sequence alignment and then to make profile models that may recognize divergence but will also give false predictions. The best method to choose is somewhere between the conservative and adventurous methods. Divergence is necessary, but the sequences chosen should be clearly related based on inspection of each pair-wise alignment and a statistical analysis. Clustering analyses of the sequences can also be useful. Questionable sequences can be left out of the analysis at one stage and added in a second to determine what effect they have on the model.

Between-proteome Comparisons Identify Orthologs, Gene Families, and Domains

Comparisons between proteomes of organisms are illustrated by the third panel in Figure 10.4A. In this analysis, each protein in the proteome is used as a query in a database similarity search against another proteome or combined set of proteomes. When the proteome of an organism is not available, an EST database may be searched for matches, but the type of search is less informative than a full-genome comparison (see below). As in the all-against-all search for paralogs, the search should identify highly conserved proteins of similar domain structure and other similar proteins that show variation in the domain structure as illustrated in Figure 10.4B. A pair of proteins in two organisms that align along most of their lengths with a highly significant alignment score are likely to be orthologs, proteins that share a common ancestry and that have kept the same function following speciation. These proteins perform the core biological functions shared by all organisms, including DNA replication, transcription, translation, and intermediary metabolism. They do not include the proteins unique to the biology of a particular organism.

Other matching sequences in this class could also be orthologs, but could also represent a match between a sequence in proteome A to a paralog of a true ortholog of the sequence in proteome B. In one method designed to identify true orthologs, the most closely related pairs of sequences in proteomes A and B are identified. Two proteins, X in proteome A and Y in proteome B, are predicted to be an orthologous pair if reciprocal searches of proteome A with Y and proteome B with X each produce the highest-scoring match with the other protein. Furthermore, the E value for each alignment should be < 0.01 and the alignment should extend over 60% of each protein (Huynen and Bork 1998).

In another method to identify the mostly closely related sequences in different proteomes, Chervitz et al. (1998) kept only matched sequences with a very conservative P value for the alignment score. The steps for identifying a group of related sequences between the yeast and worm proteomes were as follows:

1. Choose a yeast protein and perform a database similarity search of the worm proteome using WU-BLAST, a yeast-versus-worm search.

2. Group the worm sequences that match the yeast query sequence with a high *P* value (10^{-10} to 10^{-100}) and include the yeast query sequence in the group.
3. From the group in proteome B, choose a worm sequence and make a search of the yeast proteome, using the same *P* value limit as in step 2.
4. Add any matching yeast sequence to the grouping made in step 2.
5. Repeat steps 3 and 4 for all initially matched worm sequences.
6. Repeat steps 1–5 for every yeast protein.
7. Perform a comparable worm-versus-yeast analysis as outlined in steps 1–6.
8. Coalesce the groups of related sequences and remove any redundancies so that every sequence is represented only once.
9. Eliminate any matched pairs in which less than 80% of each sequence is in the alignment.

The above steps locate groups of highly related sequences in two proteomes based on high-scoring alignments among the group. These groups are then subjected to the single linkage cluster analysis described above and illustrated in Figure 10.4C. The analysis creates a multiple sequence alignment and a tree/dendrogram representation of sequence relationships very similar to that produced in a phylogenetic analysis. Orthologs appear as nearest neighbors on the tips of this tree.

The results of the above analysis with the yeast and worm proteomes are shown in Table 10.5. The numbers of sequence groups decrease about fivefold as the stringency of the *E* value of the required scores decreases from 10^{-10} to 10^{-100}, and a similar effect is observed for the subcategories shown in the table. Given that these sequences also align to the extent of 80%, they represent highly conserved sets of genes.

Clusters of orthologous groups. As described above, a pair of orthologous genes in two organisms share so much sequence similarity that they may be assumed to have arisen from a common ancestor gene. When entire proteomes of the two organisms are available, orthologs may be identified. Using the protein from one of the organisms to search the proteome of the other for high-scoring matches should identify the ortholog as the highest-scoring match, or best hit. However, in many cases, each of the orthologs belongs to a family composed of paralogous sequences related to each other by gene duplication events. Hence, in the above database search, the ortholog will not only match the orthologous sequence in the second proteome but also these other paralogous sequences. The objective of the clusters of orthologous groups (COG) approach is to identify all matching proteins in the organisms, defined as an orthologous group related by both speciation and gene duplication events. Related orthologous groups in different organisms are then clustered together to form a COG that will include both orthologs and paralogs. These clusters cor-

Table 10.5. *Numbers of closely related yeast and worm sequences*

Cut-off *P* value	$< 10^{-10}$	$< 10^{-20}$	$< 10^{-50}$	$< 10^{-100}$
Total number of sequence groups	1171	984	552	236
Number of groups with more than two members	560	442	230	79
Number and percent of all yeast proteins (6217) represented in groups	2697 (40)	1848 (30)	888 (14)	330 (5)
Number and percent of all worm proteins represented in groups	3653 (19)	2497 (13)	1094 (6)	370 (2)

Adapted, with permission, from Chervitz et al. 1998 (copyright AAAS).

respond to classes of metabolic function. A database produced by analysis of the available microbial genomes and part of the yeast genome has been made, and a newly identified microbial protein may be used as a query to search this database (see Table 10.1D). Any significant matches found will provide an indication as to the metabolic function of the query protein (Tatusov et al. 1997).

To produce COGs, similarity searches were performed among the proteomes of phylogenetically distinct clades of prokaryotes (see Fig. 6.3, p. 243 for a tree). Orthologous pairs were first defined by the best hits in reciprocal searches. A cluster of three orthologs in three different species was then represented as a triangle on a diagram. Some triangles included a common side, representing the presence of the same orthologous pair in a comparison of four or more organisms. Triangles with this feature were merged into a cluster similar in appearance to Figure 10.4C(i). Paralogs defined by sets of three matching sequences in the selected organisms were also added to these clusters. Paralogs may include a best hit or a high-scoring match of one of the sequences by another, but the reciprocal match can have low similarity that does not have to be significant (Koonin et al. 1998). Sixty percent of the original set of 720 COGs does not include paralogs, or includes paralogs from one lineage only, suggesting that there has not been extensive duplication of this group.

Some of the clusters defined in this manner include proteins having a different domain structure, as illustrated in Figure 10.4B. In other cases, examination of sequence similarity between some pairs of paralogs reveals that a particular paralog has disappeared in a particular lineage. The affected COGs have been modified to reflect more accurately the domain organization of proteins and loss of paralogs. Finally, some additional COGs not represented in the data set were produced by single linkage cluster analysis as described in Figure 10.4C and in the above sections (Tatusov et al. 1997). The proteins encoded by 13 prokaryotic organisms have been analyzed for COG relationships (Koonin et al. 1998). A COG analysis provides an initial assessment of the genome composition of prokaryotic organisms and should be followed by a more detailed analysis as described above for the worm and yeast genomes.

Comparison of proteomes to EST databases of an organism. For many eukaryotic organisms, the complete genome sequence is not available. What is available for some of these organisms is a large collection of EST sequences obtained by random sequencing of cDNA copies of cell mRNA sequences. These sequences are single DNA sequence reads that contain a small fraction of incorrect base assessments, insertions, and deletions. Many sequences arise from near the 3′ end of the mRNA, although every effort is usually made to read as far 5′ as possible into the upstream portion of the cDNA. Because not all of the genes may be expressed in the tissues chosen for analysis, the library will often not be complete. EST libraries are useful for preliminary identification of genes by database similarity searches as described in Chapter 7. A more detailed analysis may then be made by cloning and sequencing the intact cDNA.

An EST database of an organism can be analyzed for the presence of gene families, orthologs, and paralogs. A protein from the yeast or fly proteome, for example, can be used as a query of a human EST database by translating each EST sequence in all six possible reading frames. The program TBLASTN is frequently used for this purpose. The TFASTX and TFASTY programs are designed to accommodate the errors inherent in EST sequences (p. 295). The limitations to whole-proteome searches against EST libraries are that the short length of the translated EST sequence (the equivalent of 100–150 amino acids) will only match a portion of the query protein; for example, a domain or part of domain as illustrated in Figure 10.4B. Hence, it is not possible to impose the requirement of alignment with 60–80% of the query sequence that greatly improved the prediction of

orthologs. Predictions of EST relationships can be improved by identifying overlapping EST sequences so that a longer alignment can be produced, as discussed in Chapter 7. Another method is to perform an exhaustive search for a protein family, described next.

A computer script is a set of computer commands that are placed in a disk file. When the script is run, the commands are executed in the order given by the script. For example, the script may include collecting EST sequence by FTP, analyzing them by TBLASTN or TFASTY, collecting the alignment scores, ordering them, and making charts. The Perl programming language is used for producing such scripts.

Searching for orthologs to a protein family in an EST database. Searches of EST databases for matches to a query sequence routinely produce large amounts of output that must be searched manually for significant hits. Retief et al. (1999) have described an automatic method utilizing a computer script, FAST-PAN, that scans EST databases with multiple queries from a protein family, sorts the alignment scores, and produces charts and alignments of the matches found. An example of using this method is shown in Figure 10.5. A chart showing the *E* value, percent identity, fraction of query sequence matched, and type of query matched (color coded) is shown in Figure 10.5A.

In an example by Retief et al. (1999), the large family of known glutathione transferase proteins was first subjected to multiple sequence alignment, and a phylogenetic tree was made by distance methods to identify classes of proteins within the family. These proteins represented a broad range of phylogenetic context and included classes with sometimes less than 20% identity. The object was to choose class representatives for a similarity search of mammalian EST databases for paralogs and to decide which of these sequences were orthologs.

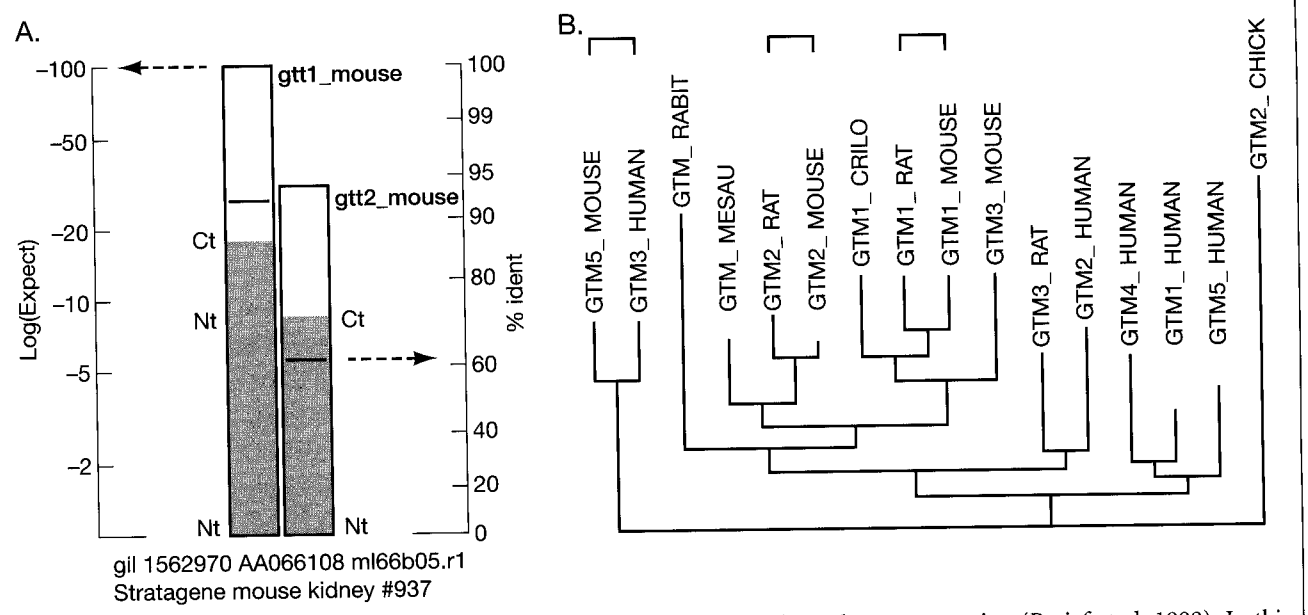


Figure 10.5. Prediction of paralogs and orthologs by searches of EST databases by gene panning (Retief et al. 1999). In this analysis, one class of glutathione transferase family members was used as queries to search mammalian EST databases for highly significant matches using TFASTY3 (Chapter 7). FAST_PAN is a Perl-script program (see Table 10.1D) that automatically searches EST databases as they are updated and compiles the results of the search. (*A*) Display of protein class matched (color), log Expect value (height of bar), length of query sequence matched (height of color bar), and percent identity (position of horizontal line in bar) on one graph as produced by FAST_PAN. Note that the log scales clearly reveal the lowest *E* value and highest identity matches. Shown are matches of two mouse ESTs to a query sequence. (*B*) Example of phylogenetic analysis to predict orthologs between species (bracketed). Amino acid sequences of ESTs in the matched regions were aligned, and this alignment was then used to direct an alignment of the EST codons. A phylogenetic tree was produced by the aligned EST sequences by the maximum likelihood method using the program DNAML in the PHYLIP package. As discussed by the authors, this method allows researchers to search rapidly and easily through EST databases to identify matching sequences and to examine the quality of the alignments found. In this example, a large number of glutathione transferase members were used as queries, allowing an exhaustive search of the EST database for representative family members. (Redrawn, with permission, from Retief et al. 1999.)

A novel feature of these searches was to use a lower-scoring PAM matrix to search for paralogs of a recently evolved group of sequences. Use of an appropriate PAM matrix that matches the expected evolutionary separation of a group of sequences provides an improved higher-scoring alignment, as described in detail in Chapter 3 (p. 82). ESTs with a high percent identity with the query sequence, a long alignment with the query sequence, and a very low *E* value of the alignment score represent groups of paralogous and orthologous genes. To identify orthologs as the most closely related sequence, ESTs were aligned using the amino acid alignment as a guide, and a phylogenetic tree was produced by the maximum likelihood method. This method, described in Chapter 6 (see flowchart for Chapter 6), is suitable for a divergent but recently evolved group of sequences. The predicted tree shown in Figure 10.5B predicts those pairs of sequences that are likely to be orthologous.

Family and Domain Analysis

As shown in the flowchart (p. 492), gene identification of predicted proteins in the genome is designed to discover the metabolic features of an organism. An important feature of proteins discussed in Chapter 9 is their organization into domains that represent modules of structure and function. Different proteins are mosaics of domains that occur in different combinations in a given protein. In a particular organism or group of organisms, one particular domain can be expanded to perform a particular function. Comparison of the domain content of an entire proteome with that of another proteome can reveal the biological roles of diverse domains in different organisms. Extensive comparisons for both prokaryotic and eukaryotic genomes have been performed (Chervitz et al. 1998; Huynen and Bork 1998; Rubin et al. 2000). A descriptive list of protein domain databases that may be used for such an analysis is given in Table 9.5. In a detailed analysis of the fly, worm, and yeast proteomes, 744 families and domains were common to all three organisms. More than 2000 fly and worm proteins are multidomain proteins, compared to about one-third this number in yeast (Rubin et al. 2000). Tekaia et al. (1999) have introduced the concept of a genome tree. A tree or dendrogram based on the proportion of proteins in one organism that is shared by another organism is produced by the single linkage clustering method described in Figure 10.4C.

Ancient Conserved Regions

Phylogenetically diverse groups of organisms have been analyzed for the presence of conserved proteins and protein domains that have been conserved over long periods of evolutionary time, called ancient conserved regions or ACRs (Green et al. 1993). The method involves database similarity searches of the SwissProt database with human, worm, yeast, or *E. coli* genes and identification of matches with sequences from a different phylum than the query sequence. An analysis of ACRs that predate the radiation of the major animal phyla some 580–540 million years ago suggested that 20–40% of coding sequences are ACRs. For example, a search with 1916 *E. coli* proteins detected 266 ACRs found in 439 sequences, roughly one-quarter of the SwissProt database. These ACRs may represent proteins present at the time of the prokaryotic–eukaryotic divergence.

With the later addition of complete genome sequences of phylogenetically diverse prokaryotic organisms, the number of ACRs may be estimated by the proportion of genes that match database sequence of known function. For the hyperthermophilic archaea *Pyro-*

coccus hirokoshii (Kawarabayasi et al. 1998), this proportion was 20%, perhaps representing an ancient set of prokaryotic ACRs. COGs described above represent sets of proteins that are conserved across distant phylogenetic lineages. For 11 prokaryotic genomes, the proportion of genes represented in COGs is approximately 50% (Koonin et al. 1998), and other studies suggest that as many as 70% of prokaryotic genomes contain ACRs (Koonin and Galperin 1997). However, one needs to take into account that horizontal transfer of genetic material discussed below increases the sharing of genes by different lineages of prokaryotes.

Horizontal Gene Transfer

The genomes of most organisms are derived by vertical transmission, the inheritance of chromosomes from parents to offspring from one generation to the next. However, in rare instances, genomes may also be modified by horizontal (sometimes called lateral) gene transfer (HT), the acquisition of genetic material from a different organism. The transferred material then becomes a permanent addition to the recipient genome. Although these exchanges do not occur very often on a generation-to-generation basis, a significant number can occur over a period of hundreds of millions of years. An extreme example is the proposed endosymbiont origin of mitochondria in eukaryotic cells and chloroplasts in plants. The endosymbiont theory proposes that these organelles were transferred from free-living bacteria to another organism with which they shared a symbiotic relationship (see Chapter 6 in Brown 1999).

Horizontal gene transfer is a significant source of genome variation in bacteria (for review, see Ochmann et al. 2000), allowing them to exploit new environments. Such transfer is rendered possible by a variety of natural mechanisms in bacteria for transferring DNA from one species to another. Detection of HT is made possible by the fact that each genome of each bacterial species has a unique base composition. Hence, transfer of a portion of a genome from one organism to another can generally be detected as an island of sequence of different composition in the recipient. If the amino acid composition of transferred genes is typical, these islands may be detected by a codon usage analysis as described in Chapter 8. Very ancient transfers may not be detectable because the base composition and codon usage of the transferred DNA will eventually blend into those of the recipient organism. The time of transfer of DNA may be estimated by the degree to which the composition of the HT DNA has blended into that of the recipient genome. Comparisons of completely sequenced bacterial genomes have revealed that they are mosaics of ancestral and horizontally transferred sequences. The proportion of the genome due to HT sequences also varies considerably roughly in proportion to genome size. A total of 12.8% of the genome of *E. coli* is due to HT DNA (the highest level found), whereas it is 0.0% in *Mycoplasma genitalium*, whose genome is less than one-quarter the size of that of *E. coli*. Mycoplasma have lost many of the genes needed to be a free-living organism and instead depend on nutrients provided by the interior of the host cell. Hence, these organisms would not be expected to carry any extra unnecessary genetic baggage. HT DNA contributes in a major way to the disease-producing ability of pathogenic bacteria, and this DNA frequently has flanking direct repeats characteristic of transposable elements. Note that when genes are clustered on the chromosome of the donor organism (described below), the recipient organism may gain an entire metabolic pathway from another by means of horizontal transfer. Hence, clustering in combination with horizontal transfers provides an evolutionary mechanism for altering metabolic pathways in diverse organisms.

Gene Annotation

Accurate annotation of genome sequences is an important first step in genome analysis. As described earlier, annotation is based on finding significant alignment to sequences of known function in database similarity searches. Matches of lesser significance provide only a tentative or hypothetical prediction and should be used as a working hypothesis of function (see Kyrpides and Ouzonis 1999). Computational tools such as MAGPIE and GENEQUIZ described below are designed to assist with accurate genome annotations.

FUNCTIONAL CLASSIFICATION OF GENES

Once sequences have been annotated, a useful next step is to classify the annotated genes by function. Genes that are significantly similar in an organism, i.e., paralogous sequences, frequently are found to have a related biological function. This discovery follows the expected origin of paralogs by gene duplication events, leaving one copy to perform the original function and producing a second copy to develop a new function not too distant from the original one under evolutionary selection. An early classification scheme for eight related groups of *E. coli* genes included categories for enzymes, transport elements, regulators, membranes, structural elements, protein factors, leader peptides, and carriers. Ninety percent of *E. coli* genes related by significant sequence similarity fell into these same broad categories (Labedan and Riley 1995).

The Enzyme Commission numbers formulated by the Enzyme Commission of the International Union of Biochemistry and Molecular Biology provide a detailed way to classify enzymes based on the biochemical reactions they catalyze (Webb 1992; Tipton and Boyce 2000). The designation ECa.b.c.d gives the following information: (a) one of six main classes of biochemical reactions, (b) the group of substrate molecule or the nature of chemical bond that is involved in the reaction, (c) designation for acceptor molecules (cofactors), and (d) specific details of the biochemical reaction. Using this system to compare sequence-related pairs of *E. coli* genes, Labedan and Riley (1995) found that 70% of them shared the first two EC designators (a and b) in the annotation of the corresponding genes, thereby indicating that they catalyze biochemically similar reactions. A third measure of functional similarity is based on a physiological characterization of *E. coli* proteins into 118 possible categories (e.g., DNA synthesis, TCA cycle, etc.) (Riley 1993). Approximately one-quarter of *E. coli* genes fall into the same category by this scheme.

An alternative approach to classification of genes that encode enzymes is to examine relationships among multiple enzymes that perform the same biochemical function in the same organism. Although catalyzing the same reaction, these enzymes showed variations in metabolic regulation of their activity. More than one-half of multiple enzymes in *E. coli* share significant sequence similarity; i.e., they are paralogs. However, the remainder do not share any sequence similarity. Either they were acquired by horizontal transfer from another bacterial species or the two enzymes were formed by convergent evolution from two different genetic starting points (Riley 1998). Accordingly, sequence similarity is frequently a good indicator of related biochemical function, but two enzymes that perform the same biochemical task may not share sequence similarity of evolutionary history.

Other functional classification schemes for genes include a broader category for genes involved in the same biological process, e.g., a three-group scheme for energy-related,

information-related, and communication-related genes has also been used. By this scheme, plants devote more than one-half of their genome to energy metabolism, whereas animals devote one-half of their genome to communication-related functions (Ouzounis et al. 1996). Another scheme, described below, is to identify proteins that physically interact in a structure or biochemical pathway.

A system for functional annotation of the yeast genome has also been produced (Cherry et al. 1997) and used in a comparison of the yeast and worm proteomes (see SGD, Table 10.4B) (Chervitz et al. 1998). *D. melanogaster* genes were classified using the Gene Ontology (GO) classification scheme (Adams et al. 2000), a collaboration among yeast, fly, and mouse informatics groups to develop a general classification scheme useful for several genomes (see GO site, Table 10.1F). This classification scheme provides a description of gene products based on function, biological role, and cellular location.

GENE ORDER (SYNTENY) IS CONSERVED ON CHROMOSOMES OF RELATED ORGANISMS

Two species that have recently diverged from a common ancestor might be expected to share a similar set of genes and also similar chromosomes with these genes positioned along the chromosomes in the same order. Over evolutionary time, the sequence of each pair of genes will slowly diverge, as the species diverge and other changes such as gene duplication and gene loss change the gene content. In addition, the order of genes also changes over evolutionary time as a result of chromosomal rearrangements. These rearrangements may be modeled by occasional chromosomal breaks, random with respect to chromosomal location, and by random rejoining of the fragments by a DNA repair mechanism. Rearrangements may be analyzed by comparing the location of orthologs, genes of highly conserved sequence and function in prokaryotic and eukaryotic proteomes from different phylogenetic lineages.

Colinearity of gene order is referred to as synteny, and a conserved group of genes in the same order in two genomes as a syntenic group or cluster.

Two important observations have been made with regard to gene order: First, order is highly conserved in closely related species but becomes changed by rearrangements over evolutionary time. As more and more rearrangements occur, there will no longer be any correspondence in the order of orthologous genes on the chromosome of one organism with that of a second organism. Second, groups of genes that have a similar biological function tend to remain localized in a group or cluster. Examples of these observations and their significance are described below.

Chromosomal Rearrangements

In Figure 10.6, a genome plot of the positions of orthologs and paralogs on the genomes of two related bacteria, *Mycoplasma pneumoniae* and *Mycoplasma genitalium*, both human pathogens, is shown (Himmelriech et al. 1997). This plot is very similar to the dot matrix plot used for sequence alignment (see Chapter 3), except that in this case a dot or symbol is shown at the intersection of the position of one member of an orthologous pair of sequences on genome 1 and the position of the other member of the pair on genome 2. The plot clearly shows that large sections of chromosome are conserved but also that a number of rearrangements have occurred, making the gene order different from that of the other genome and from the common ancestor of these two organisms. In contrast, a similar plot of orthologous genes in the genomes of the bacterial species *E. coli* and *H. influenzae* appears quite random (Tatusov et al. 1996), even though the organisms are only slightly more distant in evolution than the two *Mycoplasma* species. However, on close inspection of gene function and order, similarities can be found. By classifying genes using a nine-

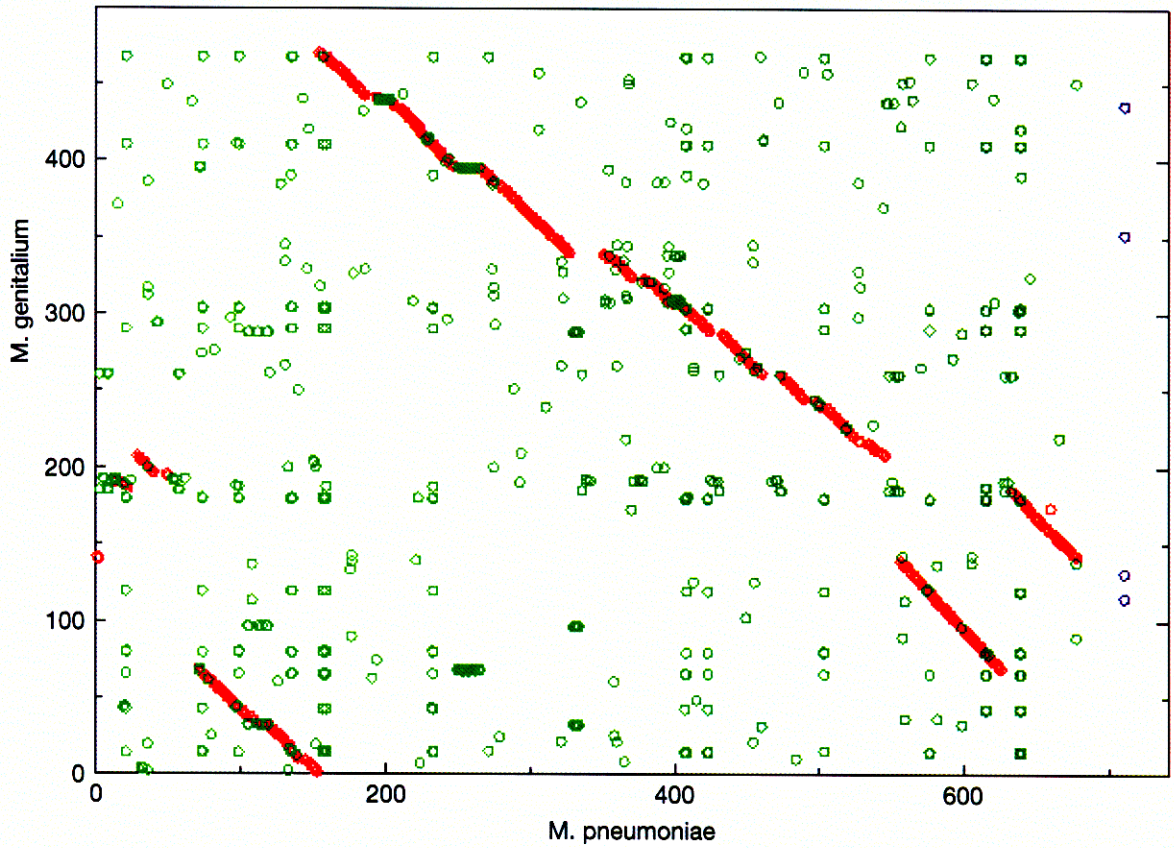


Figure 10.6. Genome plot of orthologous genes. Alignment of orthologous and paralogous genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* (Table 10.1A, comparative genome analysis in P. Bork laboratory). Horizontal axis is genome position in *M. pneumoniae*, vertical axis is genome position of *M. genitalium*. Positions of orthologs are shown in red, paralogs in green. Orthologous genes are in the same order in both genomes except for several chromosomal rearrangements. These genes are defined by high *E* values in database searches in which one of an orthologous pair is used as query of the proteome of the other species. Proteins should also align along 60% of the length of each (Huynen and Bork 1998). Paralogs are proteins that have striking, high-scoring similarity but are not the highest scoring in reciprocal proteome searches. Note also the occurrence of paralogs within the conserved stretches of orthologs, presumably representing gene duplication in these regions. In contrast to this conserved order of gene position in the *Mycoplasma* species, the orthologous genes in two other equally related species, *E. coli* and *Hemophilus influenzae*, show no detectable conservation of order on a similar genome plot.

class functional classification scheme (see above), several genes falling into the same functional category are clustered together on the chromosomes of both of these organisms, and the clusters are in a similar order (Ouzounis et al. 1996). Comparison of the number of rearrangements in a given period of evolutionary history may vary significantly from one organism to the next. In one analysis of prokaryotic organisms of diverse phylogenetic origin, it has been shown that if gene A has a neighboring gene B, then if an ortholog of A occurs in another genome, there is an increased probability of an ortholog of B also occurring in the other organism. However, the B ortholog is less likely to be a neighbor of the A ortholog of the genome of the second species if the two species are more divergent (Huynen and Bork 1998).

The TIGR Web site (Table 10.1D) includes a resource for comparing any two prokaryotic genomes of the 30+ available by means of a genome plot, as shown in Figure 10.6. In general, the order of orthologs is not well conserved in prokaryotes when the genomes have diverged sufficiently that the orthologs have <50% identity (Huynen and Bork 1998).

A similar conservation of gene order also appears to be present in closely related eukaryotic genomes. The evidence is based on chromosome painting experiments in which DNA from a section of a chromosome of one organism is labeled and then hybridized to chromosomes of a second organism. Regions of the second chromosome that are labeled reveal the presence of a homologous region. Although this method does not have the precision and sensitivity of sequence analysis methods, these experiments reveal that eukaryotic chromosomes also undergo rearrangements both within chromosomes and between chromosomes during evolution. An example of the differences between mouse and human chromosomes is shown in Figure 10.7. A much larger data collection from a variety of mammalian chromosomes suggests that each chromosome is a mosaic of a similar set of ancestral fragments (O'Brien et al. 1999). Similar studies with plant genomes have also indicated that they have a similar overall gene content but that many regional duplications and rearrangements have occurred during evolution (Bennetzen 1998, 2000; Bennetzen et al. 1998). The availability of genome sequences of plants and animals offers some exciting opportunities for determining the chromosomal changes that have occurred during evolution of the plant and animal kingdoms.

Computational Analysis of Gene Rearrangements

As genome-by-genome comparisons of the chromosomes of related species are made and the rearrangements are discovered, a further challenge to computational and evolutionary biologists is to estimate the number and types of rearrangements that have occurred and also to determine when they occurred. For example, a comparison of the mouse and human chromosomes reveals many rearrangements (Fig. 10.7). A computational approach to these questions is outlined in Figure 10.8. In aligning gene and protein sequences, one assumes a model in which no rearrangements have occurred so that lines can be drawn between the corresponding positions in the sequences and no lines will cross or intersect, as shown in Figure 10.8A. For comparing gene orders on chromosomes that have undergone rearrangements, lines joining the corresponding genes will intersect, as shown in Figure 10.8B, and the greater the amount of rearranging, the greater the number of intersects. In the random shuffling model, one tries to estimate the number of rearrangements that produces the observed number of intersections and to compare this number to one that would randomly shuffle the same fragments. The analysis shown in Figure 10.8C attempts to reconstruct the number and types of rearrangements (inversions, etc.) that have given rise to the observed variation in gene order between the chromosomes.

Clusters of Genes on Chromosomes Have a Metabolically Related Function

In a given organism or species, genes are found in a given order that is maintained on the chromosomes from one generation to the next. Genetic analysis has revealed that genes with a related function are frequently found to be clustered at one chromosomal location. Clustering of related genes presumably provides an evolutionary advantage to a species, but the underlying biological reason is not understood. One possibility is that there is genetic variation (alleles) within each gene in a cluster of a given species and that only certain allelic combinations of different genes are compatible. Another possibility is some kind of coordinated translation of the proteins that may aid their folding. In the model bacterial species *E. coli*, genes that act sequentially in a biochemical pathway are frequently found to be adjacent to each other at one chromosomal location. For example, the genes required for synthesis of the amino acid tryptophan (*trp* genes) are clustered together on the chromosome of *E. coli*, as illustrated in Figure 10.8, where their expression is coordi-

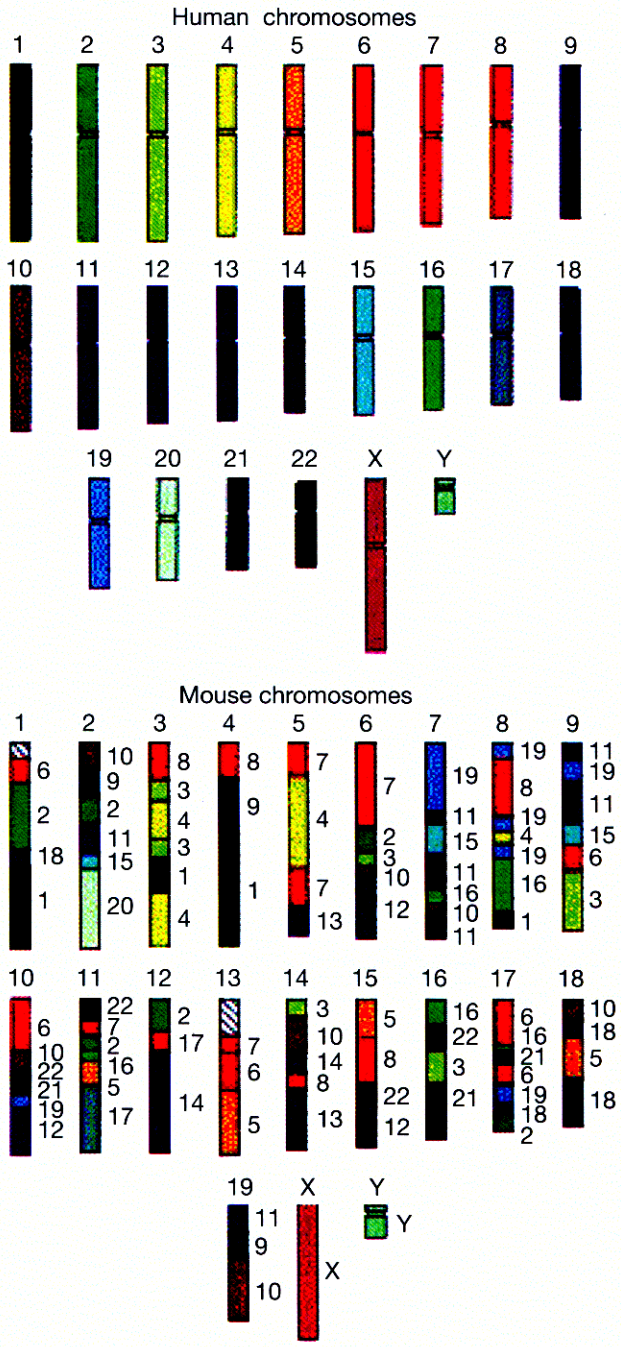


Figure 10.7. Similarity between local gene clusters in human and mouse chromosomes. Human chromosomes can be cut into >100 pieces and reassembled into a reasonable facsimile of the mouse chromosome. Only larger fragments are represented. Chromosomes of all mammals may share a similar relationship (O'Brien et al. 1999).

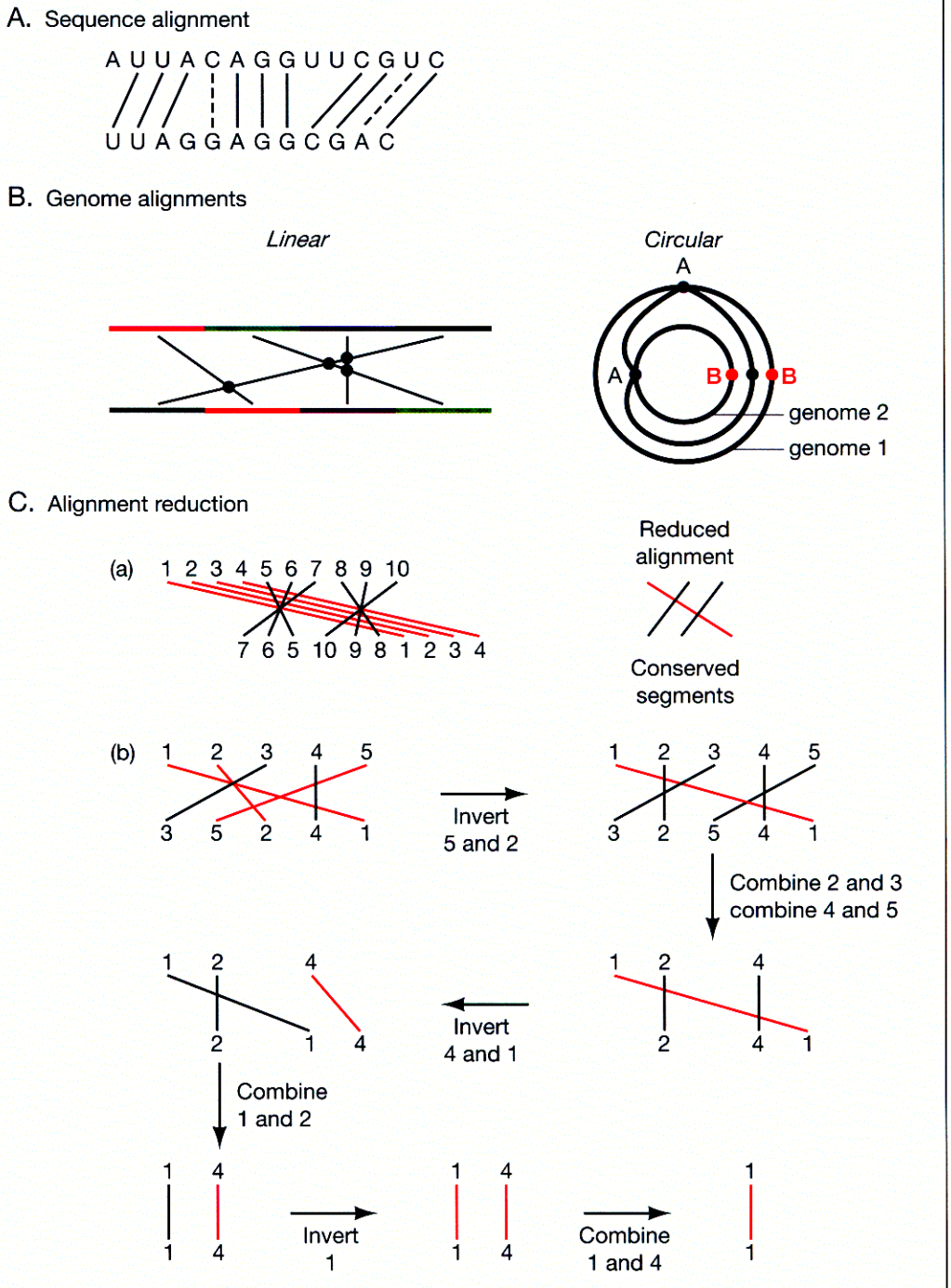


Figure 10.8. Computational analysis of genome arrangements. (*A*) In aligning two sequences, one sequence is written above the other and the highest number of consecutive matches between the sequences provides an optimal alignment as described in Chapter 3. The alignment includes matches (*solid lines*), mismatches (*dotted lines*), and insertions/deletions in order to produce an optimal number of matches. The matches are in a consecutive order in two sequences such that no rearrangements would be found. (*B*) Alignments of linear and circular chromosomes that have undergone rearrangements such as those found in mammalian chromosomes and mitochondria. In contrast to sequence alignment, lines indicating homologous positions in linear chromosomes (*left*) now cross, producing points of intersection. The more rearrangements there are, the more intersections will occur. For alignment of circular chromosomes (*right*), depending on how the chromosomes are aligned, there are two ways of showing a moved region. To go from A on the outer genome to A on the inner genome,

nately regulated by a common promoter. This coordination of expression avoids wasteful production of one enzyme when others in the same pathway are not available.

With the availability of other prokaryotic genome sequences, important metabolic genes such as *trp* can be identified in these species, and the chromosomal location of these genes can be compared with that of *E. coli*. Using the predicted tryptophan genes as an example (Fig. 10.9), the following observations were made: (1) At least some of the *trp* genes are also clustered together on the chromosomes of other species of Bacteria and Archaea; (2) the order of the genes within the cluster is conserved within the first four species listed in Figure 10.9, all of which are bacteria; (3) the order is much less conserved in the last three species, all of which are Archaea, and some of the genes have been moved to a more distant location; (4) there are multiple examples of gene fusions that give rise to a new protein that performs both biochemical functions of the single-gene, parent proteins. *trpC* has been fused independently with two other genes, *trpD* and *trpF*. Alternatively, a composite gene may produce two smaller single-component genes by fission of a parent composite gene. Fission events have only been observed in thermophiles among prokaryotes (Snel et al. 2000a). However, biochemical reasons have been presented that fission events may provide a mechanism for evolution of protein complexes (Marcotte et al. 1999b).

The term clusters has been used in two different ways in the literature and in this chapter, and the two should not be confused. One use is to represent groups of genes in one or several organisms that share a significant degree of sequence similarity. An example is Figure 10.4C. A second use of clusters is to represent a physical clustering of genes on the same chromosome. An example is the arrangement of the trp *genes in Figure 10.8.*

When a series of predicted genes in a known *E. coli* pathway is in the same order in another organism as in *E. coli*, e.g., *trpB-trpA* and *trpE-trpG* in the Archaea in Figure 10.9, then the same biochemical pathway is predicted also. Even if the genome annotation is based on a weak prediction of the biochemical function of two individual genes, the prediction is stronger if the two genes act in the same pathway and is strongest if the genes are clustered (Huynen et al. 2000). In the *trp* example shown in Figure 10.9, the presence of the genes in such a phylogenetically diverse group of organisms indicates that the pathway is an ancient one. Clustering of the genes further indicates that they probably originated as a group in the single chromosomal region of an ancient ancestor organism, assuming there has not been a driving force for repeated independent clustering events. What is also revealed in the *trp* example in Figure 10.9 is that some *trp* genes are found at a much more remote chromosomal location. The diverse location of the *trp* genes in *Methanococcus jannaschii* is an outstanding example. Apparently, rearrangements can break clusters and

the line joining them can go clockwise or counterclockwise and, as a result, there will be either 0 or 1 intersections with the line joining B. The complexity of alignments of circular chromosomes is reduced by limiting the joining lines to 180 degrees of relative genome positions. Sankoff and Goldstein (1989) devised a shuffling model for estimating the number of rearrangements when the number of intersections is known. The method is analogous to shuffling an ordered deck of cards and then predicting how much order remains. Eventually, after $n \log n$ shuffles, where n is the number of cards, the order becomes random. Given an observed remaining order, how many shuffles have occurred? The number of observed intersections is compared to the number expected for completely shuffled genomes (Sankoff et al. 1993). (*C*) Another method for determining numbers of rearrangements is to assume that they have occurred by a number of transposition or recombination events. The object of this analysis is to try to identify the rearrangements that occurred and then to undo (or derange) the alignments accordingly. The goal is to minimize the number of rearrangements, this number then representing a genetic distance between the sequences. (*a*) Alignments of genes 1–10 in two genomes where some genes are in the same order (*red lines*) and others are inverted (*blue lines*). Groups of genes such as the two joined by the blue lines may be combined into a single unit representing a conserved segment since no recombination event would be required. (*b*) Alignment that can be accounted for by these inversion events. The program DERANGEII is available from the authors and FTP from ftp.ebi.ac.uk/pub/software/unix/derange2.tar.Z. These methods have been used to analyze rearrangements in mitochondrial and bacterial genomes (Sankoff et al. 1992; Blanchette et al. 1996; Sankoff and Nadeau 1996) and additional algorithms have also been developed (Kececioglu and Sankoff 1995; Kececioglu and Gusfield 1998). (Adapted from Sankoff et al. 1992, 1993.)

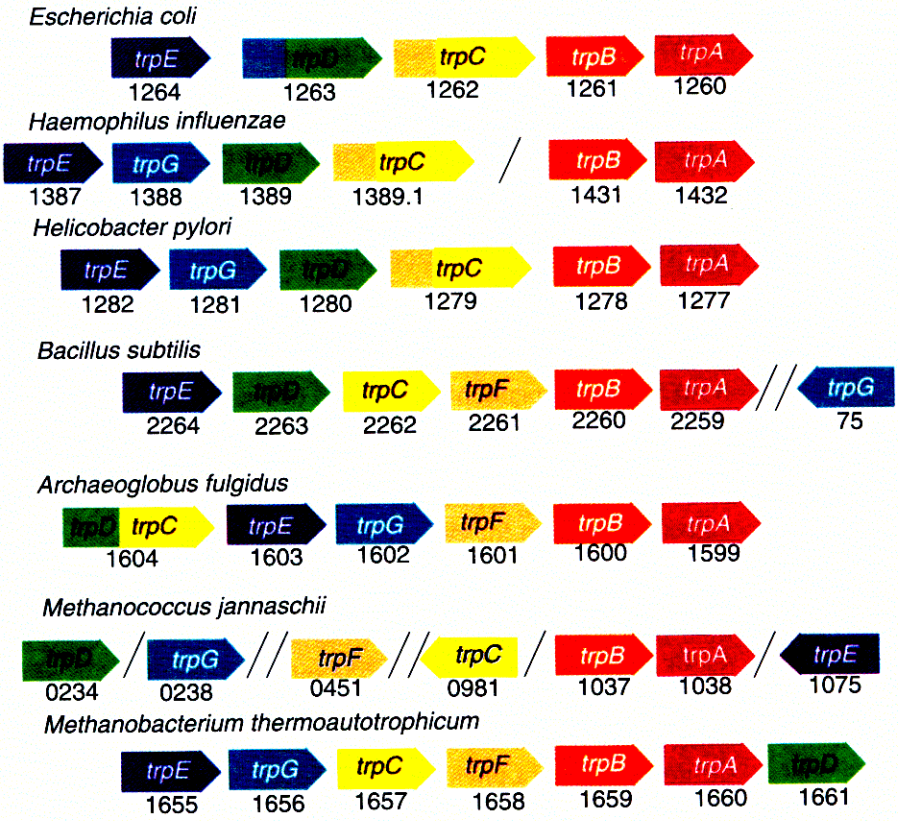


Figure 10.9. Structure of tryptophan operon in different prokaryotic organisms. Numbers indicate gene number in genome; arrows indicate direction of transcription; double lines indicate a separation of more than 50 genes due to dispersion of the operon. Shown also are examples of gene fusion so examples of domain fusions (e.g., *trpD* and *trpG*) are fused in *E. coli*. Note that only the *trpA* and *trpB* genes are genetically linked and separate genes in all of the species. (Reprinted, with permission, from Dandekar et al. 1998 [copyright Elsevier Science].)

move genes to other locations, although another possibility is that the dispersed arrangement is a more ancestral state.

Two methods have been described for identifying clusters or coordinately regulated genes. In one study with three separate groups of three distantly related prokaryotes (Dandekar et al. 1998), approximately 100 genes were found to be conserved as a cluster of two pairs. (Looking for a pair in three species avoided possible complications from horizontal transfer.) The direction of transcription was the same for all genes, implying a regulatory relationship as in an operon. For approximately 75% of the genes, a physical interaction between the genes had previously been demonstrated and could be predicted for almost all proteins based on additional sequence comparisons. These conserved proteins have core biological functions such as transcription, translation, and cell division.

In a second method (Overbeek et al. 1999), a full reciprocal search like that used in Figure 10.4 for comparing yeast, worm, and fly genomes and for making COGs was performed between the proteomes of two prokaryotes: Each protein of one proteome was used to search the proteome of the second. Protein pairs that gave a best hit with the other genome and that had an *E* value of less than 10^{-5} were identified, called a bidirectional best

hit (BBH). Pairs of close bidirectional best hits (PCBBH) that are within 300 bp of each other on the chromosomes of the respective organisms and that are transcribed from the same strand, i.e., are in a "typical" operon, were then identified. A score for these pairs was formulated that is higher when the number of organisms in which the pair is observed is greater and the phylogenetic distance between the organisms is larger. Forty percent of a set of higher-scoring pairs corresponded to proteins that are known to act in a common metabolic pathway, as defined in metabolic function databases (see Table 10.1D). Hence, a significant proportion of the pairs of PCBBH correspond to genes that have a related function and lie on the same pathway. This same approach could play an important role in assigning a function to uncharacterized genes in genomes based on proximity to other genes of known function.

Composite Genes with a Multiple Set of Domains Predict Physical Interactions and Functional Relationships between Protein Pairs That Share the Same Domains

As illustrated in Figure 10.9, single *trp* genes can be fused into larger composite genes. Observation of such evolutionary events provided a major step forward in understanding relationships among the proteins of diverse organisms (Enright et al. 1999; Marcotte et al. 1999b). The occurrence of a fused or composite gene in one organism is called a "Rosetta Stone sequence" because it provides evidence that the single component genes in a separate organism encode proteins that physically interact (Marcotte et al. 1999b). For example, if a composite human gene has two domains A and B, the analysis assumes that A and B physically interact within the protein. If two separate genes in other organisms (yeast or *E. coli*) make two proteins, one with domain A and a second with domain B, then these two proteins are assumed to interact because A and B interact. These sequence relationships may be found by sequence alignment of the composite AB protein with each of the single-component A and B proteins. However, A and B will not align with each other. If A and B do not interact in composite proteins, the prediction is a false-positive result. However, these proteins are still predicted to have related functions based on the gene fusion result.

Composite proteins were found by searching SwissProt for statistically significant matches to domains in the ProDom domain database (see Table 9.5, p. 430). Six percent of the Rosetta Stone proteins were found to be represented in the DIP database of interacting proteins (see Table 9.5). Rosetta Stone predictions of interacting proteins were compared to predictions by another method for predicting related proteins, the phylogenetic profile method (Pelligrini et al. 1999; see also "bag of genes" concept in Huynen and Bork 1998). This method is based on the assumption that proteins that function together in a biochemical pathway should evolve in a correlated fashion. Databases are searched for significant matches to two proteins A and B. If A and B have related functions, they should be found together in a large proportion of genomes, whereas if they do not, they will be found to have a random association in genomes.

Enright et al. (1999) used reciprocal searches among three complete prokaryotic proteomes, as described above in Figure 10.4, and identified related proteins that have the expected alignments for composite (AB) and component (A or B) proteins. These proteins interact functionally, act in the same biochemical pathway, or are coregulated. Predictions are stronger when component proteins (A and B) have few paralogs, since the interacting pair can be more readily identified. Conversely, the presence of paralogs of the composite proteins increases the strength of the prediction because the number of possible interactions is increased (Enright et al. 1999).

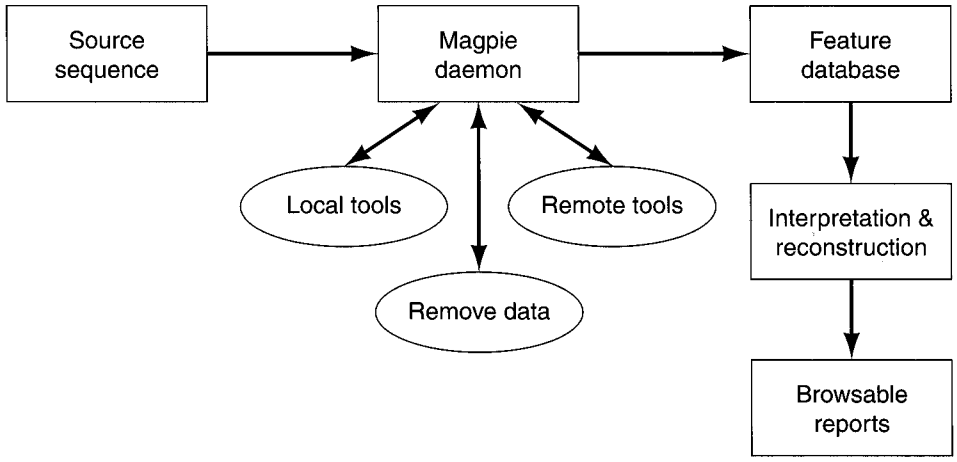


Figure 10.10. Automatic analysis of genome sequences with the MAGPIE system. The sequence is input into an automatic system (a daemon) that sends the sequence to local and remote resources for analysis (BLAST search, PROSITE search, etc.). The information retrieved is stored in a feature database, and the data are interpreted by a set of rules and placed in Web-browsable reports. (Redrawn, with permission, from Gasterland and Sensen 1998 [copyright Kluwer Academic/Plenum].)

Resources for Genome Analysis

The above types of analyses depend on a labor-intensive annotation of the genome and functional analysis of the predicted proteins. Computational tools have been made available to automate some of these steps. Examples are MAGPIE and GeneQuiz, listed in Table 10.1F.

MAGPIE analyzes the genome using a set of automated processes that are illustrated in Figure 10.10. Designed for high-throughput genome sequence analysis, MAGPIE automatically annotates genomic sequence data and maintains a daily up-to-date record in response to user queries about one or more genomes. The system also uses a set of rules in logic programming to make decisions that may be used to interpret information from various sources. MAGPIE has been used to locate potential promoters, terminators, start codons, Shine-Dalgarno sites, DNA motif sites, co-transcription units, and putative operons in microbial genomes. These sites are shown on a map display of the genome that may be edited.

GeneQuiz is an integrated system for large-scale biological sequence analysis that uses a variety of search and analysis methods using current sequence databases. By applying expert rules to the results of the different methods, GeneQuiz creates a compact summary of findings. It focuses on deriving a predicted protein function, based on a variety of available evidence, including the evaluation of the similarity to the closest homolog in a database.

GLOBAL GENE REGULATION

One way to obtain useful information about a genome is to determine which genes are induced or repressed in response to a phase of the cell cycle, a developmental phase, or a response to the environment, such as treatment with a hormone. Sets of genes whose expression rises and falls under the same condition are likely to have a related function. In addition, a pattern of gene expression may also be an indicator of abnormal cellular regulation and is a useful tool in cancer diagnosis (see, e.g., Golub et al. 1999; Perou et al. 1999). Because genomes, especially eukaryotic genomes, are so large, a new technology has been developed for studying the regulation of thousands of genes on a microscope slide.

Microarray (or microchip) analysis is a new technology in which all of the genes of an organism are represented by oligonucleotide sequences spread out in an 80 × 80 array on microscope slides, but can also be synthesized directly on the slide at densities of up to one million per square centimeter. The oligonucleotides are collectively hybridized to a labeled cDNA library prepared by reverse-transcribing mRNA from cells. The amount of label binding to each oligonucleotide spot reflects the amount of mRNA in the cell. The analysis of the data collected in this type of experiment is depicted and described in Figure 10.11. Genes that are responding the same way to an environmental signal, in this case the addition of serum to serum-starved skin cells, are clustered together in a display. From this analysis, a set of genes that responds in an identical manner may be identified. Automatic methods for clustering related sets of genes have been devised, and three representative methods are shown and described in Figure 10.12. The first of these methods, hierarchical clustering (Eisen et al. 1998), is commonly used, but the other two methods are better designed to detect differences in patterns over a set of time points or samples. The derivation of clustering algorithms for microarray analysis has become an active area of bioinformatics.

Once a set of genes that are coregulated has been found, the promoter regions of these genes may be analyzed for conserved patterns that represent sites of interaction with specific transcription factors. This type of analysis is described in detail in Chapter 8 (Table 8.6, p. 370), and additional resources are given in Table 10.1E.

Microarray analysis is designed to detect global changes in transcription in a genome but does not provide information about the levels of protein products of the genes, which may also be subject to translational regulation. Labeled protein samples may also be extracted from treated cells and separated by two-dimensional gel electrophoresis. The proteins are first separated in a column on the basis of size and then across a second dimension on a slab on the basis of charge. The amount of protein in each spot is then determined. This method also can resolve thousands of proteins based on size and charge. There are databases of the patterns found in different organisms; these are listed in Table 10.1E. The technology can also be extended to further purification and microsequencing of the protein spots or of proteins in complexes so that the genes encoding the protein may be identified by proteome similarity searches.

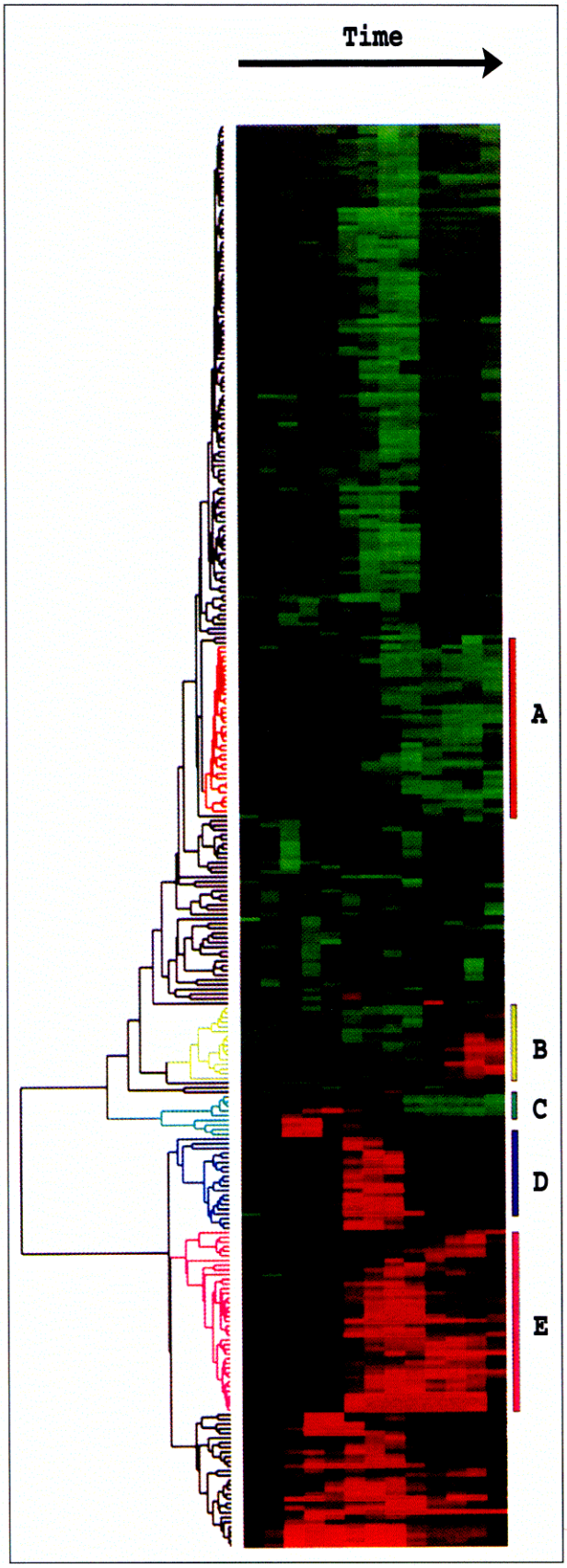

Time
A
B
C
D
E

Figure 10.11. Example of cluster analysis of microarray data. Rows represent changes in an individual hybridization signal for a single gene on a cDNA microarray display system. Columns show changes in the expression of a selected 9800 human cDNA set. These genes change their level of expression in human skin fibroblasts that have been deprived of serum growth factors and serum then added back (time 0). The time points vary from 0 to 24 hours, left to right, and the last column is a control. RNA was removed from cells and the amount was measured by quantitative reverse transcription in the presence of the fluorescent dye Cy5. A reference time 0 sample was labeled in parallel with the green fluorescent dye Cy3 and mixed with samples taken at a later time. The labeled cDNA preparations were then hybridized to the cDNA microarray and the Cy5/Cy3 fluorescence ratio of each spot was measured. Each ratio is expressed as a log odds ratio to the base 2. Thus, a value of +4 at time t indicates 16 times more mRNA at time t than at time 0; 0 means no change and −4 means 16 times less RNA at time t than at time 0. Tables of these raw data are kept (see http://rana.stanford.edu/clustering). The color display in the figure varies from saturated green (log odds − 3.0) to saturated red (log odds + 3.0) with black as the intermediate color (log odds 0). The dendrogram on the right of the color display was made by a hierarchical clustering algorithm that is similar to the single-linkage cluster analysis described in Fig. 10.4. The object of clustering is to identify genes that respond the same way to the environmental treatment. Each gene is compared to every other gene and a gene similarity score (metric) is produced. If X_i is the log odds value for gene X at time i, then for two genes X and Y and N observations, a similarity score is calculated. (Reprinted, with permission, from Eisen et al. 1998 [copyright National Academy of Sciences].)

$$S(X, Y) = 1/N \sum_{i=1,N} \left(\frac{X_i - X_{\text{offset}}}{\Phi_X}\right)\left(\frac{Y_i - Y_{\text{offset}}}{Q_Y}\right)$$

$$\text{where } \Phi_X = \sqrt{\sum_{i=1,N} \frac{(X_i - X_{\text{offset}})^2}{N}}$$

$$\text{and } Q_Y = \sqrt{\sum_{i=1,N} \frac{(Y_i - Y_{\text{offset}})^2}{N}}$$

$S(X,Y)$ is also known as the Pearson correlation coefficent. X_{offset} and Y_{offset} can be the mean of the observations on X or Y, respectively, in which case Φ is the standard deviation, or else X_{offset} and Y_{offset} can be set to zero when a reference state is used (as in the present example). After values of $S(X,Y)$ have been calculated for all gene combinations, the most closely related pairs are identified in an above-diagonal scoring matrix. A node is created between the highest-scoring pair, and the gene-expressed profiles of these two genes are averaged and the joined elements are weighted by the number of elements they contain. The matrix is then updated replacing the two joined elements by the node. For n genes, the process is repeated $n - 1$ times until a single element remains. In the final dendrogram, the order of genes within a cluster is determined by simple weighting schemes, e.g., average dendrogram level (Eisen et al. 1998). The software availability is given in Table 10.1E, microarray guide. This image is available at http://rana.stanford.edu/clustering/serum.html. On the left side of the color display are letters A–E which identify clusters of genes that show clearly distinct responses to the treatment.

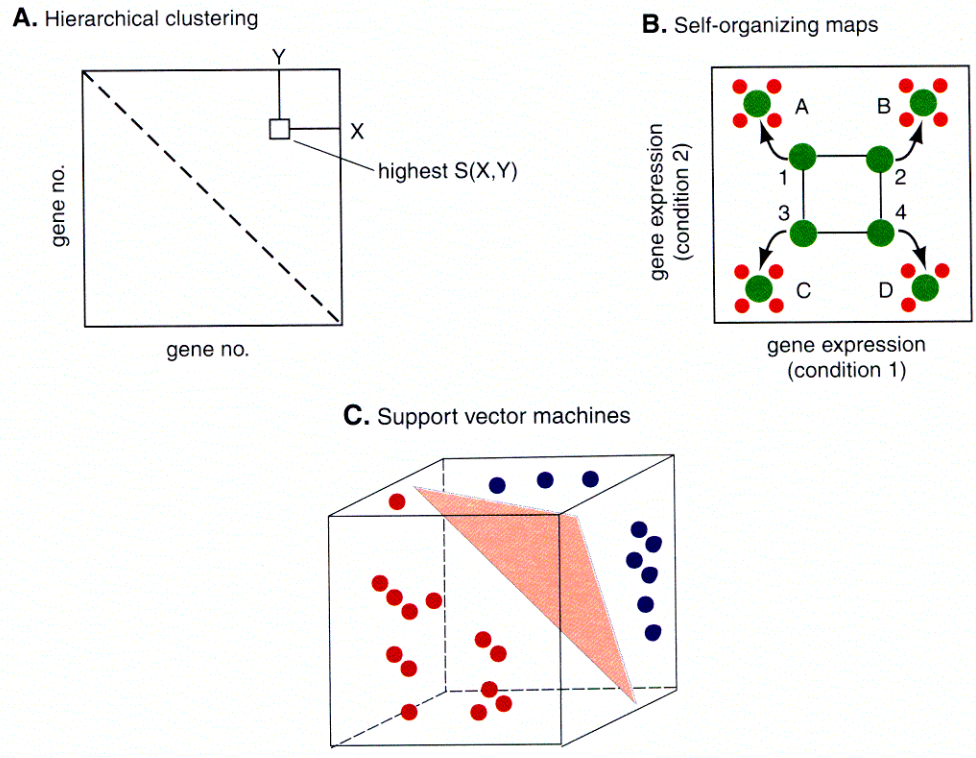


Figure 10.12. Examples of methods of cluster analysis for microarray data. (*A*) Hierarchical clustering (Eisen et al. 1998). (*B*) Self-organizing maps (Tamayo et al. 1999). (*C*) Support vector machines (Brown et al. 2000). (*A*) The hierarchical clustering method is described in detail in Fig. 10.11. Basically, the method generates a similarity score [$S(X,Y)$] for all gene combinations, places the scores in a matrix, joins those genes that have the highest score, and then continues to join progressively less similar pairs. The disadvantage of this method is that it fails to discriminate between different patterns of variation. For example, a gene expression pattern for which a high value is found at an intermediate time point will be clustered with another for which a high value is found at a late time point in the experiment. These variations have to be separated in a subsequent step (see Fig. 10.11). Methods B and C below are able to discriminate such differences. (*B*) In the SOM method, a choice is made of a number of clusters by which to organize the data. Shown is a 2 × 2 SOM comprising nodes 1–4 that assumes the presence of four clusters. Only two data points are shown in the example, but more data can be included by adding more dimensions to the analysis. The object is to move each node to the center of a cluster of data points. At each iteration, a data point P is selected, and the node closest to that point is identified. The location of that node is then moved slightly toward the point. Thus, node 1 in the example will gradually migrate to the center of cluster A, node 2 to cluster B, and so on. In practice, the size of the SOM is gradually increased until clearly different sets of expression patterns are identified. The computer program that performs the SOM analysis is GENECLUSTER (Table 10.1E). (*C*) SVMs are a binary classification method to discriminate one set of data points from another. They are similar to the types of discriminant analyses described for gene prediction in Chapter 9. For microarray analysis, sets of genes are identified that represent a target pattern of gene expression. The SVM is then trained to discriminate between the data points for that pattern and other data points

PREDICTION OF GENE FUNCTION BASED ON A COMPOSITE ANALYSIS

When two proteins share a considerable degree of sequence identity throughout the sequence alignment, they are likely to share the same function. A considerable fraction of a genome may encode proteins whose function may not be identified in this manner because the proteins are not related to another of known function. In the above sections, other types of evidence for a relationship between two genes are also given that are not dependent in sequence similarity. These include (1) genes are closely linked on the same chromosomes and transcribed from the same DNA strand, implying coordinated regulation in an operon-like structure; (2) gene fusions are observed between otherwise separate genes (suggests the encoded proteins are physically associated in a common complex); and (3) phylogenetic profiles reveal the genes are both commonly present in many organisms (implying they have interdependent metabolic functions). Three additional types of data have been used as evidence for gene relatedness: (1) the encoded proteins each have homologs in another organism that operate in a common metabolic pathway, (2) experimental data suggest an interaction between the proteins (stored in databases of interacting proteins; Table 9.5, p. 430), and (3) patterns of mRNA expressions are found to be correlated in microarray data. The results of using the above tests for the identification of a group of related genes in yeast are shown in Figure 10.13. In an examination of the entire yeast proteome, proteins that share a relationship with the yeast Sup35 protein based on one or more of the above tests are shown as points in a two-dimensional cluster where the distances between the points are proportional to the weight of the evidence for a relationship between the protein pair and the strength of the connection is proportional to the amount of evidence for a relationship. These types of predictions can be an important basis for hypotheses that can be tested experimentally.

that do not show the pattern. Shown in the diagram are two sets of data points (*red* and *blue*) in a three-dimensional plot that illustrate these two classes of data points. As the SVM learns to discriminate between the data sets, a hyper-plane (*pink*) is drawn between the sets. The hyper-plane is then used as a basis for classifying unknown data points. Only three dimensions are shown for illustrative purposes, but additional ones can be included, adding more dimensions to the analysis. SVMs were used to categorize genes based on 79 different sets of data points from studies of the yeast cell cycle and are particularly useful for such complex data sets. Data points are log-transformed and normalized as in method A, where for N observations of a gene i, the log transform X_i of the expression level E_i and reference level R_i is:

$$X_i = \frac{\log (E_i/R_i)}{\sqrt{\sum_{j=1,N} \log_z (E_j/R_j)}}$$

so that X_i is positive if the gene is more strongly expressed than in the reference condition, and negative if expression is reduced. Gene combinations averaged over all experimental conditions are then examined by a multidimensional analysis (see http://www.cse.ucsc.edu/research/compbio/genex). A tutorial on SVMs is available through http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT/svm_light.eng.html. (Adapted, with permission, from Gaasterland and Bekinanov 2000 [copyright Nature Publishing].)

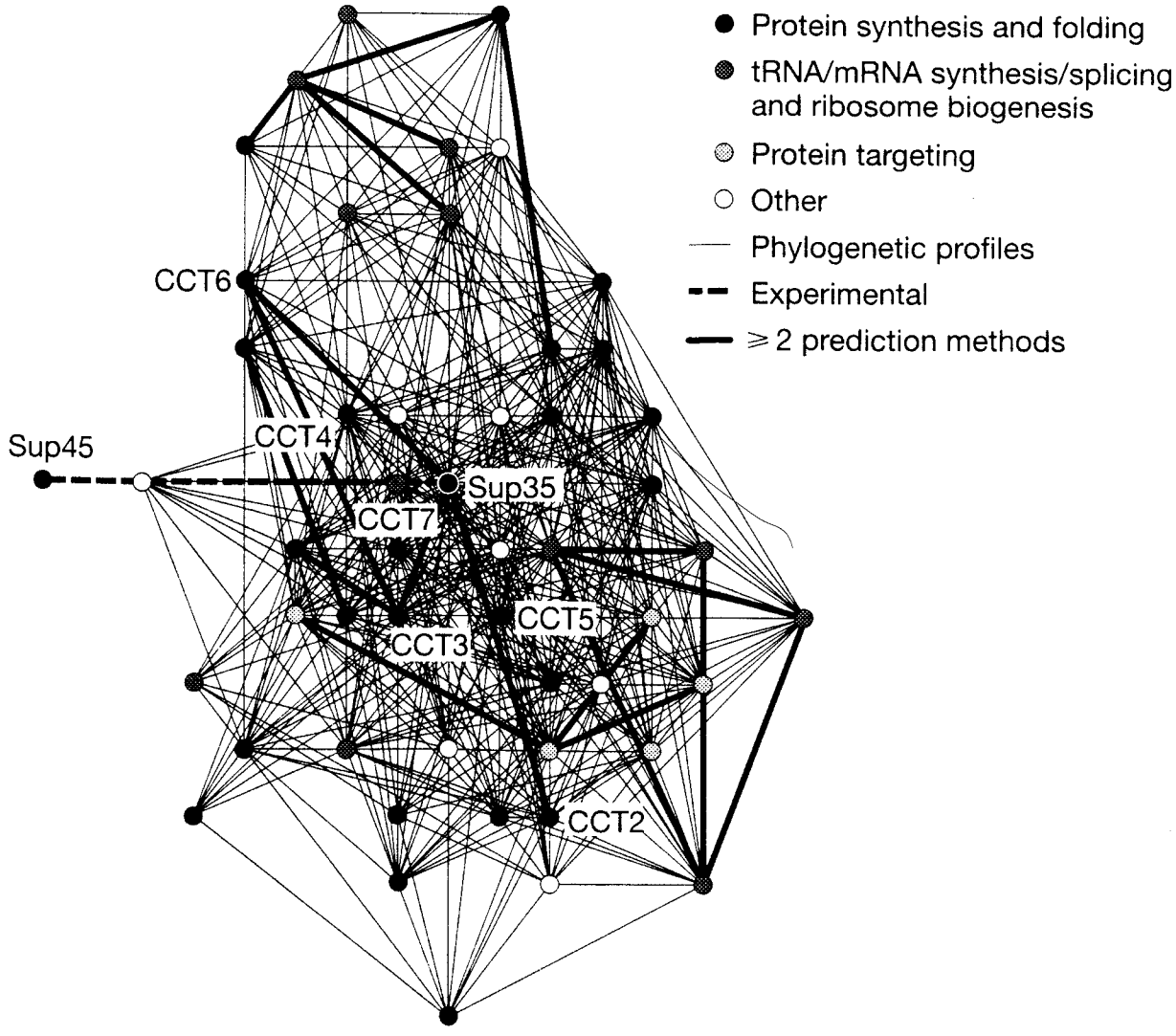


Figure 10.13. Genome-wide prediction of protein functions by a combinatorial method (Enright et al. 1999; Marcotte et al. 1999a). This figure shows the network of yeast proteins that are linked to the yeast prion and translation factor Sup35 (double circle in center of network). Each point represents a yeast protein, and branches between proteins indicate a relationship by one of several criteria indicated in the legend. Branch lengths are shorter for closely related proteins and thicker when two or more prediction methods indicate a relationship. Related to Sup35 protein are proteins involved in protein folding and targeting. The links are based on experimental data, proteins whose homologs are known to operate sequentially in metabolic pathways, proteins that evolved in a correlated fashion as evidenced by presence in fully sequenced genomes (see Snel et al. 1999), proteins whose homologs are fused into a single protein in another organism, and proteins whose mRNA expression profiles are similar under a range of cellular and environmental conditions. (Reprinted, with permission, from Marcotte et al. 1999a [copyright Macmillan].)

FUNCTIONAL GENOMICS

Genome analysis depends to a large extent on sequence analysis methods that identify gene function based on similarity between proteins of unknown function and proteins of known function. Known functions are derived from experimental evidence in molecular biology and genetic studies with model organisms. Orthologous genes between biologically distinct species (for example, yeast and fruit flies) can be identified, and the high sequence similarity between them is strong evidence for a related function. However, given the more complex multicellular biology of flies, the fly gene could have an additional function that is not predictable by the yeast model. In other cases, the occurrence of families of paralogous genes that share common domains can make a precise guess of function of one of these proteins more difficult because all match a model protein to some degree. Sequence-based methods of gene prediction can be augmented by the types of genome comparisons described above that are designed to identify related genes based on common patterns of expression, evolutionary profiles, chromosomal locations, and other features. However, all of the above methods can fail to provide a precise determination of gene function. Hence, methods have been devised for directing mutations into specific genes that inactivate or modify the gene function, and the effect is then analyzed in the mutant organism.

Two general types of approaches illustrated in Figure 10.14 are used—one in which a genetic construct is made that interferes with the expression of a particular gene (and sometimes a set of related genes) and a second in which a large number of random mutations are generated in a population of organisms. The individual with a mutation in a particular gene is then identified. Once mutants are obtained, the effect of the mutant genes on phenotype is determined. The gene function may then be predicted on the basis of the observed alterations. Because such extreme genetic experiments cannot be performed with humans, the mouse model for the human genome serves the same purpose. Web sites that compare the mouse and human genomes listed in Table 10.1C provide an important basis for analyzing the human genome. An orthologous gene is identified in the mouse genome, the sequence or expression of the gene is disrupted in some fashion, and a transgenic mouse homozygous for the mutant gene is then produced. Using this technology, one can systematically go through genes that regulate cell division, for example, and determine the significance of these genes in normal versus abnormal (tumor) growth.

PUTTING TOGETHER ALL OF THE INFORMATION INTO A GENOME DATABASE

A genome database may also be interfaced with other types of data, such as clinical data. This type of organization, termed data warehousing, can facilitate the search for novel relationships among the data by data-mining methods. These methods include genetic algorithms, neuronetworks, and others described elsewhere in this text.

The ultimate step in genome analysis is to collect the information found on gene and protein sequences, alignments, gene function and location, protein families and domains, relationships of genes to those in other organisms, chromosomal rearrangements, and so on, into a comprehensive database. This database should be logically organized so that all types of information are readily accessible and easily retrievable by users who have widely divergent knowledge of the organism. This goal is best achieved by using controlled vocabularies that can identify the same genetic or biochemical function in different organisms without ambiguity. Examples of groups that are developing systematic ways of defining terms and of collecting and organizing data are given in Table 10.1E. Other examples of database tools used to express biological information are given in Chapter 2 (page 44). The genome sites of model organisms listed in Table 10.1B, especially SGD and Flybase, provide examples for further study. In addition to the care needed in organizing genome databases, a

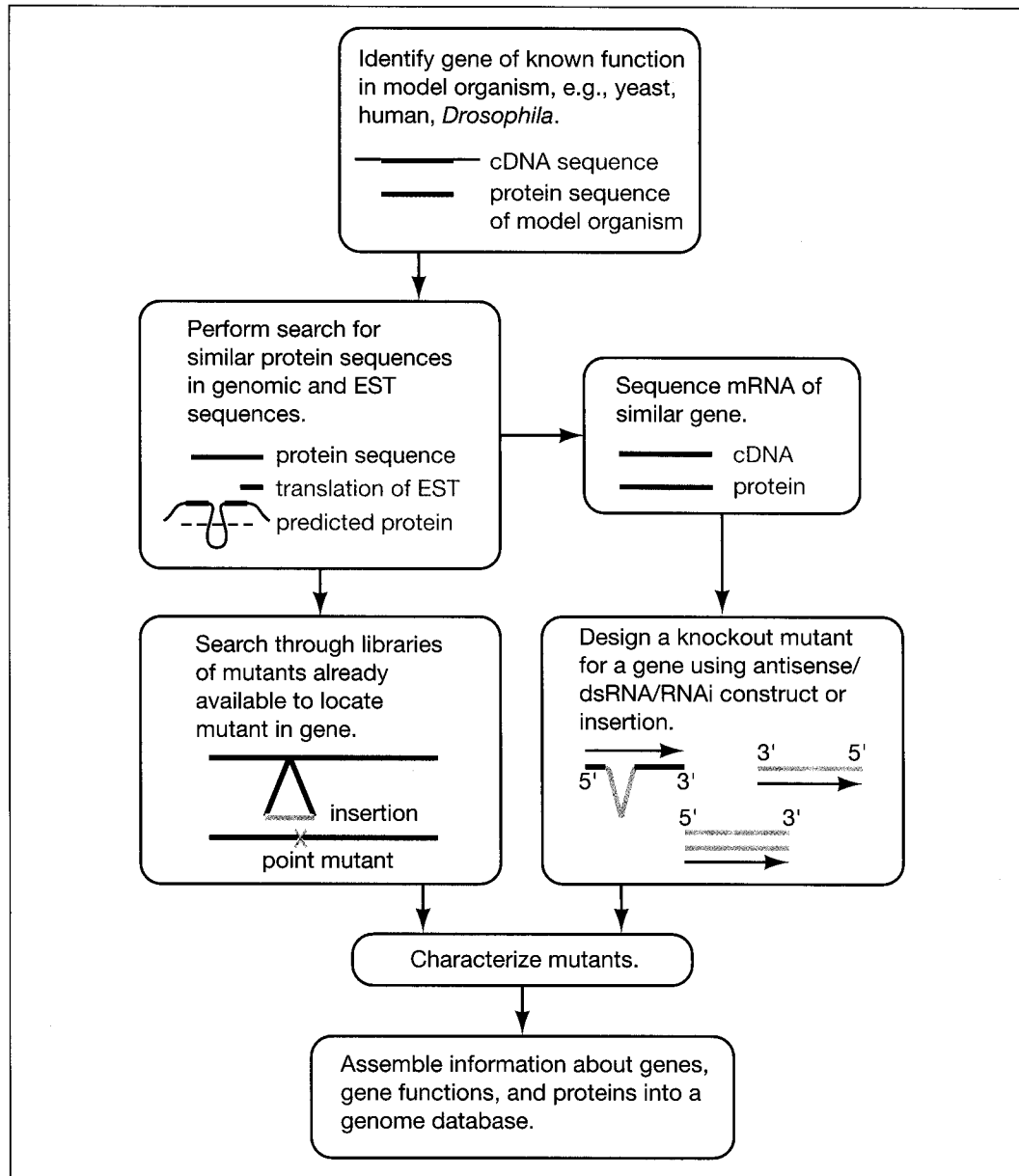


Figure 10.14. Reverse-genetics analysis of gene function. Steps for identification of gene function in an organism are identified. Even though a particular gene may be a highly predicted ortholog of a gene of known function in another organism, that gene may be acquired by a novel function. For example, a defect in a plant or animal gene that is a homolog of a yeast gene may have an effect on a developmental process or other biologically unique function of multicellular organisms. Information on knockout mutants in model organisms is available through the genome Web sites given in Table 10.1C. Directed gene knockout and mutagenesis methods are described in Fire et al. (1998) and McCallum et al. (2000), respectively.

great deal of human input is needed to annotate the genome manually with information about individual genes and proteins, effects of mutations in these genes, and other types of genome variations that cannot be readily incorporated into the database by automated methods. For the human genome, this activity will occupy the time of many scientists for many years to come.

REFERENCES

Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D., Amanatides P.G., Scherer S.E., Li P.W., Hoskins R.A., Galle R.F., et al. 2000. The genome sequence of *Drosophila melanogaster. Science* **287:** 2185–2195.

Andrade M.A., Brown N.P., Leroy C., Hoersch S., de Daruvar A., Reich C., Franchini A., Tamames J., Valencia A., Ouzounis C., and Sander C. 1999. Automated genome sequence analysis and annotation. *Bioinformatics* **15:** 391–412.

Bailey L.C., Jr., Fischer S., Schug J., Crabtree J., Gibson M., and Overton G.C. 1998. GAIA: Framework annotation of genomic sequence. *Genome Res.* **8:** 234–250.

Baker P.G., Goble C.A., Bechhofer S., Paton N.W., Stevens R., and Brass A. 1999. An ontology for bioinformatics applications. *Bioinformatics* **15:** 510–520.

Bansal A.K. 1999. An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics* **15:** 900–908.

Bennetzen J.L. 1998. The structure and evolution of angiosperm nuclear genomes. *Curr. Opin. Plant Biol.* **1:** 103–108.

———. 2000. Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. *Plant Cell* **12:** 1021–1030.

Bennetzen J.L., SanMiguel P., Chen M., Tikhonov A., Francki M., and Avramova Z. 1998. Grass genomes. *Proc. Natl. Acad. Sci.* **95:** 1975–1978.

Bernardi G. 1995. The human genome: Organization and evolutionary history. *Annu. Rev. Genet.* **29:** 445–476.

Blanchette M., Kunisawa T., and Sankoff D. 1996. Parametric genome rearrangements. *Gene* **172:** GC11–17.

Blattner F.R., Plunkett G., III, Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B., and Shao Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1462.

Bork P. 1999. Powers and pitfalls in sequence analysis: The 70% hurdle. *Genome Res.* **10:** 398–400.

Brown M.P., Grundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares M., Jr., and Haussler D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.* **97:** 262–267.

Brown T.A. 1999. *Genomes.* Wiley-Liss, New York.

Bult C.J., White O., Olsen G.J., Zhou L., Fleischmann R.D., Sutton G.G., Blake J.A., FitzGerald L.M., Clayton R.A., Gocayne J.D., Kerlavage A.R., Dougherty B.A., Tomb J.F., Adams M.D., Reich C.I., Overbeek R., Kirkness E.F., Weinstock K.G., Merrick J.M., Glodek A., Scott J.L., Geoghagen N.S.M., and Venter J.C. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273:** 1058–1073.

Cawthon R.M., O'Connell P., Buchberg A.M., Viskochil D., Weiss R.B., Culver M., Stevens J., Jenkins N.A., Copeland N.G., and White R.1990. Identification and characterization of transcripts from the neurofibromatosis 1 region: The sequence and genomic structure of EVI2 and mapping of other transcripts. *Genomics* **7:** 555–565.

Cherry J.M., Ball C., Weng S., Juvik G., Schmidt R., Adler C., Dunn B., Dwight S., Riles L., Mortimer R.K., and Botstein D. 1997. Genetic and physical maps of *Saccharomyces cerevisiae. Nature* (suppl. 6632) **387:** 67–73.

Chervitz S.A., Aravind L., Sherlock G., Ball C.A., Koonin E.V., Dwight S.S., Harris M.A., Dolinski K., Mohr S., Smith T., Weng S., Cherry J.M., and Botstein D. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282:** 2022–2028.

Chung S.-Y. and Wong L. 1999. Kleisli: A new tool for data integration in biology. *Trends Biotechnol.* **17:** 351–355.

Copertino D.W. and Hallick R.B. 1993. Group II and group III introns of twintrons: Potential relationships with nuclear pre-mRNA introns. *Trends Biochem. Sci.* **18:** 467–471.

Dandekar T., Snel B., Huynen M., and Bork P. 1998. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23:** 324–328.

de Bruijn F.J., Lupski J.R., and Weinstock G.M., Eds. 1998. *Bacterial genomes: Physical structure and analysis.* Chapman and Hall, New York.

Deckert G., Warren P.V., Gaasterland T., Young W.G., Lenox A.L., Graham D.E., Overbeek R., Snead M.A., Keller M., Aujay M., Huber R., Feldman R.A., Short J.M., Olsen G.J., and Swanson R.V. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus. Nature* **392:** 353–358.

Devos K.M. and Gale M.D. 2000. Genome relationships. The grass model in current research. *Plant Cell* **12:** 637–646.

Duran B.S. and Odell P.L. 1974. Cluster analysis: A survey. In *Lecture notes in economics and mathematical systems* (ed. M. Beckmann and H.P. Künzi). Springer-Verlag, New York.

Eisen M.B., Spellman P.T., Brown P.O., and Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Enright A.J. and Ouzounis C.A. 2000. GeneRAGE: A robust algorithm for sequence clustering and domain detection. *Bioinformatics* **16:** 451–457.

Enright A.J., Iliopoulos I., Kyrpides N.C., and Ouzounis C.A. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402:** 86–90.

Fire A., Xu S., Montgomery M.K., Kostas S.A., Driver S.E., and Mello C.C. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans. Nature* **391:** 806–811.

Fitch W.M. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19:** 99–113.

Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.-F., Dougherty B.A., Merrick J.M., et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Force A., Lynch M., Pickett F.B., Amores A., Yan Y.L., and Postlethwait J. 1999. Nucleotide preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.

Gaasterland T. and Bekiranov S. 2000. Making the most of microarray data. *Nat. Genet.* **24:** 204–206.

Gaasterland T. and Sensen C.W. 1998. MAGPIE: A multipurpose automated genome project investigation environment for ongoing sequencing projects. In *Bacterial genomes: Physical structure and analysis* (ed. F.J. de Bruijn et al.), pp. 559–582. Chapman and Hall, New York.

Gilbert W., de Souza S.J., and Long M. 1997. Origin of genes. *Proc. Natl. Acad. Sci.* **94:** 7698–7703.

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., and Lander E.S. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286:** 531–537.

Green P., Lipman D., Hillier L., Waterston R., States D., and Claverie J.-M. 1993. Ancient conserved regions in new gene sequences and the protein databases. *Science* **259:** 1711–1716.

Henikoff S., Greene E.A., Pietrokovski S., Bork P., Attwood T.K., and Hood L. 1997. Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278:** 609–614.

Himmelreich R., Plagens H., Hilbert H., Reiner B., and Herrmann R. 1997. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium. Nucleic Acids Res.* **25:** 701–712.

Himmelreich R., Hilbert H., Plagens H., Pirkl E., Li B.C., and Herrmann R. 1996. Complete sequence of the genome of the bacterium *Mycoplasma pneumoniae. Nucleic Acids Res.* **24:** 4420–4449.

Hoersch S., Leroy C., Brown N.P., Andrade M.A., and Sander C. 2000. The GeneQuiz web server: Protein functional analysis through the Web. *Trends Biochem. Sci.* **25:** 33–35.

Hoogland C., Sanchez J.C., Tonella L., Binz P.A., Bairoch A., Hochstrasser D.F., and Appel R.D. 2000. The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.* **28:** 286–288.

Hughes J.D., Estep P.W., Tavazoie S., and Church G.M. 2000. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae. J. Mol. Biol.* **296:** 1205–1214.

Huttley G.A., Easteal S., Southey M.C., Tesoriero A., Giles G.G., McCredie M.R., Hopper J.L., and Venter D.J. 2000. Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian breast cancer family study. *Nat. Genet.* **25:** 410–413.

Huynen M.A. and Bork P. 1998. Measuring genome evolution. *Proc. Natl. Acad. Sci.* **95:** 5449–5856.

Huynen M., Snel B., Lathe W., III, and Bork P. 2000. Exploitation of gene context. *Curr. Opin. Struct. Biol.* **10:** 366–370.

Kanehisa M. and Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28:** 27–30.

Kaneko T., Katoh T., Sato S., Nakamura Y., Asamizu E., Kotani H., Miyajima N., and Tabata S. 1999. Structural analysis of *Arabidopsis thaliana* chromosome 5. IX. Sequence features of the regions of 1,011,550 bp covered by seventeen P1 and TAC clones. *DNA Res.* **6:** 183–195.

Kaneko T., Sato S., Kotani H., Tanaka A., Asamizu E., Nakamura Y., Miyajima N., Hirosawa M., Sugiura M., Sasamoto S., Kimura T., Hosouchi T., Matsuno A., Muraki A., Nakazaki N., Naruo K., Okumura S., Shimpo S., Takeuchi C., Wada T., Watanabe A., Yamada M., Yasuda M., and Tabata S. 1996a. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3:** 109–136.

———. 1996b. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* (suppl.) **3:** 185–209.

Karlin S., Campbell A.M., and Mrázek J. 1998. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32:** 185–225.

Karp P.D., Riley M., Saier M., Paulsen I.T., Paley S., and Pellegrini-Toole A. 2000. The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* **28:** 56–59.

Kawarabayasi Y., Sawada M., Horikawa H., Haikawa Y., Hino Y., Yamamoto S., Sekine M., Baba S., Kosugi H., Hosoyama A., Nagai Y., Sakai M., Ogura K., Otsuka R., Nakazawa H., Takamiya M., Ohfuku Y., Funahashi T., Tanaka T., Kudoh Y., Yamazaki J., Kushida N., Oguchi A., Aoki K., and Kikuchi H. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3 (supplement). *DNA Res.* **5:** 147–155.

Kececioglu J. and Gusfield D. 1998. Reconstructing a history of recombinations from a set of sequences. *Discrete Appl. Math.* **88:** 239–260.

Kececioglu J. and Sankoff D. 1995. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica* **13:** 180–210.

Kidwell M.G. and Lisch D.R. 1997. Transposable elements as sources of variation in plants and animals. *Proc. Natl. Acad. Sci.* **94:** 7704–7711.

———. 2000. Transposable elements and host genome evolution. *Trends Ecol. Evol.* **15:** 95–99.

Koonin E.V. and Galperin M.Y. 1997. Prokaryotic genomes: The emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.* **7:** 757–763.

Koonin E.V., Tatusov R.L., and Galperin M.Y. 1998. Beyond complete genomes: From sequence to structure and function. *Curr. Opin. Struct. Biol.* **8:** 355–363.

Kunst F., Ogasawara N., Moszer I., Albertini A.M., Alloni G., Azevedo V., Bertero M.G., Bessieres P., Bolotin A., Borchert S., et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature* **390:** 249–256.

Kyrpides N.C. 1999. Genomes OnLine database (GOLD 1.0): A monitor of complete and ongoing genome projects world-wide. *Bioinformatics* **15:** 773–774.

Kyrpides N.C. and Ouzounis C.A. 1999. Whole-genome sequence annotation: "Going wrong with confidence." *Mol. Microbiol.* **32:** 886–887.

Labedan B. and Riley M. 1995. Gene products of *Escherichia coli:* Sequence comparisons and common ancestries. *Mol. Biol. Evol.* **12:** 980–987.

Li W.-H. 1997. *Molecular evolution.* Sinauer Associates, Sunderland, Massachusetts.

Liang F., Holt I., Pertea G., Karamycheva S., Salzberg S.L., and Quackenbush J. 2000. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25:** 239–240.

Lin X., Kaul S., Rounsley S., Shea T.P., Benito M.I., Town C.D., Fujii C.Y., Mason T., Bowman C.L., Barnstead M., et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana. Nature* **402:** 761–768.

Link A.J., Eng J., Schieltz D.M., Carmack E., Mize G.J., Morris D.R., Garvik B.M., and Yates J.R., III. 1999. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17:** 676–682.

Lupski J.R. 1998. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Genetics* **14:** 417–422.

Marcotte E.M., Pellegrini M., Thompson M.J., Yeates T.O., and Eisenberg D. 1999a. A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83–86.

Marcotte E.M., Pellegrini M., Ng H., Rice W.D., Yeates T.O., and Eisenberg D. 1999b. Detecting protein function and protein–protein interactions from genome sequences. *Science* **285:** 751–753.

McCallum C.M., Comai L., Greene E.A., and Henikoff S. 2000. Targeting induced local lesions in genomes (TILLING) for plant functional genomics. *Plant Physiol.* **123:** 439–442.

McGuire A.M., Hughes J.D., and Church G.M. 2000. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10:** 744–757.

Myers E.W., Sutton G.G., Delcher A.L., Dew I.M., Fasulo D.P., Flanigan M.J., Kravitz S.A., Mobarry C.M., Reinert K.H.J., Remington K.A., et al. 2000. A whole-genome assembly of *Drosophila. Science* **287:** 2196–2204.

O'Brien S.J., Menotti-Raymond M., Murphy W.J., Nash W.G., Wienberg J., Stanyon R., Copeland N.G., Jenkins N.A., Womack J.E., and Marshall Graves J.A. 1999. The promise of comparative genomics in mammals. *Science* **286:** 458–462, 479–481.

Ochmann H., Lawrence J.G., and Groisman E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405:** 299–304.

Oliver S.G., van der Aart Q.J., Agostoni-Carbone M.L., Aigle M., Alberghina L., Alexandraki D., Antoine G., Anwar R., Ballesta J.P., Benit P., et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357:** 38–46.

Ouzounis C., Casari G., Sander C., Tamames J., Valencia A. 1996. Computational comparisons of genomes. *Trends Biotechnol.* **14:** 280–285.

Overbeek R., Fonstein M., D'Souza M., Pusch G.D., and Maltsev N. 1999. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci.* **96:** 2896–2901.

Overbeek R., Larsen N., Pusch G.D., D'Souza M., Selkov E., Jr., Kyrpides N., Fonstein M., Maltsev N., and Selkov E. 2000. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids. Res.* **28:** 123–125.

Paterson A.H., Lin Y.-R., Li Z., Schertz K.F., Doebley J.F., Pinson S.R.M., Liu S.-C., Stansel J.W., and Irvine J.E. 1995. Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science* **269:** 1714–1718.

Pellegrini M., Marcotte E.M., Thompson J.M., Eisenberg D., and Yeates T.O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* **96:** 4285–4288.

Perou C.M., Jeffrey S.S., van de Rijn M., Rees C.A., Eisen M.B., Ross D.T., Pergamenschikov A., Williams C.F., Zhu S.X., Lee J.C., Lashkari D., Shalon D., Brown P.O., and Botstein D. 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci.* **96:** 9212–9217.

Pruitt K.D., Katz K.S., Sicotte H., and Maglott D.R. 2000. Introducing RefSeq and LocusLink: Curated human genome resources at the NCBI. *Trends Genet.* **16:** 44–47.

Retief J.D., Lynch K.R., and Pearson W.R. 1999. Panning for genes: A visual strategy for identifying novel gene orthologs and paralogs. *Genome Res.* **9:** 373–382.

Riley M. 1993. Functions of the gene products of *Escherichia coli. Microbiol. Rev.* **57:** 862–952.

———. 1998. *E. coli* genes: Ancestry and map locations. In *Bacterial genomes: Physical structure and analysis* (ed. F.J. de Bruijn et al.), pp. 187–195. Chapman and Hall, New York.

Roth F.P., Hughes J.D., Estep P.W., and Church G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16:** 939–945.

Rowen L., Koop B.F., and Hood L. 1996. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* **272:** 1755–1762.

Rubin G.M., Yandell M.D., Wortman J.R., Gabor Miklos G.L., Nelson C.R., Hariharan I.K., Fortini M.E., Li P.W., Apweiler R., Fleischmann W., et al. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204–2215.

Ruiz M., Giudicelli V., Ginestoux C., Stoehr P., Robinson J., Bodmer J., Marsh S.G., Bontrop R., Lemaitre M., Lefranc G., Chaume D., and Lefranc M.P. 2000. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* **28:** 219–221.

Sanger F., Coulson A.R., Hong G.F., Hill D.F., and Petersen G.B. 1982. Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* **162:** 729–773.

Sankoff D. and Goldstein M. 1989. Probabilistic models of genome shuffling. *Bull. Math. Biol.* **51:** 117–124.

Sankoff D. and Nadeau J.H. 1996. Conserved synteny as a measure of genomic distance. *Discrete Appl. Math.* **71:** 247–257.

Sankoff D., Cedergren R., and Abel Y. 1993. Genome divergence through gene rearrangement. *Methods Enzymol.* **183:** 428–438.

Sankoff D., Leduc G., Antoine N., Paquin B., Lang B.F., and Cedergren R. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci.* **89:** 6575–6579.

SanMiguel P., Gaut B.S., Tikhonov A., Nakajima Y., and Bennetzen J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20:** 43–45.

SanMiguel P., Tikhonov A., Jin Y.K., Motchoulskaia N., Zakharov D., Melake-Berhan A., Springer P.S., Edwards K.J., Lee M., Avramova Z., and Bennetzen J.L. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274:** 765–768.

Schultz J., Copley R., Doerks T., Ponting C.P., and Bork P. 2000. SMART: A web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28:** 231–234.

Smit A.F. 1996. The origin of interdispersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6:** 743–748.

Snel B., Bork P., and Huynen M. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21:** 108–110.

———. 2000a. Genome evolution: Gene fusion versus gene fission. *Trends Genet.* **16:** 9–11.

Snel B., Lehmann G., Bork P., and Huynen M.A. 2000b. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* (in press).

Stoeckert C.J., Jr., Salas F., Brunk B., and Overton G.C. 1999. EpoDB: A prototype database for the analysis of genes expressed during vertebrate erythropoiesis. *Nucleic Acids Res.* **27:** 200–203.

Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., and Golub T.R. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* **96:** 2907–2912.

Tatusov R.L., Koonin E.V., and Lipman D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.

Tatusov R.L., Galperin M.Y., Natale D.A., and Koonin E.V. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **8:** 33–36.

Tatusov R.L., Mushegian A.R., Bork P., Brown N.P., Hayes W.S., Borodovsky M., Rudd K.E., and Koonin E.V. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli. Curr. Biol.* **6:** 279–291.

Tekaia F., Lazcano A., and Dujon B. 1999. The genomic tree as revealed from whole genome comparisons. *Genome Res.* **9:** 550–557.

Tipton K. and Boyce S. 2000. History of the enzyme nomenclature system. *Bioinformatics* **16:** 34–40.

Tomita M., Hashimoto K., Takahashi K., Shimizu T.S., Matsuzaki Y., Miyoshi F., Saito K., Tanida S., Yugi K., Venter J.C., and Hutchison C.A., III. 1999. E-CELL: Software environment for whole-cell simulation. *Bioinformatics* **15:** 72–84.

Volff J.N. and Altenbuchner J. 2000. A new beginning with new ends: Linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol Lett.* **86:** 143–150.

Webb E.C. 1992. *Enzyme nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Academic Press, San Diego, California.

Weiner A.M. 2000. Do all SINEs lead to LINEs? *Nat. Genet.* **24:** 332–333.