



ATTGACTAGTAC

010011011100010

$P(S < x) = \exp[-e^{-x}]$

# 11

# Excerpt from Chapter 11

## Genome Analysis

### C O N T E N T S

#### INTRODUCTION, 496

- Chapter Guide for Biologists, 496
- Chapter Guide for Computational Scientists, 497
- What Should Be Learned in This Chapter?, 498
- Glossary Terms, 498

#### GENOME ANALYSIS PRESENTS MANY CHALLENGES, 499

- Genome Sequences of Individuals Vary, 500
- Genome Duplication Influences Genetic Variation, 502
- Analysis of the Proteome, 503
- Web Resources and Computational Tools for Genome Analysis, 503

#### GENOME ANATOMY HAS BEEN STUDIED IN PROKARYOTES AND EUKARYOTES, 503

- Prokaryotic Genomes, 504
- Eukaryotic Genomes, 506
  - Sequence Repeats, 506
  - Gene Structure Varies in Eukaryotes, 509
  - Pseudogenes, 510

#### HOW GENOME SEQUENCE IS ASSEMBLED AND GENES ARE IDENTIFIED, 511

- Searching the Genome for Protein-encoding Genes, 511

#### HOW IS GENOME ANALYSIS PERFORMED?, 516

#### GENOMES CAN BE COMPARED FOR ORTHOLOGS, PARALOGS, AND PROTEOMES, 519

- Proteome Analysis, 521
  - All-against-All Self-comparison Reveals Numbers of Gene Families and Duplicated Genes, 521

*Between-Proteome Comparisons to Identify Orthologs, Gene Families, and Domains, 523*

Ancient Conserved Regions, 527

Horizontal Gene Transfer, 527

#### GENES CAN BE CLASSIFIED ACCORDING TO FUNCTION, 528

#### GENE ORDER (SYNTENY) IS CONSERVED ON CHROMOSOMES OF RELATED ORGANISMS, 529

#### GENOMES CAN BE USED TO PREDICT EVOLUTIONARY RELATIONSHIPS, 530

Visualizing Chromosomal Rearrangements, 530

Computational Analysis of Gene Rearrangements, 532

Clusters of Metabolically Related Functions on Prokaryotic Chromosomes, 536

Composite Genes with a Multiple Set of Domains Predict Physical Interactions and Functional Relationships between Protein Pairs That Share the Same Domains, 538

Resources for Genome Analysis, 538

#### MICROARRAY ANALYSIS PROVIDES INFORMATION ABOUT GLOBAL GENE REGULATION, 539

#### GENE FUNCTION CAN BE PREDICTED USING A COMPOSITE ANALYSIS, 540

#### FUNCTIONAL GENOMICS APPROACHES IDENTIFY GENE FUNCTION, 540

#### COLLECTING ALL OF THE INFORMATION INTO A GENOME DATABASE, 542

#### WEB SEARCH TERMS FROM THIS CHAPTER, 543

#### REFERENCES, 544

#### PROBLEMS FOR CHAPTER 11, 548

## INTRODUCTION

A major application of bioinformatics is analysis of the full genomes of organisms that have been sequenced starting in the late 1990s, including over 100 microbial genomes, the budding yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, the plant *Arabidopsis thaliana*, the fruit fly *Drosophila*, the human genome, as well as many organisms that are human pathogens. Many additional genome sequencing projects have been completed and others are planned or under way.

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.

## CHAPTER GUIDE FOR BIOLOGISTS

This chapter describes applications of many of the sequence analysis methods that have been introduced in previous chapters and illustrates how to use these methods for genome analysis. The methods start with genome fragment assembly into full-length genome sequences, sometimes guided with a physical map and, at other times, without a physical map. To analyze a single genome, DNA and protein sequence alignments and similarity searches are used to find and analyze repetitive DNA sequences, including highly repetitive sequences, gene families (encoding proteins that are paralogs resulting from gene duplication), gene intron structure, and gene/chromosomal duplications. Between-genome sequence similarity searches, sequence alignments, and methods to cluster genes that are similar in principle to phylogenetic analyses are used to identify strongly alike genes encoding proteins with the same function (orthologs). Protein analysis methods are also used for analysis of protein function, three-dimensional structure, and domain analysis.

Genome analysis also introduces some new computational challenges not covered previously in this book. One challenge is analyzing genome rearrangements between closely related species, e.g., human and mouse, to determine where the rearrangements occurred in the genome and in what order. In aligning DNA and protein sequences, it is assumed that the order of DNA bases or amino acids in the sequences has not changed. In aligning genomes, the gene order is often not conserved, so that new alignment methods are needed. The problem is discovering how to turn the gene order of one genome into the gene order of the other in the least (most parsimonious) number of steps. Once this number is computed, the evolutionary distance between genomes can be calculated.

A second challenge is analyzing single-nucleotide polymorphisms (SNPs) within a species population. SNPs in the human genome are linked into haplotype blocks defined by preferred recombination sites during meiosis. Genes within these blocks have been found by traditional genetic analysis to be in linkage disequilibrium; i.e., they are seldom separated by meiotic recombination events. If a human population is screened for SNPs in one chromosomal region, then they should fall into linkage groups that represent the haplotype block structure. Discovering these blocks in large sequence data sets and defining a representative set of SNPs that can be used to assess an individual's haplotype are problems of considerable interest to the computational biology community.

Developing standard formats for data storage, display, and sharing is another group of computational challenges in genome analysis. For example, as SNPs and other information bearing upon the functional organization of genomes are collected, these data need to be stored in an appropriate database format and made available for distribution on Web sites.

## CHAPTER GUIDE FOR COMPUTATIONAL SCIENTISTS

Biologists have collected the genome sequence, which is the complete DNA sequence of all of an organism's chromosomes, of over 100 different organisms ranging from simple, one-celled organisms to multicellular organisms with complex developmental and life cycles. These DNA sequences include genes that specify the amino acid sequences of proteins, sequences that encode RNA molecules needed for making proteins and performing other biological functions, and highly repetitive sequences of no apparent function that are carried along in the genome. Genes that encode proteins and RNA sequences needed for basic life processes such as protein synthesis and reproduction are conserved in most of these organisms, reflecting the importance of these processes and their common evolutionary origin. Other sets of genes, including those needed for development in multicellular organisms, are also conserved as a group. These gene sequences and the amino acid sequences of the encoded proteins may be used to predict the evolutionary history of genomes. Genes frequently occur in families resulting from repeated duplication and subsequent diversification of the biological function of these genes through sequence variation.

One issue to consider in genome analysis is the content of genes and how it might change. Genome analysis has revealed that introns (sequences that interrupt the amino acid–specifying regions of messenger RNA molecules) are present in the genes of most higher organisms (eukaryotes) but are noticeably absent in most single-celled organisms (prokaryotes). The size and number of introns in particular genes can be tracked in the genomes of closely related species. The function of introns is not known, but it has been suggested that they play an important role in genome evolution of eukaryotic organisms. The birth of new introns and the death of introns in genes by processes not yet understood can potentially add or remove amino acids to the protein product of the gene, change the expression level of a gene in different tissues, and cause other changes that promote biological diversity.

There are several classes of sequences (transposable elements) that can move from one genome location to another, thus affecting gene content. Highly repetitive sequences in the genome are derived from such classes of sequences that move (transpose) from one genome location to another. These sequences fall into classes based on length, DNA sequence, and the biochemical mechanisms for producing them. Moving to new chromosomal locations potentially interrupts other sequences, but in some cases, provides a biological advantage. These sequence elements are thought to have played an important role in genome evolution by promoting genome rearrangements.

In addition to gene content, the number and arrangement of genes in genomes may also vary. Sequence similarity analysis reveals that individual genes, groups of many genes, and whole chromosomal regions can be duplicated and rearranged during the evolution of genomes. Duplications provide more diverse groups of biological functions that account for species differences; rearrangements can influence the timing and extent of gene expression. When mammalian genomes are compared, e.g., mouse and human, a large number of rearrangements are found. Modeling these rearrangements between species to discover how many there were, in what order they occurred, and how random they were is a challenging computational problem. One of the most interesting areas of genome analysis is comparing genomes of similar organisms, especially the primates that are the closest relatives of humans. These types of analyses can help to discover those sequences that are the most conserved and therefore most important for function. Human variations in these sequences can then be used as potential disease markers.

One particular benefit of the human genome project to human genetics and medicine is the information provided regarding sequence variation among individuals. Each individual has a change in DNA sequence of approximately 1 DNA base in every 500 bases compared to any other individual. These changes are called single-nucleotide polymorphisms (SNPs), and, collectively, they are responsible for most inherited differences between individuals. However, DNA sequences are transmitted from one generation to the next as haplotype blocks of sequence. Each block contains a few thousand to hundreds of thousands of DNA bases. It is these blocks that determine genetic variation between individuals by keeping groups of SNPs together in the human population, with human offspring receiving new block combinations from each parent. The challenge is to discover subsets of SNPs that indicate the block makeup of individuals. Although this analysis is primarily focused on the human genome, most other species also have sequence variations that determine biological properties in their populations, e.g., stress and disease resistance in plants.

This chapter explores the application of sequence analysis methods to these issues of genome analysis.

## WHAT SHOULD BE LEARNED IN THIS CHAPTER?

- Types of sequences found in genomes and how they vary between genomes.
- Methods of sequence analysis that are used to compare genomes.
- Roles of transposable elements and gene duplication in genome evolution.
- How to discover orthologous and paralogous genes, gene families, and conserved protein domains.
- How to model genome rearrangements.
- How to collect and use information on gene function.

## Glossary Terms

**Annotation** is the process of marking a genome sequence with information regarding function, such as, for instance, the location of exons and introns within a eukaryotic gene.

**Archaea** is a third class of organisms that is different from the classes Bacteria and Eukarya. It was originally determined by analysis of the evolution of ribosomal RNA structure.

**Centromeres** are specific chromosomal regions that serve as points of attachment between newly replicated daughter chromosomes and that are also used for pulling the chromosomes apart during cell division.

**COGs (clusters of orthologous groups)** analysis shows gene relationships as clusters of orthologous groups, a type of graphical representation of the sequence relationships among related genes in a group of organisms, usually prokaryotic organisms.

**Computer script** is a line of code in certain computer languages, e.g., Perl, that is immediately interpreted by the machine, as opposed to computer languages such as C, which require compilation to produce an executable program.

**Conserved regions** refers to regions of common DNA or protein sequence in two genome contexts.

**Core proteome** is the set of basic biological functions required by an organism for survival.

**Eukaryotes** are a class of organisms composed of cells that have a microscopically visible nucleus separated from the cellular cytoplasm by a nuclear membrane and containing the chromosomes of the organism.

**E value (expectation value)** is a statistical value provided by programs that perform genome analysis by sequence

database similarity searches. The *E* value of the alignment score between a query sequence and a database sequence is the number of unrelated sequences in the database that are expected to achieve as good an alignment score with the query sequence.

**Gene cluster** is a group of related genes shown as points on a graph joined by lines (edges) that indicate close gene relationships.

**Gene duplication** is the process of making a new copy of an existing gene in the genome of an organism.

**Genome** refers to the entire DNA sequence of one set of chromosomes of an organism.

**Genome comparison** is a comparison between the genomes of different organisms in regard to a variety of features such as the set of encoded proteins (proteome), the order of genes on the genome, the presence of transposable elements, etc. Comparisons are also made between the genomes of individuals of the same species including sequence variation (SNPs, repeats, etc.) and association with disease, e.g., cancer.

**Genome tree** is a tree based on genome similarities and differences as, for example, the fraction of proteins that are similar in organisms.

**Graph** is a representation of relationships among a group of objects, as, for example, the genes in two genomes. The objects are represented by nodes (vertices) that may include information about the objects, and related objects are joined by edges that are labeled according to degrees of relationship. The graph may be used to find clusters of objects that are most closely related.

**Haplotype block** is one region of one chromosome that is transmitted to the genome of the next generation.

**Haplotype map (hapmap)** is a genome map that shows blocks of chromosomal sequence of variable length that are passed along from one generation to the next and maintained in the population of an organism. Discovered as conserved patterns of sequence variations (see SNPs), they are of interest in the human genome because they cause genetic variation and influence disease risk.

**Horizontal gene transfer** (also called lateral gene transfer) is the transfer and insertion of DNA sequences from the genome of one species into the genome of a second species.

**Isochores** are genome regions having a distinct level of GC-richness.

**LINES** are long interspersed nuclear elements of length 4–7 kilobases found in large numbers in eukaryotic genomes including the human genome. They can promote insertion of other sequences, e.g., pseudogenes into the genome.

**MITES** are miniature, inverted repeat transposable elements often found in association with sequences that regulate transcription.

**Orthologs** are proteins of highly conserved sequence, function, and structure that are found to be the most-alike pair in whole-genome comparisons.

**Paralogs** are proteins that share sequence similarity and originated from gene duplication events.

**Prokaryotes** are single-celled organisms, such as bacteria, that lack an observable nucleus and generally have small genomes of several million DNA bases.

**Protein domains** are conserved amino acid sequences that are found in proteins and often represent a conserved function and structure.

**Protein families** are groups of proteins that are found to be at least 50% identical in sequence alignments.

**Proteome** is the entire complement of proteins encoded in the genome of an organism.

**Pseudogene** is a nonfunctional duplicate copy of a gene, usually produced by making a DNA copy of the spliced mRNA of a gene and inserting the copy into the genome.

**QTLs (quantitative trait loci)** are sequence variations that are so close to biologically important genes that they stay together from one generation to the next. The presence of such nearby genes is revealed by a strong association of the QTL with an important biological property, e.g., yield in plants.

**Ridges** are genome domains that have a high gene density, GC content, SINE repeat density, and a low LINE repeat density, as well as shorter introns than genes outside of ridges. Genes that are strongly expressed cluster into ridges. Antiridges contain clusters of weakly expressed genes and have opposite sequence characteristics to ridges.

**Shotgun sequencing** is the process of sequencing a genome based on assembly of random fragments based on sequence overlaps into a linear genome sequence without using a physical map as a guide.

**SINES** are short interspersed nuclear elements 80–400 DNA bases long found in large numbers in eukaryotic genomes. An example is *Alu* sequences in the human.

**Single-nucleotide polymorphisms (SNPs)** are sequence variations at a single base position that are quite common between individuals of the same species.

**Synten**y is a colinearity of gene order in two species: a conserved group of genes in the same order in two genomes as a syntenic group or cluster. In genome analysis, synteny will be an extended local alignment between two genome sequences that may not necessarily be in the same chromosomal location in the species.

**Telomeres** are sequences composed of short repeated elements found at the ends of chromosomes and necessary for chromosomal replication.

**Transposable elements** are DNA sequences that move from one chromosomal location to another. Often many copies representing many transpositions will be found in a genome.

## GENOMES CAN BE COMPARED FOR ORTHOLOGS, PARALOGS, AND PROTEOMES

---

Comparative genomics includes a comparison of all of the predicted proteins, i.e., the proteomes of two or more organisms, gene locations, and the number and location of sequence repeats, as possible influences on

genome evolution. Another area of genome comparison is the degree of diversity within a species population (e.g., SNPs, microsatellite variations, gene expression levels) and the association of these sequence vari-

ations with environmental responses and disease. The availability of complete genome sequences makes possible a comparison of the proteome of one organism with that of another. Because the genome sequence provides both the sequence and the map location of each gene, both the sequence and the location of each predicted protein can be compared. As described in Chapter 10, proteins are frequently found in families that have similar biological function and three-dimensional structure. Genome analysis provides additional information on the evolutionary relationships among proteins by exploring orthologs, paralogs, and protein domain changes.

Sequence comparisons between genomes provide information on gene relationships. Genes are called orthologs when the number of genes in two organisms are so similar that they must have the same function and evolutionary history (Fitch 1970). Map locations of orthologs in two or more genomes may also be compared. If a set of genes is grouped together at a particular chromosomal location in one organism, and if a set of similar genes is also grouped together in the chromosome of another organism, these

groups share an evolutionary history. Some of the group may come from a common ancestor gene whose function has been conserved over periods of evolutionary time. Others may have arisen by rare gene duplication events.

Gene families originating from rare gene duplication events over evolutionary time, called paralogs, are also found within a genome. Gene duplication events give rise to this type of gene relationship, which is found by genome comparisons. Proteins with new functions may be produced by such gene duplication events. Two tandem copies of a gene are produced (see Fig. 3.3). Through mutation and natural selection, one of the copies can develop a new function, leaving the second 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). Long chromosomal regions or

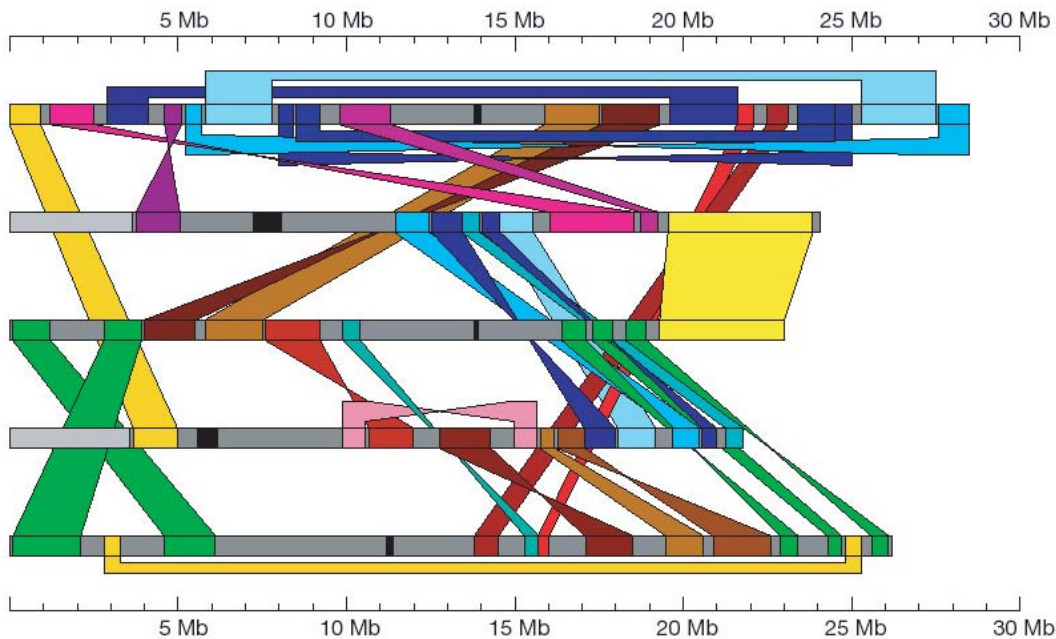


FIGURE 11.7. Presence of duplicated segments in the genome of the plant *Arabidopsis*. Multiple regions of sequence similarity are found in the same or opposite orientation by performing extensive comparisons of sections of genome sequence using the methods described in Chapter 3. These regions indicate the occurrence of genome duplications in the evolutionary history of the *Arabidopsis* genome. (Reprinted, with permission, from *Arabidopsis* Genome Initiative 2000 [©2000 Macmillan Magazines, Ltd.]) (Image by Dirk Haase.) For a similar map of duplications in the human genome, see Bailey et al. (2002).

even entire genomes can also be involved in genome duplications that create additional copies of genes in the duplicated regions, thus providing a similar opportunity for functional divergence of genes. As an example, extensive duplication is found in the *Arabidopsis* genome, shown in Figure 11.7. Examples of duplication of parts of chromosomes to other locations in the same chromosome or a different chromosome are apparent, sometimes in the same orientation and sometimes in the reverse orientation. Partial gene sequences can also be duplicated or joined with sequences of other genes, resulting in proteins with an altered sequence domain structure.

Comparisons of proteomes of different organisms can identify the types of protein domain changes over evolutionary time and provide an indication as to what biological role the domain changes may have in a particular organism. As discussed in Chapters 5 and 10, proteins are modular and comprise separate structural and functional 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. 2000). Domain variations of these kinds are illustrated in Figure 11.6 with examples shown below in Figure 11.12. The assortment and reassortment of protein domains and domain swapping between proteins takes place in individual genomes and any beneficial results are passed along to newly derived species over evolutionary time.

The processes of domain assortment and gene duplication produce families of proteins in organisms. During speciation, a newly derived genome will inherit the families of ancestor organisms, but will also develop new ones to meet new 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. 11.4B), 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 are members of families made up of paralogs resulting from gene duplication events. The analysis 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 11.6A (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 6, with each alignment score receiving a statistical evaluation (*E* value). As shown in Figure 11.6B, 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 have a complicated history of domain shuffling and sequence change that is difficult to reconstruct.

*Cluster analysis.* In order to sort out relationships among all proteins that are found to be related in a series of searches of the types shown in Figure 11.6B, the proteins are subjected to a clustering analysis shown in Figure 11.6C. Only the relationships revealed by the hypothetical set of searches illustrated in Figure 11.6B 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  $E$  value from BLAST searches). The lower this value, the better the alignment. There will be a cutoff  $E$  value at which the matches in the BLAST search are no longer considered significant. A value of  $E > 0.01$ – $0.05$  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 each pair of aligned sequences. Deciding the criteria to be used for clustering proteins into the same or different clusters can be a difficult problem that is discussed in the box “Criteria for Clustering Protein Sequences.” Two clustering methods are described below.

*Clustering by making subgraphs.* Figure 11.6C indicates two ways of clustering related sequences based on the above criteria. Figure 11.6C, part i, illustrates matching a protein pair according to the criterion of the statistical significance of their alignment score. In the graph shown, 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  $E$  values (dotted edges). As weaker and weaker links are removed, the remaining combinations of vertices and edges represent the 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.* The method for clustering related sequences using the distance criterion for sequence relationships is shown in Figure 10.6C, part (ii). First, a group of related sequences found in the all-against-all proteome comparison is subjected to a multiple sequence alignment as described in Chapter 5. 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 7. These methods produce a tree (Fig. 11.6C, part ii, left) or a different representation of the tree called a dendrogram (Fig. 11.6C, 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 subtrees 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 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. Tekaiia et al. (1999) have used single-linkage clustering to determine the proportion of proteins in one organism that is shared with another organism to produce a genome tree. Methods of clustering data sets are further elaborated in Chapter 13.

*Core proteome.* The above types of all-against-all analyses provide an indication regarding the number



TABLE 11.4. Numbers of gene families and duplicated genes in model organisms

Organism	Total number of genes	Number of gene families <sup>a</sup>	Number of duplicated genes <sup>b</sup>
<i>Hemophilus influenzae</i> (bacteria)	1,709	1,425 <sup>c</sup>	284
<i>Saccharomyces cerevisiae</i> (yeast)	6,241	4,383	1,858
<i>Caenorhabditis elegans</i> (worm)	18,424	9,453	8,971
<i>Drosophila melanogaster</i> (fly)	13,600	8,065	5,536

Source: Rubin et al. (2000).

<sup>a</sup> 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.

<sup>b</sup> Count of number of duplicated genes within the protein family clusters.

<sup>c</sup> 178 families have paralogs.

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 11.4. 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.

#### *Between-Proteome Comparisons to Identify Orthologs, Gene Families, and Domains*

Comparisons between proteomes of different organisms are illustrated by the third panel in Figure 11.6A. 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 11.6B. 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.

### CRITERIA FOR CLUSTERING PROTEIN 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. This problem was also addressed in Chapter 5 (p. 192), where the ability of a scoring matrix or profile to distinguish known protein family members from nonmembers is analyzed. 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.

Other matching sequences in this class could 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 statistical value (*E* value or equivalent statistical score) 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, a yeast-versus-worm search.
2. Make a list of the worm sequences that give a high-scoring aligning with a low statistical value ( $10^{-10}$  to  $10^{-100}$ ) and include the yeast query sequence in the list.
3. From the list in step 2, choose a worm sequence and make a search of the yeast proteome, using the same criteria as in step 2.
4. Add any matching yeast sequence to the list 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 11.6C. 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 11.5. The numbers of sequence groups decrease about fivefold as the stringency of the statistical value of the alignment score 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.

*Family and domain analysis.* Extensive protein domain analyses have been performed for both prokaryotic and eukaryotic genomes (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 10.5. In a detailed analysis of the fly, worm, and yeast proteomes, 744 families and domains were common to all

TABLE 11.5. *Numbers of closely related yeast and worm sequences*

Cutoff <i>P</i> (or <i>E</i> ) 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 (©1998 AAAS).

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).

*Clusters of orthologous groups.* As described above, a pair of orthologous genes in two organisms shares so much sequence similarity that the genes may be assumed to have arisen from a common ancestor gene. When entire proteomes of the two organisms are available, orthologs may be identified as the most-alike sequences in reciprocal proteome similarity searches, as described in the above section. 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 match not only the orthologous sequence in the second proteome, but also these other paralogous sequences. The objective of the clusters of orthologous groups (COGs) 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 clustered together to form a COG that includes both orthologs and paralogs. These clusters correspond to classes of metabolic functions. 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. 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. 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 11.6C, part i. Paralogs defined by sets of three matching sequences in the selected organisms were also added to these clusters. The proteins encoded by many prokaryotic organisms have been analyzed for COG relationships (Koonin et al. 1998). A COGs 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 some eukaryotic organisms, the complete genome sequence is not available. What is available 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 cells or 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 6. A more detailed analysis may then be made by cloning and sequencing the intact full-length 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. 41). 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 11.6B. Hence, it is not possible to impose the requirement of alignment with 60–80% of the query sequence, which greatly improves 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 6. Another method is to perform an exhaustive search for a protein family, described next.

*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-

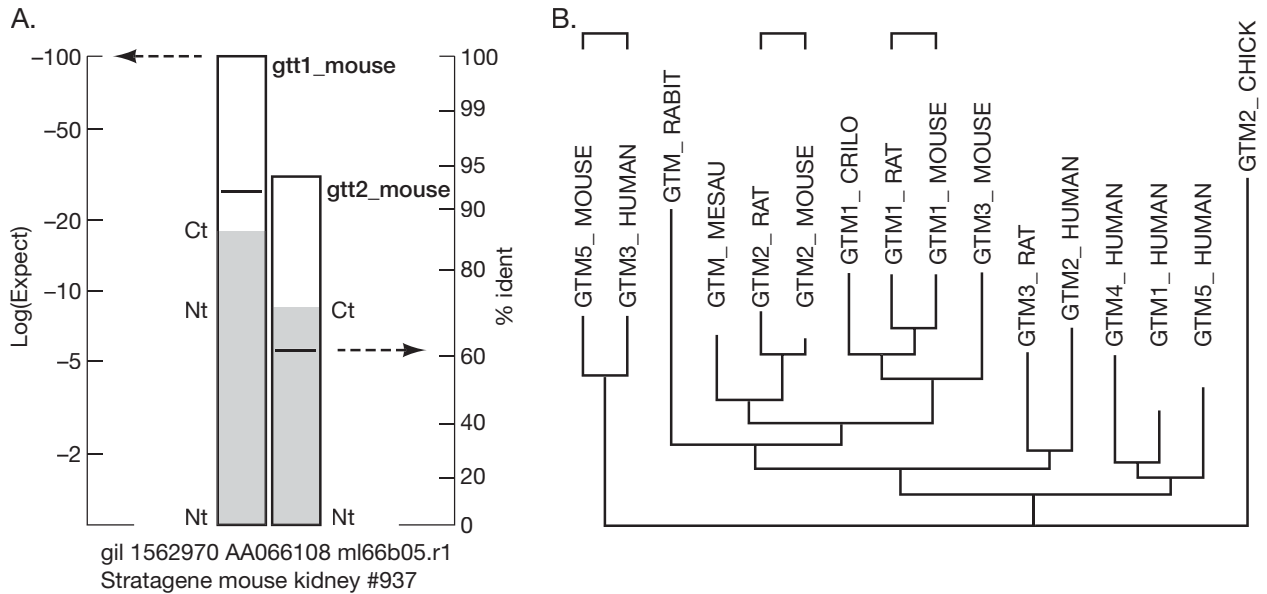


FIGURE 11.8. Prediction of paralogs and orthologs by searches of EST databases by gene panning (Retief et al. 1999). In this analysis, one class of glutathione transferase family members was used as queries to search mammalian EST databases for highly significant matches using TFASTY3 (Chapter 6). FAST\_PAN is a Perl-script program (use as Web search term) that automatically searches EST databases as they are updated and compiles the results of the search. (A) Display of protein class matched (shaded), log Expect value (height of bar), length of query sequence matched (height of shaded 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.)

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 11.8. A chart showing the  $E$  value, percent identity, fraction of query sequence matched, and type of query matched (color coded) is shown in Figure 11.8A.

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 represen-

tatives for a similarity search of mammalian EST databases for related sequences and to decide which of these sequences were orthologs and which were paralogs.

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. 98). 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 7 (see flowchart for Chapter 7), is suitable for a divergent but recently evolved group of sequences. The predicted tree shown in Figure 11.8B predicts those pairs of sequences that are likely to be orthologous.

### Ancient Conserved Regions

Phylogenetically diverse groups of organisms have been analyzed for the presence of conserved proteins and protein domains called ancient conserved regions or ACRs, domains that have been conserved over long periods of evolutionary time (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 protein 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 could be estimated by the proportion of genes that match database sequence of known function. For the hyperthermophilic archaeal organism *Pyrococcus 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 50–70% (Koonin and Galperin 1997; Koonin et al. 1998). 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 (HGT), the acquisition of genetic material from a different organism. (For a review of this topic, see Bushman 2002.) The transferred material then becomes a permanent addition to the recipient genome and subsequently is transmitted vertically. Although these exchanges do not occur very often on a generation-to-generation basis, a significant number of such exchanges 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).

HGT is a significant source of genome variation in bacteria (Ochmann et al. 2000; Bushman 2002 for overview), 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 HGT 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 9. 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 HGT 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 HGT sequences also varies considerably, roughly in proportion to genome size. A total of 12.8% of the genome of *E. coli* is due to HGT 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

## GENOME ANNOTATION

Accurate annotation of genome sequences is an important step in genome analysis. An initial step in annotation is predicting the location of genes that encode proteins. Any significant alignment of the predicted protein sequences to sequences of known function is then found 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. After the genes have been identified, the genome sequence can be annotated with a great deal of information about the genome sequence—the location and structure of genes (exons and introns), links to related sequences such as cDNA and encoded protein sequences, detailed information about the function of genes and the biochemistry and structure of proteins and other gene products, literature links, links to genetic maps, location of sequence repeats including transposable elements, the location of sequence tagged sites (STS) for amplification of local genome sequences by polymerase chain reaction (PCR), and sequence polymorphisms in the population (SNPs). These annotated sequences and their Web links to a variety of related information are a valuable resource to the biological community.

not be expected to carry any extra unnecessary genetic baggage.

HGT 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 HGT. Hence, clustering in combination with HGT provides an evolutionary mechanism for altering metabolic pathways in diverse organisms