CHAPTER SEVENTEEN

Genome Mapping, Assembly, and Comparison

Genomics is the study of genomes. Genomic studies are characterized by simultaneous analysis of a large number of genes using automated data gathering tools. The topics of genomics range from genome mapping, sequencing, and functional genomic analysis to comparative genomic analysis. The advent of genomics and the ensuing explosion of sequence information are the main driving force behind the rapid development of bioinformatics today.

Genomic study can be tentatively divided into structural genomics and functional genomics. *Structural genomics* refers to the initial phase of genome analysis, which includes construction of genetic and physical maps of a genome, identification of genes, annotation of gene features, and comparison of genome structures. This is the major theme of discussion of this chapter. However, it should to be mentioned that the term *structural genomics* has already been used by a structural biology group for an initiative to determine three-dimensional structures of all proteins in a cell. Strictly speaking, the initiative of structural determination of proteins falls within the realm of *structural proteomics* and should not be confused as a subdiscipline of genomics. The structure genomics discussed herein mainly deals with structures of genome sequences. *Functional genomics* refers to the analysis of global gene expression and gene functions in a genome, which is discussed in Chapter 18.

GENOME MAPPING

The first step to understanding a genome structure is through genome mapping, which is a process of identifying relative locations of genes, mutations or traits on a chromosome. A low-resolution approach to mapping genomes is to describe the order and relative distances of genetic markers on a chromosome. *Genetic markers* are identifiable portions of a chromosome whose inheritance patterns can be followed. For many eukaryotes, genetic markers represent morphologic phenotypes. In addition to genetic linkage maps, there are also other types of genome maps such as physical maps and cytologic maps, which describe genomes at different levels of resolution. Their relations relative to the DNA sequence on a chromosome are illustrated in Figure 17.1. More details of each type of genome maps are discussed next.

Genetic linkage maps, also called *genetic maps*, identify the relative positions of genetic markers on a chromosome and are based on how frequent the markers are inherited together. The rationale behind genetic mapping is that the closer the two

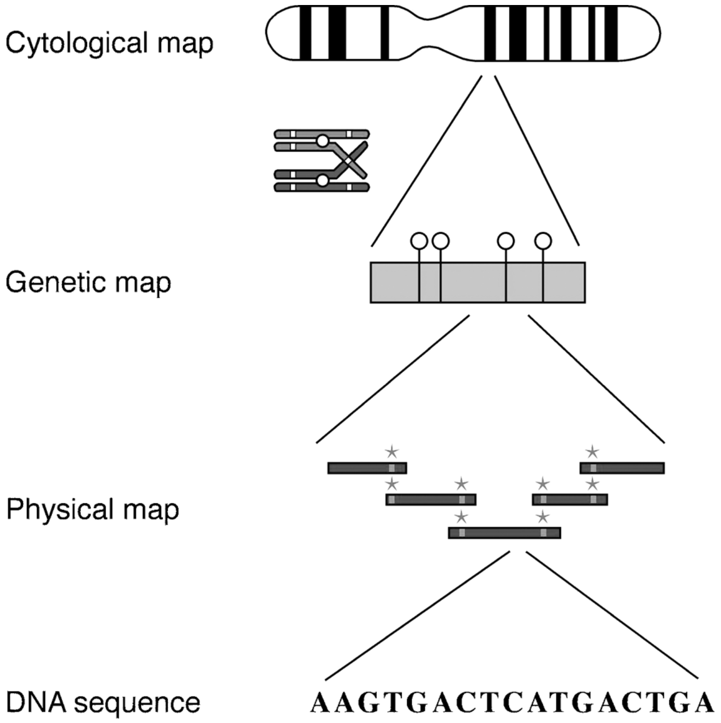

Figure 17.1: Overview of various genome maps relative to the genomic DNA sequence. The maps represent different levels of resolution to describe a genome using genetic markers. Cytologic maps are obtained microscopically. Genetic maps (*grey bar*) are obtained through genetic crossing experiments in which chromosome recombinations are analyzed. Physical maps are obtained from overlapping clones identified by hybridizing the clone fragments (grey bars) with common probes (grey asterisks).

genetic markers are, the more likely it is that they are inherited together and are not separated in a genetic crossing event. The distance between the two genetic markers is measured in centiMorgans (cM), which is the frequency of recombination of genetic markers. One centiMorgan is defined as one percentage of the total recombination events when separation of the two genetic markers is observed in a genetic crossing experiment. One centiMorgan is approximately 1 Mb in humans and 0.5 Mb in *Drosophila*.

Physical maps are maps of locations of identifiable landmarks on a genomic DNA regardless of inheritance patterns. The distance between genetic markers is measured directly as kilobases (Kb) or megabases (Mb). Because the distance is expressed in physical units, it is more accurate and reliable than centiMorgans used in genetic maps. Physical maps are constructed by using a chromosome walking technique, which uses a number of radiolabeled probes to hybridize to a library of DNA clone fragments. By identifying overlapping clones probed by common probes, a relative order of the cloned fragments can be established.

Cytologic maps refer to banding patterns seen on stained chromosomes, which can be directly observed under a microscope. The observable light and dark bands are the visually distinct markers on a chromosome. A genetic marker can be associated with a specific chromosomal band or region. The banding patterns, however, are not always constant and are subject to change depending on the extent of chromosomal contraction. Thus, cytologic maps can be considered to be of very low resolution and hence somewhat inaccurate physical maps. The distance between two bands is expressed in relative units (Dustin units).

GENOME SEQUENCING

The highest resolution genome map is the genomic DNA sequence that can be considered as a type of physical map describing a genome at the single base-pair level. DNA sequencing is now routinely carried out using the Sanger method. This involves the use of DNA polymerases to synthesize DNA chains of varying lengths. The DNA synthesis is stopped by adding dideoxynucleotides. The dideoxynucleotides are labeled with fluorescent dyes, which terminate the DNA synthesis at positions containing all four bases, resulting in nested fragments that vary in length by a single base. When the labeled DNA is subjected to electrophoresis, the banding patterns in the gel reveal the DNA sequence.

The fluorescent traces of the DNA sequences are read by a computer program that assigns bases for each peak in a chromatogram. This process is called *base calling*. Automated base calling may generate errors and human intervention is often required to correct the sequence calls.

There are two major strategies for whole genome sequencing: the shotgun approach and the hierarchical approach. The *shotgun approach* randomly sequences clones from both ends of cloned DNA. This approach generates a large number of sequenced DNA fragments. The number of random fragments has to be very large, so large that the DNA fragments overlap sufficiently to cover the entire genome. This approach does not require knowledge of physical mapping of the clone fragments, but rather a robust computer assembly program to join the pieces of random fragments into a single, whole-genome sequence. Generally, the genome has to be redundantly sequenced in such a way that the overall length of the fragments covers the entire genome multiple times. This is designed to minimize sequencing errors and ensure correct assembly of a contiguous sequence. Overlapping sequences with an overall length of six to ten times the genome size are normally obtained for this purpose.

Despite the multiple coverage, sometimes certain genomic regions remain unsequenced, mainly owing to cloning difficulties. In such cases, the remainder gap sequences can be obtained through extending sequences from regions of known genomic sequences using a more traditional PCR technique, which requires the use of custom primers and performs genome walking in a stepwise fashion. This step of genome sequencing is also known as *finishing*, which is followed by computational assembly of all the sequence data into a final complete genome.

The hierarchical genome sequencing approach is similar to the shotgun approach, but on a smaller scale. The chromosomes are initially mapped using the physical mapping strategy. Longer fragments of genomic DNA (100 to 300 kB) are obtained and cloned into a high-capacity bacterial vector called bacterial artificial chromosome (BAC). Based on the results of physical mapping, the locations and orders of the BAC clones on a chromosome can be determined. By successively sequencing adjacent BAC clone fragments, the entire genome can be covered. The complete sequence of each individual BAC clone can be obtained using the shotgun approach. Overlapping BAC clones are subsequently assembled into an entire genome sequence. Major differences between the hierarchical and the full shotgun approaches are shown in Figure 17.2.

During the era of human genome sequencing, there was a heated debate on the merits of each of the two strategies. In fact, there are advantages and disadvantages in either. The hierarchical approach is slower and more costly than the shotgun approach because it involves an initial clone-based physical mapping step. However, once the map is generated, assembly of the whole genome becomes relatively easy and less error prone. In contrast, the whole genome shotgun approach can produce a draft sequence very rapidly because it is based on the direct sequencing approach. However, it is computationally very demanding to assemble the short random fragments. Although the approach has been successfully employed in sequencing small microbial genomes, for a complex eukaryotic genome that contains high levels of repetitive sequences, such as the human genome, the full shotgun approach becomes less accurate and tends to leave more "holes" in the final assembled sequence than the hierarchical approach. Current genome sequencing of large organisms often uses a combination of both approaches.

GENOME SEQUENCE ASSEMBLY

As described, initial DNA sequencing reactions generate short sequence reads from DNA clones. The average length of the reads is about 500 bases. To assemble a whole genome sequence, these short fragments are joined to form larger fragments after removing overlaps. These longer, merged sequences are termed *contigs*, which are usually 5,000 to 10,000 bases long. A number of overlapping contigs can be further merged to form scaffolds (30,000–50,000 bases, also called *supercontigs*), which are unidirectionally oriented along a physical map of a chromosome (Fig. 17.3). Overlapping scaffolds are then connected to create the final highest resolution map of the genome.

Correct identification of overlaps and assembly of the sequence reads into contigs are like joining jigsaw puzzles, which can be very computationally intensive when dealing with data at the whole-genome level. The major challenges in genome assembly are sequence errors, contamination by bacterial vectors, and repetitive sequence regions. Sequence errors can often be corrected by drawing a consensus from an alignment of multiple overlapped sequences. Bacterial vector sequences can be removed using filtering programs prior to assembly. To overcome the problem of sequence repeats, programs such as RepeatMasker (see Chapter 4) can be used to detect and

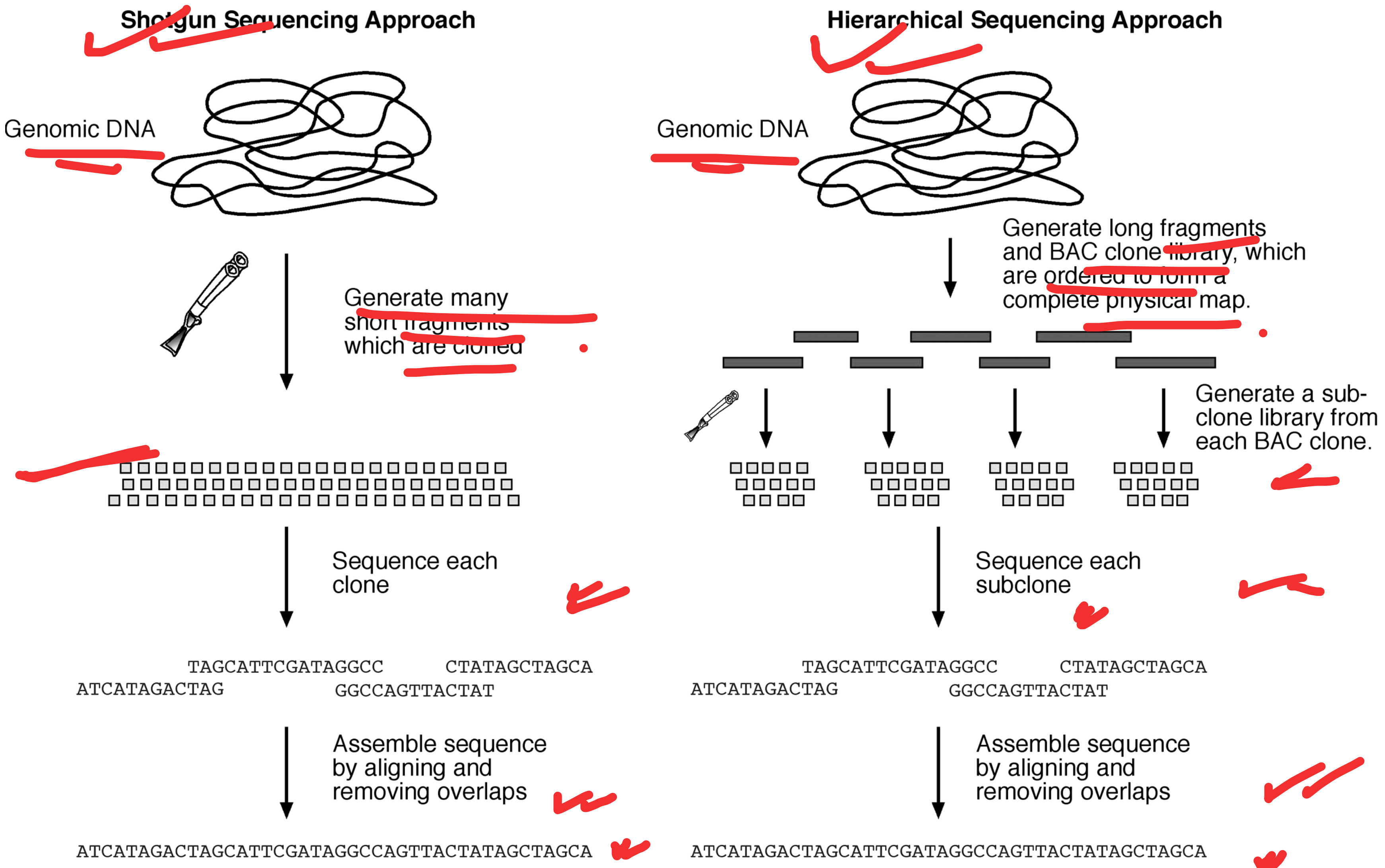
Shotgun Sequencing Approach
Genomic DNA
Generate many short fragments which are cloned
Sequence each clone
TAGCATTCGATAGGCC
CTATAGCTAGCA
ATCATAGACTAG
GGCCAGTTACTAT
Assemble sequence by aligning and removing overlaps
ATCATAGACTAGCATTCGATAGGCCAGTTACTATAGCTAGCA
Hierarchical Sequencing Approach
Genomic DNA
Generate long fragments and BAC clone library, which are ordered to form a complete physical map.
Generate a sub-clone library from each BAC clone.
Sequence each subclone
TAGCATTCGATAGGCC
CTATAGCTAGCA
ATCATAGACTAG
GGCCAGTTACTAT
Assemble sequence by aligning and removing overlaps
ATCATAGACTAGCATTCGATAGGCCAGTTACTATAGCTAGCA

Figure 17.2: Schematic comparison of the two whole genome sequencing approaches. The full shotgun approach cuts DNA into ∼2 kB fragments, which are cloned into small vectors and sequenced individually. The sequenced fragments are then put together into a final sequence in one step. The hierarchical approach cuts DNA into intermediate size fragments (∼150 kB). The DNA fragments are cloned into BACs. A physical map has to be built based on the BAC clones. Each BAC clone is then subject to the shotgun approach.

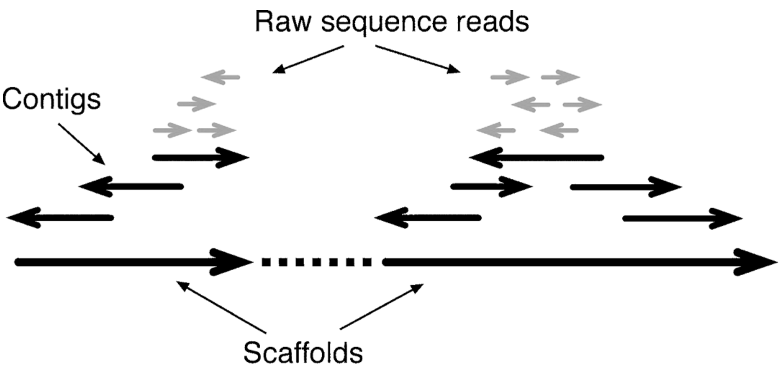

Figure 17.3: Schematic diagram showing three different levels of sequence assembly. Contigs are formed by combining raw sequence reads of various orientations after removing overlaps. Scaffolds are assembled from contigs and oriented unidirectionally on a chromosome. Because sequence fragments generated can be in either of the DNA strands, arrows are used to represent directionality of the sequences written in 5′ → 3′ orientation.

mask repeats. Additional constraints on the sequence reads can be applied to avoid misasembly caused by repeat sequences.

A commonly used constraint to avoid errors caused by sequence repeats is the so-called forward–reverse constraint. When a sequence is generated from both ends of a single clone, the distance between the two opposing fragments of a clone is fixed to a certain range, meaning that they are always separated by a distance defined by a clone length (normally 1,000 to 9,000 bases). When the constraint is applied, even when one of the fragments has a perfect match with a repetitive element outside the range, it is not able to be moved to that location to cause missassembly. An example of assembly with or without applying the forward–reverse constraints is shown in Figure 17.4.

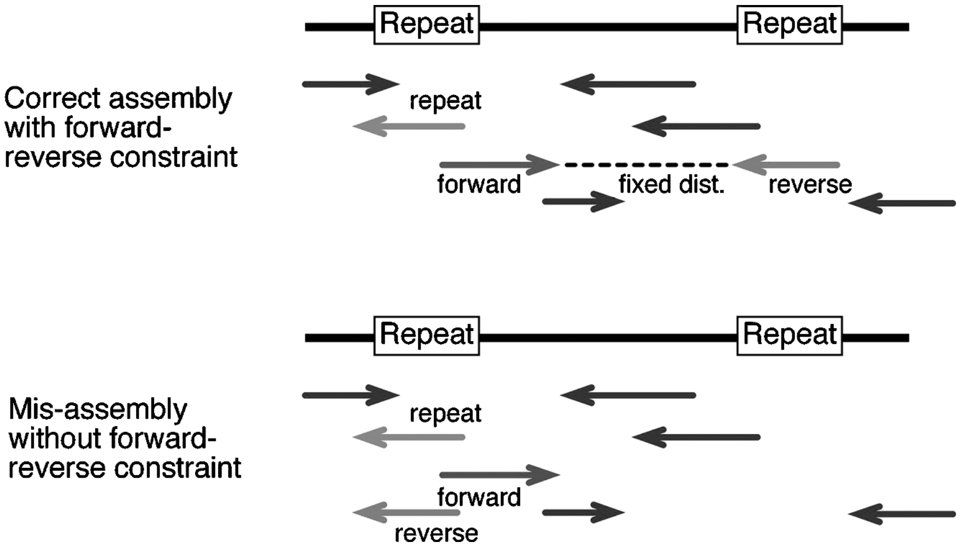
Repeat
Repeat
Correct assembly with forward-reverse constraint
repeat
forward
fixed dist.
reverse
Repeat
Repeat
Mis-assembly without forward-reverse constraint
repeat
forward
reverse

Figure 17.4: Example of sequence assembly with or without applying forward–reverse constraint, which fixes the sequence distance from both ends of a subclone. Without the restraint, the red fragment is misassembled due to matches of repetitive element in the middle of a fragment (see color plate section).

Base Calling and Assembly Programs

The first step toward genome assembly is to derive base calls and assign associated quality scores. The next step is to assemble the sequence reads into contiguous sequences. This step includes identifying overlaps between sequence fragments, assigning the order of the fragments and deriving a consensus of an overall sequence. Assembling all shotgun fragments into a full genome is a computationally very challenging step. There are a variety of programs available for processing the raw sequence data. The following is a selection of base calling and assembly programs commonly used in genome sequencing projects.

Phred (www.phrap.org/) is a UNIX program for base calling. It uses a Fourier analysis to resolve fluorescence traces and predict actual peak locations of bases. It also gives a probability score for each base call that may be attributable to error. The commonly accepted score threshold is twenty, which corresponds to a 1% chance of error. The higher the score, the better the quality of the sequence reads. If the score value falls below the threshold, human intervention is required.

Phrap (www.phrap.org/) is a UNIX program for sequence assembly. It takes Phred base-call files with quality scores as input and aligns individual fragments in a pairwise fashion using the Smith–Waterman algorithm. The base quality information is taken into account during the pairwise alignment. After all the pairwise sequence similarity is identified, the program performs assembly by progressively merging sequence pairs with decreasing similarity scores while removing overlapped regions. Consensus contigs are derived after joining all possible overlapped reads.

VecScreen (www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) is a web-based program that helps detect contaminating bacterial vector sequences. It scans an input nucleotide sequence and compares it with a database of known vector sequences by using the BLAST program.

TIGR Assembler (www.tigr.org/) is a UNIX program from TIGR for assembly of large shotgun sequence fragments. It treats the sequence input as clean reads without consideration of the sequence quality. A main feature of the program is the application of the forward–reverse constraints to avoid misassembly caused by sequence repeats. The sequence alignment in the assembly stage is performed using the Smith–Waterman algorithm.

ARACHNE (www-genome.wi.mit.edu/wga/) is a free UNIX program for the assembly of whole-genome shotgun reads. Its unique features include using a heuristic approach similar to FASTA to align overlapping fragments, evaluating alignments using statistical scores, correcting sequencing errors based on multiple sequence alignment, and using forward–reverse constraints. It accepts base calls with associated quality scores assigned by Phred as input and produces scaffolds or a fully assembled genome.

EULER (http://nbcr.sdsc.edu/euler/) is an assembly algorithm that uses a Eulerian Superpath approach, which is a polynomial algorithm for solving puzzles such as the famous "traveling salesman problem": finding the shortest path of visiting a given number of cities exactly once and returning to the starting point. In this approach, a sequence fragment is broken down to tuples of twenty nucleotides. The tuples are distributed in a diagram with numerous nodes that are all interconnected. The tuples are converted to binary vectors in the nodes. By using a Viterbi algorithm (see Chapter 6), the shortest path among the vectors can be found, which is the best way to connect the tuples into a full sequence. Because this approach does not directly rely on detecting overlaps, it may be advantageous in assembling sequences with repeat motifs.

GENOME ANNOTATION

Before the assembled sequence is deposited into a database, it has to be analyzed for useful biological features. The genome annotation process provides comments for the features. This involves two steps: gene prediction and functional assignment. Some examples of finished gene annotations in GenBank have been described in the Biological Database section (see Chapter 2). The following example illustrates the overall process employed in annotating the human genome.

human genome gene annotat

As a real-world example, gene annotation of the human genome employs a combination of theoretical prediction and experimental verification. Gene structures are first predicted by ab initio exon prediction programs such as GenScan or FgenesH (see Chapter 8). The predictions are verified by BLAST searches against a sequence database. The predicted genes are further compared with experimentally determined cDNA and EST sequences using the pairwise alignment programs such as GeneWise, Spidey, SIM4, and EST2Genome. All predictions are manually checked by human curators. Once open reading frames are determined, functional assignment of the encoded proteins is carried out by homology searching using BLAST searches against a protein database. Further functional descriptions are added by searching protein motif and domain databases such as Pfam and InterPro (see Chapter 7) as well as by relying on published literature.

Gene Ontology

A problem arises when using existing literature because the description of a gene function uses natural language, which is often ambiguous and imprecise. Researchers working on different organisms tend to apply different terms to the same type of genes or proteins. Alternatively, the same terminology used in different organisms may actually refer to different genes or proteins. Therefore, there is a need to standardize protein functional descriptions. This demand has spurred the development of the gene ontology (GO) project, which uses a limited vocabulary to describe molecular functions, biological processes, and cellular components. The controlled vocabulary is organized such that a protein function is linked to the cellular function through a hierarchy of descriptions with increasing specificity. The top of the hierarchy provides an overall picture of the functional class, whereas the lower level

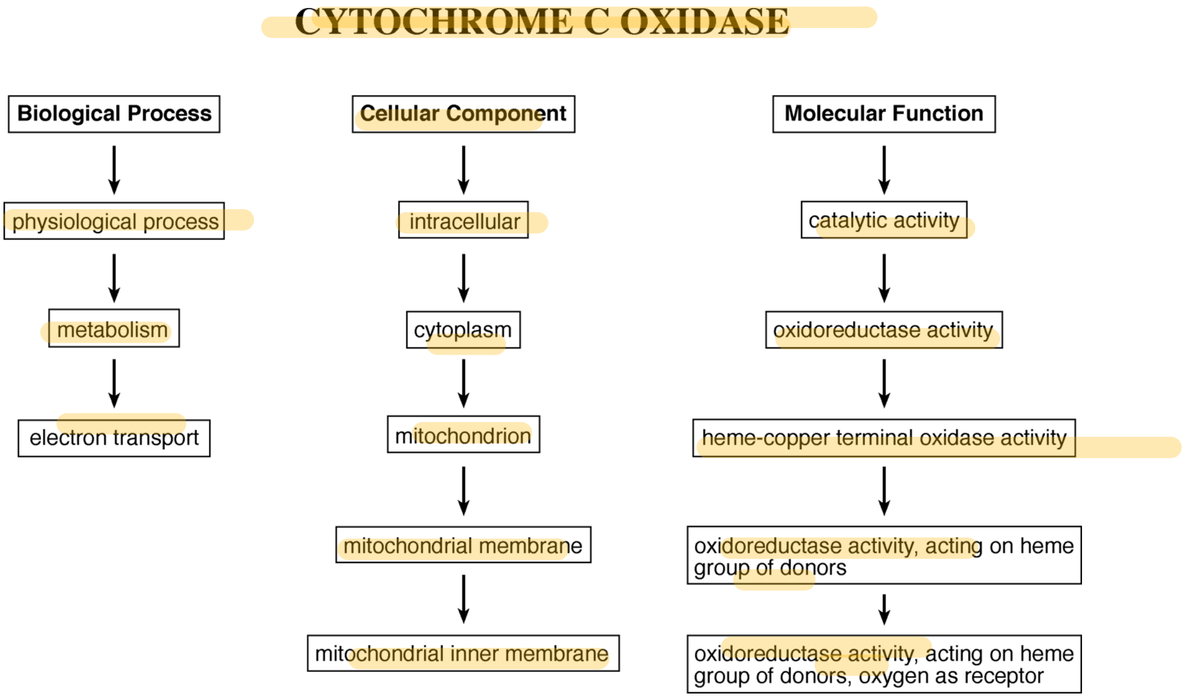

Figure 17.5: Example of GO annotation for cytochrome *c* oxidase. The functional and structural terms are arranged in three categories with a number of hierarchies indicating the levels of conceptual associations of protein functions.

in the hierarchy specifies more precisely the functional role. This way, protein functionality can be defined in a standardized and unambiguous way.

A GO description of a protein provides three sets of information: *biological process*, *cellular component*, and *molecular function*, each of which uses a unique set of nonoverlapping vocabularies. The standardization of the names, activities, and associated pathways provides consistency in describing overall protein functions and facilitates grouping of proteins of related functions. A database searching using GO for a particular protein can easily bring up other proteins of related functions in much the same way as using a thesaurus. Using GO, a genome annotator can assign functional properties of a gene product at different hierarchical levels, depending on how much is known about the gene product.

At present, the GO databases have been developed for a number of model organisms by an international consortium, in which each gene is associated with a hierarchy of GO terms. These have greatly facilitated genome annotation efforts. A good introduction of gene ontology can be found at www.geneontology.org. An example of GO annotation for cytochrome *c* oxidase is shown in Figure 17.5.

Automated Genome Annotation

With the genome sequence data being generated at an exponential rate, there is a need to develop fast and automated methods to annotate the genomic sequences. The automated approach relies on homology detection, which is essentially heuristic sequence similarity searching. If a newly sequenced gene or its gene product has significant matches with a database sequence beyond a certain threshold, a transfer of functional assignment is taking place. In addition to sequence matching at the full length, detection of conserved motifs often offers additional functional clues.

Because using a single database searching method is often incomplete and error prone, automated methods have to mimic the manual process, which takes into consideration multiple lines of evidence in assigning a gene function, to minimize errors. The following algorithm is an example that goes a step beyond examining sequence similarity and provides functional annotations based on multiple protein characteristics.

GeneQuiz (http://jura.ebi.ac.uk:8765/ext-genequiz/) is a web server for protein sequence annotation. The program compares a query sequence against databases using BLAST and FASTA to identify homologs with high similarities. In addition, it performs domain analysis using the PROSITE and Blocks databases (see Chapter 7) as well as analysis of secondary structures and supersecondary structures that includes prediction of coiled coils and transmembrane helices. Multiple search and analysis results are compiled to produce a summary of protein function with an assigned confidence level (clear, tentative, marginal, and negligible).

Annotation of Hypothetical Proteins

Although a large number of genes and proteins can be assigned functions by the sequence similarity based approach, about 40% of the genes from newly sequenced genomes have no known functions and can only be annotated as genes encoding "hypothetical proteins." Experimental discovery of the functions of these genes and proteins is often time consuming and difficult because of lack of hypotheses to design experiments. In this situation, more advanced tools can be used for functional predictions by searching for remote homologs.

One way to obtain functional hints of genes encoding hypothetical proteins is by searching for remote homologs in databases. Detecting remote homologs typically involves combined searches of protein motifs and domains and prediction for secondary and tertiary structures. Conserved functional sites can be identified by profile and hidden Markov model–based motif and domain search tools such as SMART and InterPro (see Chapter 7). The prediction can also be performed using structure-based approaches such as threading and fold recognition (see Chapter 15). If the distant homologs detected using the structural approach are linked with well-defined functions, a broad functional class of the query protein if not the precise function of the protein can be inferred. In addition, prediction results for subcellular localization, protein–protein interactions can provide further functional hints (see Chapter 19).

These suggestions do not guarantee to provide correct annotations for the "hypothetical proteins," but they may provide critical hypotheses of the protein function that can be tested in the laboratory. The remote homology detection helps to shed light on the possible functions of the proteins that previously have no functional information at all. Thus, the bioinformatic analysis can spur an important advance in knowledge in many cases. Some hypothetical proteins, because of their novel

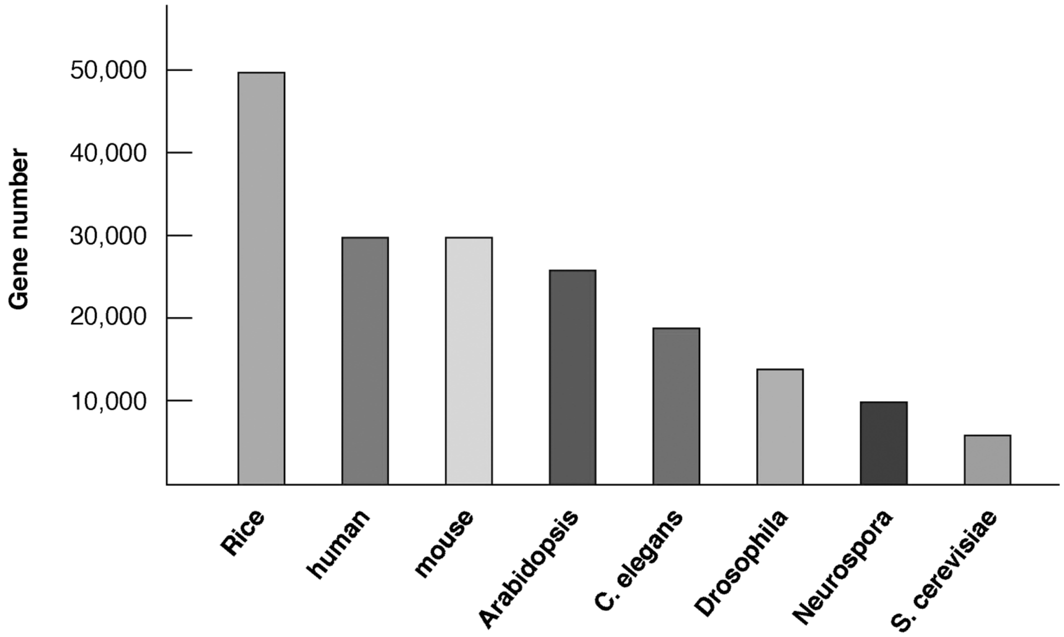

Figure 17.6: Gene numbers estimated from several sequenced eukaryotic genomes. (Data from Integrated Genomics Online Database http://ergo.integratedgenomics.com/GOLD/.)

structural folds, still cannot be predicted even with the advanced bioinformatics approaches and remain challenges for both experimental and computational work.

How Many Genes in a Genome?

One of the main tasks of genome annotation is to try to give a precise account of the total number of genes in a genome. This may be more feasible for prokaryotes as their gene structures are relatively simple. However, the number of genes in eukaryotic genomes, in particular the human genome, has been a subject of debate. This is mainly because of the complex structures of these genomes, which obscure gene prediction. Before the human genome sequencing was completed, the estimated gene numbers ranged from 20,000 to 120,000. Since the completion of the sequencing of the human genome, with the use of more sophisticated gene finding programs, the total number of human genes now dropped to close to 25,000 to 30,000. Although no exact number is agreed upon by all researchers, it is now widely believed that the total number of human genes will be no more than 30,000. This compares to estimates of 50,000 in rice, 30,000 in mouse, 26,000 in *Arabidopsis*, 18,400 in *C. elegans*, and 6,200 in yeast (Fig. 17.6).

The discovery of the low gene count in humans may be ego defeating to some as they realize that humans are only five times more complex than baker's yeast and apparently equally as complex as the mouse. What is worse, the food in their rice bowls has twice as many genes. The finding seriously challenges the view that humans are a superior species on Earth. As in many discoveries in scientific history, such as Darwin's evolutionary theory suggesting that humans arose from a "subhuman" ancestor, recent genomic discoveries have moved humans further away from this exalted status. However, before we are overwhelmed by the humble realization, we should also realize that the complexity of an organism simply cannot be represented by gene numbers. As will soon become clear, gene expression and regulation, protein expression, modification, and interactions all contribute to the overall complexity of an organism.

Genome Economy

One level of genetic complexity is manifested at the protein expression level in which there are often more expressed proteins than genes available to code for them. For example, in humans, there are more than 100,000 proteins expressed based on EST analysis (see Chapter 18) compared to no more than 30,000 genes. If the "one gene, one protein" paradigm holds true, how could this discrepancy exist? Where does the extra coding power come from?

The answer lies in "genome economy," a phenomenon of synthesizing more proteins from fewer genes. This is a major strategy that eukaryotic organisms use to achieve a myriad of phenotypic diversities. There are many underlying genetic mechanisms to help account for genome economy. A major mechanism responsible for the protein diversity is *alternative splicing*, which refers to the splicing event that joins different exons from a single gene to form different transcripts. A related mechanism, known as *exon shuffling*, which joins exons from different genes to generate more transcripts, is also common in eukaryotes. It is known that, in humans, about two thirds of the genes exhibit alternative splicing and exon shuffling during expression, generating 90% of the total proteins. In *Drosophila*, the *DSCAM* gene contains 115 exons that can be alternatively spliced to produce 38,000 different proteins. This remarkable ability to generate protein diversity and new functions highlights the true complexity of a genome. It also illustrates the evolutionary significance of introns in eukaryotic genes, which serve as spacers that make the molecular recombination possible.

There are more surprising mechanisms responsible for genome economy. For example, trans-splicing can occur between RNAs produced from both DNA strands. In the *Drosophila mdg4* mutant, RNA transcribed from four exons in the sense strand and two exons in the antisense strand are joined to form a single mRNA. With different exon combinations, four different proteins can be produced. In some circumstances, one mRNA transcript can lead to the translation of more than one protein. For example, human dentin phosphoprotein and dentin sialoprotein are proteins involved in tooth formation. An mRNA transcript that includes coding regions from both proteins is translated into a precursor protein that is cleaved to produce two different mature proteins. Another situation, called "gene within gene," can be found in a gene for human prostate-specific antigen (PSA). In addition to regular PSA, humans can produce a similar protein, called PSA-LM, that functions antagonistically to PSA and is important for prostate cancer diagnosis. PSA-LM turns out to be encoded by the fourth intron of the PSA gene.

These are just a few known mechanisms of condensing the coding potential of genomic DNA to achieve increased protein diversity. From a bioinformatics point of view, this makes gene prediction based on computational approaches all the more complicated. It also highlights one of the challenges that faces software program developers today. A number of databases have recently been established to archive alternatively spliced forms of eukaryotic genes. The following is one such example for human genes.

ProSplicer (http://prosplicer.mbc.nctu.edu.tw/) is a web-based database of human alternative spliced transcripts. The spliced variants are identified by aligning each known human protein, mRNA, and EST sequence against the genomic sequence using the SIM4 and TBLASTN program. The three sets of alignment are compiled to derive alternative splice forms. The database organizes data by tissue types and can be searched using keywords.

COMPARATIVE GENOMICS

Comparison of whole genomes from different organisms is comparative genomics, which includes comparison of gene number, gene location, and gene content from these genomes. The comparison helps to reveal the extent of conservation among genomes, which will provide insights into the mechanism of genome evolution and gene transfer among genomes. It helps to understand the pattern of acquisition of foreign genes through lateral gene transfer. It also helps to reveal the core set of genes common among different genomes, which should correspond to the genes that are crucial for survival. This knowledge can be potentially useful in future metabolic pathway engineering.

As alluded to previously, the main themes of comparative genomics include whole genome alignment, comparing gene order between genomes, constructing minimal genomes, and lateral gene transfer among genomes, each of which is discussed in more detail.

Whole Genome Alignment

With an ever-increasing number of genome sequences available, it becomes imperative to understand sequence conservation between genomes, which often helps to reveal the presence of conserved functional elements. This can be accomplished through direct genome comparison or genome alignment. The alignment at the genome level is fundamentally no different from the basic sequence alignment described in Chapters 3, 4, and 5. However, alignment of extremely large sequences presents new complexities owing to the sheer size of the sequences. Regular alignment programs tend to be error prone and inefficient when dealing with long stretches of DNA containing hundreds or thousands of genes. Another challenge of genome alignment is effective visualization of alignment results. Because it is obviously difficult to sift through and make sense of the extremely large alignments, a graphical representation is a must for interpretation of the result. Therefore, specific alignment algorithms are needed to deal with the unique challenges of whole genome alignment. A number of alignment programs for "super-long" DNA sequences are described next.

MUMmer (Maximal Unique Match, www.tigr.org/tigr-scripts/CMR2/webmum/mumplot) is a free UNIX program from TIGR for alignment of two entire genome sequences and comparison of the locations of orthologs. The program is essentially a modified BLAST, which, in the seeding step (see Chapter 4), finds the longest approximate matches that include mismatches instead of finding exact *k*-mer matches as in regular BLAST. The result of the alignment of whole genomes is shown as a dot plot with lines of connected dots to indicate collinearity of genes. It is optimized for pairwise comparison of closely related microbial genomes.

BLASTZ (http://bio.cse.psu.edu/) is a UNIX program modified from BLAST to do pairwise alignment of very large genomic DNA sequences. The modified BLAST program first masks repetitive sequences and searches for closely matched "words," which are defined as twelve identical matches within a stretch of nineteen nucleotides. The words serve as seeds for extension of alignment in both directions until the scores drop below a certain threshold. Nearby aligned regions are joined by using a weighted scheme that employs a unique gap penalty scheme that tolerates minor variations such as transitions in the seeding step of the alignment construction to increase its sensitivity.

LAGAN (Limited Area Global Alignment of Nucleotides; http://lagan.stanford.edu/) is a web-based program designed for pairwise alignment of large genomes. It first finds anchors between two genomic sequences using an algorithm that identifies short, exactly matching words. Regions that have high density of words are selected as anchors. The alignments around the anchors are built using the Needleman–Wunsch global alignment algorithm. Nearby aligned regions are further connected using the same algorithm. The unique feature of this program is that it is able to take into account degeneracy of the genetic codes and is therefore able to handle more distantly related genomes. Multi-LAGAN, an extension of LAGAN, available from the same website, performs multiple alignment of genomes using a progressive approach similar to that used in Clustal (see Chapter 5).

PipMaker (http://bio.cse.psu.edu/cgi-bin/pipmaker?basic) is a web server using the BLASTZ heuristic method to find similar regions in two DNA sequences. It produces a textual output of the alignment result and also a graphical output that presents the alignment as a percent identity plot as well as a dot plot. For comparing multiple genomes, MultiPipMaker is available from the same site.

MAVID (http://baboon.math.berkeley.edu/mavid/) is a web-based program for aligning multiple large DNA sequences. MAVID is based on a progressive alignment algorithm similar to Clustal. It produces an NJ tree as a guide tree. The sequences are aligned recursively using a heuristic pairwise alignment program called AVID. AVID works by first selecting anchors using the Smith–Waterman algorithm and then building alignments for the sequences between nearby anchors. Connected alignments are treated as new anchors for building longer alignments. The process is repeated iteratively until the entire sequence pair including weakly conserved regions are aligned.

GenomeVista (http://pipeline.lbl.gov/cgi-bin/GenomeVista) is a database searching program that searches against the human, mouse, rat, or *Drosophila* genomes using a large piece of DNA as query. It uses a program called BLAT to find anchors and extends the alignment from the anchors using AVID. (BLAT is a fast local alignment algorithm that aligns short sequences of forty bases with more than 95% similarity.) It produces a graphical output that shows the sequence percent identity.

Finding a Minimal Genome

One of the goals of genome comparison is to understand what constitutes a minimal genome, which is a minimal set of genes required for maintaining a free-living cellular organism. Finding minimal genomes helps provide an understanding of genes constituting key metabolic pathways, which are critical for a cell's survival. This analysis involves identification of orthologous genes shared between a number of divergent genomes.

Coregenes (http://pasteur.atcc.org:8050/CoreGenes1.0//) is a web-based program that determines a core set of genes based on comparison of four small genomes. The user supplies NCBI accession numbers for the genomes of interest. The program performs an iterative BLAST comparison to find orthologous genes by using one genome as a reference and another as a query. This pairwise comparison is performed for all four genomes. As a result, the common genes are compiled as a core set of genes from the genomes.

Lateral Gene Transfer

Lateral gene transfer (or *horizontal gene transfer*) is defined as the exchange of genetic materials between species in a way that is incongruent with commonly accepted vertical evolutionary pathway. Lateral gene transfer mainly occurs among prokaryotic organisms when foreign genes are acquired through mechanisms such as transformation (direct uptake of foreign DNA from environment), conjugation (gene uptake through mating behavior), and transduction (gene uptake mediated by infecting viruses). The transmission of genes between organisms can occur relatively recently or as a more ancient event.

If lateral transfer events occurred relatively recently, one would expect to discover traces of the transfer by detecting regions of genomic sequence with unusual properties compared to surrounding regions. 1 The unusual characteristics to be examined include nucleotide composition, codon usage, and amino acid composition. This can be considered a "within-genome" approach. 2 Another way to discern lateral gene transfer is through phylogenetic analysis (see Chapters 10 and 11), referred to as an "among-genome" approach, which can be used to discover both recent and ancient lateral gene transfer events. Abnormal groupings in phylogenetic trees are often interpreted as the possibility of lateral gene transfer events. Because phylogenetic analyses have been described in detail in previous chapters, the following introduces basic tools for identifying genomic regions that may be a result of lateral gene transfer events using the within-genome approach.

1) within genome

2) Among genome

Within-Genome Approach

This approach is to identify regions within a genome with unusual compositions. Single or oligonucleotide statistics, such as G–C composition, codon bias, and

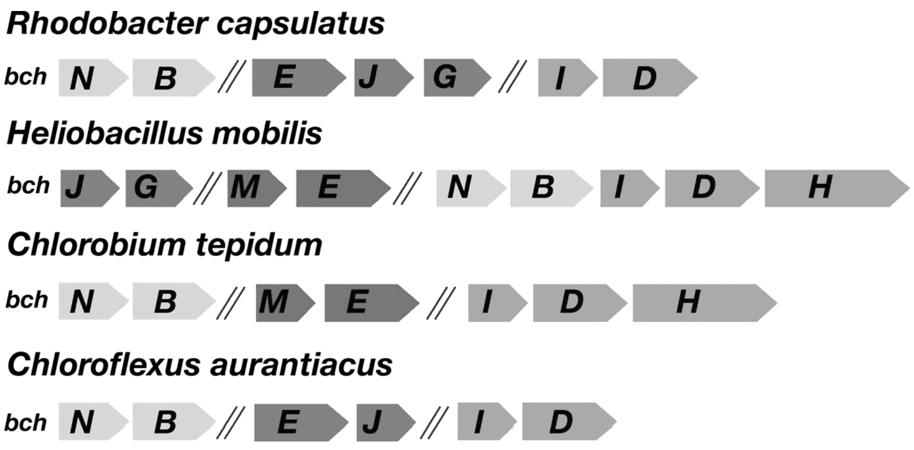

Figure 17.7: Schematic diagram showing a conserved linkage pattern of photosynthesis genes among four divergent photosynthetic bacterial groups. The synteny reveals potential physical interactions of encoded proteins, some of which have been experimentally verified. All the genes shown (*bch*) are involved in the pathway of bacteriochlorophyll biosynthesis. Intergenic regions of unspecified lengths are indicated by forward slashes (//). (*Source:* from Xiong et al., 2000; reproduced with permission from *Science*).

oligonucleotide frequencies are used. Unusual nucleotide statistics in certain genomic regions versus the rest of the genome may help to identify “foreign” genes in a genome. A commonly used parameter is GC skew ((G − C)/(G + C)), which is compositional bias for G in a DNA sequence and is a commonly used indicator for newly acquired genetic elements.

ACT (Artemis Comparison Tool; www.sanger.ac.uk/Software/ACT) is a pairwise genomic DNA sequence comparison program (written in Java and run on UNIX, Macintosh, and Windows) for detecting gene insertions and deletions among related genomes. The pairwise sequence alignment is conducted using BLAST. The display feature includes showing collinear as well as noncollinear (rearrangement) regions between two genomes. It also calculates GC biases to indicate nucleotide patterns. However, it is up to the genome annotators to determine whether the observations constitute evidence for lateral gene transfer, as this requires combining evidence from multiple approaches.

Swaap (http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm) is a Windows program that is able to distinguish coding versus noncoding regions and measure GC skews, oligonucleotide frequencies in a genomic sequence.

Gene Order Comparison

Another aspect of comparative genomics is the comparison of gene order. When the order of a number of linked genes is conserved between genomes, it is called *synteny*. Generally speaking, gene order is much less conserved compared with gene sequences. Gene order conservation is in fact rarely observed among divergent species. Therefore, comparison of syntenic relationships is normally carried out between relatively close lineages. However, if syntenic relationships for certain genes are indeed observed among divergent prokaryotes, they often provide important clues to functional relationships of the genes of interest. For example, genes involved in the same metabolic pathway tend to be clustered among phylogenetically diverse organisms. The preservation of the gene order is a result of the selective pressure to allow the genes to be coregulated and function as an operon. Furthermore, the synteny of genes from divergent groups often associates with physical interactions of the encoded gene products. The use of conserved gene neighbors as predictors of protein interactions is discussed in Chapter 18. An example of synteny of bacterial photosynthesis genes coupled with protein interactions is illustrated in Figure 17.7.

GeneOrder (http://pumpkins.ib3.gmu.edu:8080/geneorder/) is a web-based program that allows direct comparison of a pair of genomic sequences of less than 2 Mb. It displays a dot plot with diagonal lines denoting collinearity of genes and lines off the diagonal indicating inversions or rearrangements in the genomes.

SUMMARY

Genome mapping using relative positions of genetic markers without knowledge of sequence data is a low-resolution approach to describing genome structures. A genome can be described at the highest resolution by a complete genome sequence. Whole-genome sequencing can be carried out using full shotgun or hierarchical approaches. The former requires more extensive computational power in the assembly step, and the latter is inefficient because of the physical mapping process required. Among the genome sequence assembly programs, ARACHNE and EULER are the best performers. Genome annotation includes gene finding and assignment of function to these genes. Functional assignment depends on homology searching and literature information. GO projects aim to facilitate automated annotation by standardizing the descriptions used for gene functions. The exact number of genes in the human genome is unknown, but is likely to be in the same range as most other eukaryotes. The gene number, however, does not dictate complexities of a genome. One example is exhibited in protein expression in which a larger number of proteins are produced than genes available to code for them. This is the so-called genome economy. The main mechanisms responsible for genome economy are alternative splicing and exon shuffling. Genomes can be compared on the basis of their gene content and gene order. Many specialized genome comparison programs for cross-genome alignment have been developed. Among them, BLASTZ and LAGAN may be the best in terms of speed and accuracy. Gene order comparison across genomes often helps to discover potential operons and assign putative functions. Conserved gene order among prokaryotes is often indicative of protein physical interactions.

FURTHER READING

Bennetzen, J. 2002. Opening the door to comparative plant biology. *Science* 296:60–3.

Chain, P., Kurtz, S., Ohlebusch, E., and Slezak, T. 2003. An applications-focused review of comparative genomics tools: Capabilities, limitations and future challenges. *Brief. Bioinform.* 4:105–23.