11 GENE IDENTIFICATION AND PREDICTION

11.1 INTRODUCTION

The objective of gene prediction is to identify regions of genomic DNA that encode proteins. It is based on the statistical analysis of sequence bias in genome coding regions. Gene prediction programs are used to sift through new sequences and then annotate the sequence database entry with this information.

Many of the gene-finding programs are similarity-based, i.e. sequence similarity is done at either the protein or nucleotide level to known proteins or Expressed Sequence Tags (ESTs) to identify the genome regions likely to be protein coding. If a region has similarity to a known EST or protein, it provides support that the region likely encodes an exon. This information can be used to refine regions that are likely exons for closer examination in building a gene model.

Many programs have some commonness and possess the ability to differentiate between gene sequences characteristic of exons, introns, splicing sites, and other regulatory sites in expressed genes from other non-gene sequences that lack these patterns. However, a program trained on one organism may not work as efficiently for the other. cDNA sequencing can be used to confirm gene identification. If EST sequences are available, covering a large amount of genome, they can also be used for confirmation of predicted gene sequences.

11.2 BASIS OF GENE PREDICTION

A gene is a segment of DNA that is expressed to yield a functional product like an RNA or a polypeptide. The summary of the important features of the gene structure, that is important for gene recognition, is given in Table 11.1 and shown in Figure 11.1.

Translation is the process of building proteins according to the template RNA. Translation begins from the 5′ ends of mRNA and the proteins are made in an N to C direction (Amino H_2N to carboxylic terminal) (see Figure 11.4). Ribosomes bind to a region of the mRNA that lies just upstream of the initiation AUG (methionine) codon and are thus

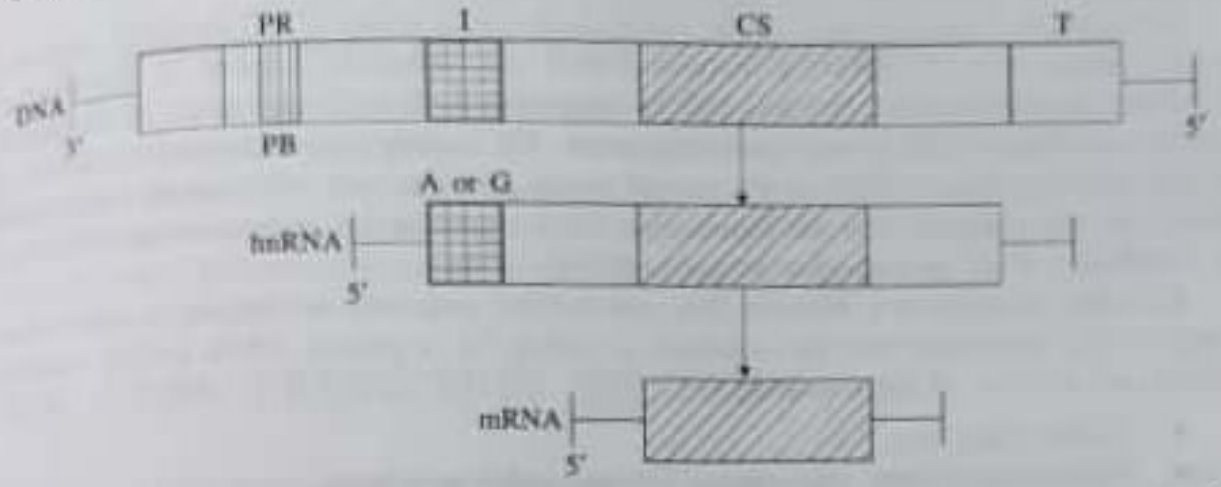

FIGURE 11.4 Transcription begins at the specific promoter site on the DNA template—Prinbow box. PB—Prinbow, Pr—promoter, I—initiator, CS—coding sequence, T—terminator, hnRNA—heterogeneous RNA and mRNA—messenger RNA.

positioned ready to begin translation. This sequence (which is analogous to the promoter in DNA) is called a *Shine-Dalgarno sequence*, or *Ribosome Binding Site* (RBS). The typical sequence is AGGAGGA and it is complementary to the 3′ end of the 16S RNA, UCCUCCUA-OH3′. In polycistronic messengers this sequence often occurs just upstream of each initiation codon. There is no similar sequence found in eukaryotes where the 5′ end of the mRNA is modified by a cap. Ribosomes are believed to bind and scan until they find the first AUG. The sequences surrounding the AUG help in recognition by the ribosome. ACC AUG G (Kozak consensus sequence) is optimal.

Codon Bias

Genes are usually predicted as the regions of a genome that encode proteins and have biased sequence composition compared to non-coding regions. There is a bias in sequence due to the properties of coding regions or on the tendencies for bases to occur at particular positions within some codons. The unequal use of possible codons in the coding regions is a universal characteristic of organisms and is known as *codon bias*. This bias is in terms of unequal usage of amino acids in proteins and the uneven usage of synonymous codons. A comparison of codon frequency in a region of the genome of a species with the typical codon frequency in the same species may give an idea of the regions coding for a particular protein. Thus, regions in which codons are used with frequencies similar to the typical species codon frequencies observed in known protein coding regions can be used to tentatively identify the region as a possible coding region.

pattern matching software may allow several options for making the pattern search more specific or general, as the case may be. These options include:

- Presence of ambiguous symbols in the specified patterns
- Variable spacing between matched positions
- Choice of alternative matches to particular positions
- Matches that include gaps in the pattern or target sequence

Sequences of related protein families sometimes also have multiple consensus patterns that increase the prediction of function (for example, using the BLOCKS database). Pattern matching in sequences is the basis for performing rapid searches through a sequence database for the closest matches to a given sequence by the BLAST and FASTA programs. FASTA has been successful in locating unidentified DNA binding sites for *E. coli* LexA protein.

11.4 GENE PREDICTION METHODS

There are several methods of gene prediction, differing in the approach, algorithm, and the efficacy of prediction methods. Important methods of gene prediction are as follows:

- Laboratory-based approaches
- Feature-based approaches
- Homology-based approaches
- Statistical and HMM-based approaches

Laboratory-based Approaches

This is the traditional method to find a gene, which includes experimental procedures for locating genes in a sample of DNA. These can now be discussed.

Identification using blotting methods

Blotting is a technique used for detection of nucleic acids and proteins. The technique employs the transfer of biomolecules on to a membrane support that usually accomplishes blotting.

The entire procedure involves the following steps:

1. Preparing a cell-free extract containing the biomolecule(s) of interest
2. Resolving the mixture by gel electrophoresis
3. Transferring the resolved mixture onto a membrane support such as nitrocellulose paper
4. Incubating the paper with a detection system that specifically hybridizes to the molecule of interest.

When DNA is blotted, it is termed as Southern blotting, whereas Northern and Western blotting techniques are coined for RNA and protein, respectively.

the cell or tissue. Since cellular structure is maintained throughout the procedure, ISH provides information about the location of mRNA within the tissue sample.

The procedure begins by fixing samples in neutral-buffered formalin, and embedding the tissue in paraffin. The samples are then sliced into thin sections and mounted onto microscope slides. Alternatively, tissue can be sectioned frozen and post-fixed in paraformaldehyde. After a series of washings to de-wax and rehydrate the sections, a proteinase K digestion is performed to increase probe accessibility. A labeled probe is then hybridized to the sample sections. (Proteinase K digestion is a crucial step for successful ISH. Insufficient digestion will result in a diminished hybridization signal. On the other hand, if the sample is over digested, tissue morphology will be poor and would make localization of the hybridization signal difficult. The concentration of Proteinase K needed is dependent on the tissue type, length of fixation, and size of tissue core.) Radio-labeled probes are visualized with liquid film dried onto the slides, while non-isotopically labeled probes are conveniently detected with colorimetric or fluorescent reagents.

The major drawback to ISH is the procedure itself. Standard protocols are cumbersome, time-consuming and laborious and require specialized equipment for preparing samples and visualizing results of the experiment. Additionally, quantitation of gene expression is not as straightforward as with the other techniques.

Feature-based Approach

Feature-based approaches are based on pattern recognition, treating DNA fragments as sequences (see Table 11.1).

Gene finding by ORF prediction

ORFs without stop codons are strongly suggestive of genes. ORF has the presence of a long series of codons in a DNA sequence without the series being interrupted by a termination codon. An ORF signal is enhanced even further by the presence of sequence patterns for starting and stopping transcription before and after the ORF. Dynamic programming can be used to identify the highest scoring regions. The best gene recognition systems tend to be species-specific, trained on examples of known genes in the given organism. The initiation site of box is always an ATG codon and it is always about 30 base pairs downstream from a TAATAA sequence.

GRAIL (Gene Recognition and Analysis Internet Link) (**http://compbio.ornl.gov/Grail-1.3/**) is perhaps the most widely used ORF identification tool. (It was also one of the first to be made available). It provides analysis of protein coding potential of a DNA sequence. GRAIL uses variable-length windows tailored to each potential exon candidate defined as an open reading frame bounded by a pair of start/donor, acceptor/donor or acceptor/stop sites. This scheme facilitates the use of more genomic context information (splice junctions, translation starts, non-coding scores of 60-base regions on either side of a putative exon) in the exon recognition process. GRAIL finds about 91% of all coding regions with an apparent false positive rate of 8.6%.

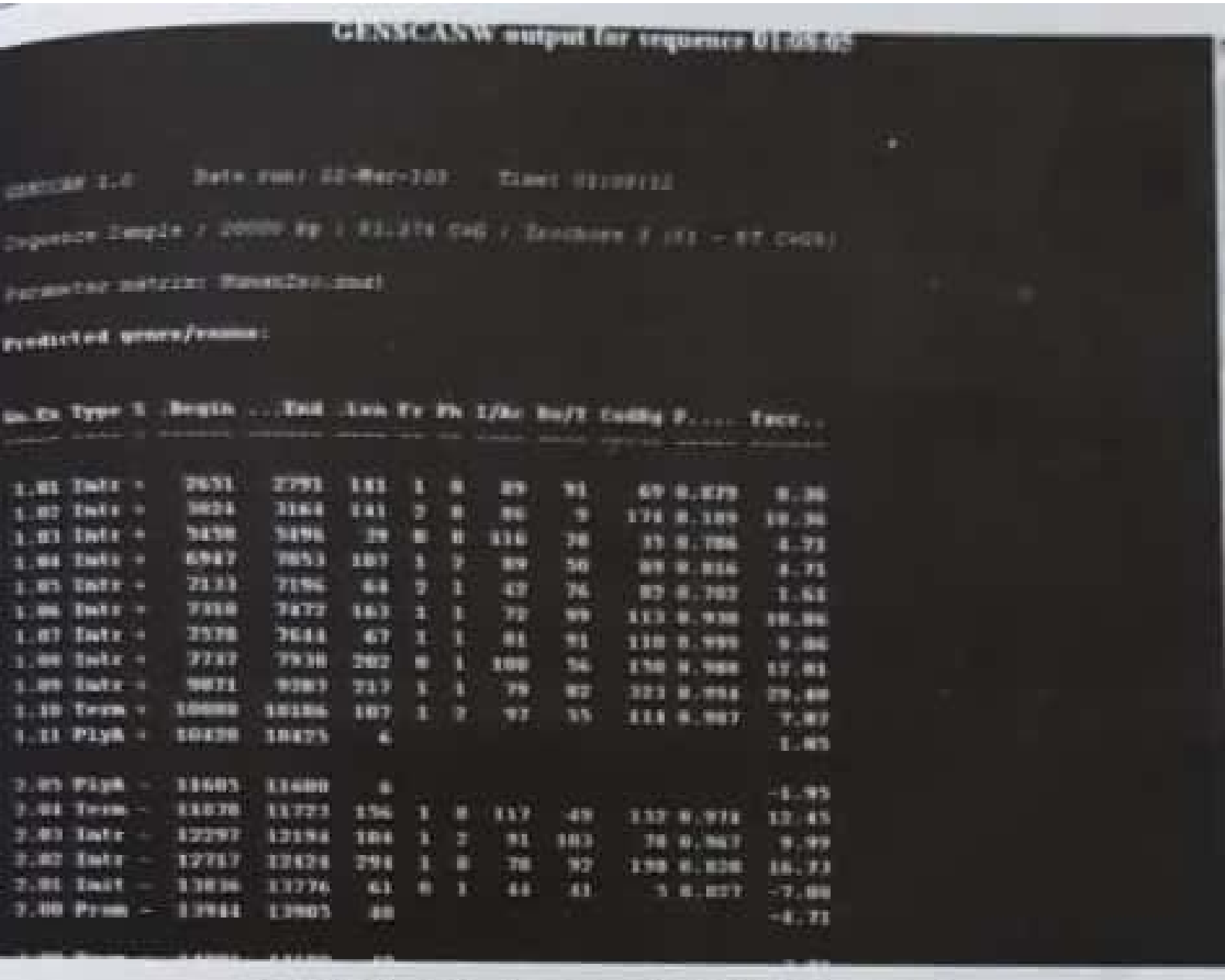

FIGURE 11.8 Sample of GENSCAN output.

GeneParser

GeneParser (**http://beagle.colorado.edu/~eesnyder/GeneParser.html**) predicts the most likely combination of exons and introns in a genomic sequence by a DP approach. It uses a likelihood score for each sequence position being in an intron and exon. The intron and exon positions are then aligned with the constraint that they must alternate within a gene structure. In this manner, a combination of the most likely intron and exon regions that comprise a gene structure are found. GeneParser uses a scheme for adjusting the weights used for several types of sequence patterns that make up the intron and exons.

GeneId

GeneID (**http://www1.imim.es/software/geneid/asd.html**) is a program to predict genes in unknown genomic sequences designed with a hierarchical structure. In the first step, splice sites, start and stop codons are predicted and scored along the sequence using Position Weight Arrays (PWAs). In the second step, exons are built from the sites. Exons are scored as the sum of the scores of the defining sites, plus the log-likelihood ratio of a Markov Model for coding DNA. Finally, from the set of predicted exons, the gene structure is assembled, maximizing the sum of the scores of the assembled exons.

11.5 OTHER GENE PREDICTION TOOLS

Some of the other tools available for searching for various patterns are given below:

Poly-A site prediction

HCpolyA (**http://l25.itba.mi.cnr.it/~webgene/wwwHC_polya.html**) is used for Poly-A Site Prediction using the Hamming Clustering Method on the WebGene Server.

TATA signaling, promoter & trans-factor bind site prediction

HCtata (**http://l25.itba.mi.cnr.it/~webgene/wwwHC_tata.html**) is a similar tool using Hamming-Clustering Method for TATA Signal Prediction in Eukaryotic Genes. This is also from WebGene.

MatInspector (**http://www.genomatix.de/cgi-bin/matinspector/matinspector.pl**)—Search Sequence for Transcription Factor.

Signal Scan (**http://bimas.dcrt.nih.gov/molbio/signal/**) is used to find and list homologies of published signal sequences with the input DNA sequence.

Tfsitescan (**http://www.ifti.org/cgi-bin/ifti/Tfsitescan.pl**)—This tool is intended for promoter sequence analysis and works best with sequences of ~500 nt.

Exon (ORF) prediction

FramePlot (**http://watson.nih.go.jp/~jun/cgi-bin/frameplot-3.0b.pl**) Protein Coding Region Prediction in Bacterial DNA—NIH-NET.

ORF Finder (**http://www.ncbi.nlm.nih.gov/gorf/gorf.html**) graphical analysis tool that finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes.

Genefinder is a tool for exon prediction. It has the following facilities: HEXON—internal exons, FEXH—all exons and FGENEH—gene structure.

The Exon Prediction Program (Perceval) (**http://compbio.ornl.gov/grailexp/gxpfaq2.html**) stands for Protein-coding Exon, Repetitive, and CpG-Island EVALuator. Perceval reads in a DNA sequence and produces a list of possible Grail Exon Candidates. It also filters these candidates against a repetitive element database. It also locates repetitive elements and CpG islands.

Splice site prediction

Genefinder is a tool for exon prediction. It has the following facilities:

HSPL—splice sites

Repetitive DNA & CpG isles analyses

RepeatMasker2 (**http://ftp.genome.washington.edu/cgi-bin/RepeatMasker**) is used for analysis of Repetitive Elements in DNA Sequences. RepeatMasker screens DNA sequences in fasta format against a library of repetitive elements and returns a masked query sequence ready for database searches as well as a table annotating the masked regions.

tRNA gene prediction

tRNAscan is used for genomic tRNA Identification (**http://www.genetics.wustl.edu/eddy/tRNAscan-SE/**).

A comparison of the various tools in the terms of their performance is given in Table 11.4.

TABLE 11.4 Comparison of Various Gene Prediction Tools

Tool	*Prediction Type*	*Sensitivity (%) Nucl.*	*Specificity (%) Nucl.*	*Sensitivity (%) Exact Exon*	*Specificity (%) Exact Exons*	*Mixed Exon*
FGENES	Gene structure	83	93	73	78	15
GeneID	Gene structure	69	77	42	46	28
Gene Parser	Gene structure	66	79	35	40	29
GENSCAN	Gene structure	93	93	78	81	9
GRAIL II	Gene structure	83	87	–	52	25
MZEF	Internal exons	87	95	78	86	14

Searching for protein binding sites in DNA sequences

DNA and protein sequences, which have a related function, may share consensus patterns that can be found by sequence analysis methods. An example is a set of DNA sequences that contains signals for transcriptional promoters. Sequences of proteins in families may also have conserved amino acid patterns or motifs. These patterns serve as a signature of function and may be used to identify other sequences that may have the same function.

SIGNAL SCAN (**http://www.cbs.umn.edu/software/sigscan.html**) is a program that utilizes a database of transcription factor sequences to find potential transcription factor binding sites in DNA sequences.

If the members of a set of sequences are similar to each other, the simplest method of finding consensus patterns is to align the sequences by Multiple Sequence Alignment and make a profile and search for patterns with a statistical method such as the expectation maximization method.

DNA binding sites for proteins may be composed of several conserved patterns separated by variable spaces between the patterns. Expectation Maximization Algorithm has been devised for finding such regions in unaligned sequence fragments. It goes through the following four steps:

1. The best scoring comparison matrix is obtained.
2. This matrix is then used to find the approximate locations of the binding sites in the original sequences.
3. The predicted binding sites are then used to make a new matrix.
4. This matrix is again used to define even better the locations of the binding sites in the sequences.

This process is repeated until the method converges on a single set of patterns in the sequences.

GenLang

GenLang (**http://www.cbil.upenn.edu/genlang/genlang_home.html**) is a syntactic pattern recognition system, which uses the tools and techniques of computational linguistics to find genes and other higher-order features in biological sequence data. Patterns are specified by means of rule sets called grammars, and a general purpose parser, implemented in the logic programming language Prolog, then performs the search.

BCM GeneFinder

GeneFinder (originally from Baylor Institute and now with Sanger Institute—not freely available now) offers some unusual custom algorithms. The algorithm first predicts all possible potential internal exons, and potential 5′ and 3′-exon for each internal by linear discriminant functions combining characteristics describing various contextual features of these exons. Then, by the method of dynamic programming, it searches for optimal combination of these exons and construct gene model.

splice sites and start/stop codons. If some features of a sequence are known, such as hits to ESTs, proteins, or repeat elements, these regions can be locked as coding or non-coding and then the program will find the best gene structure under these constraints. Apart from reporting the best prediction, HMMgene can also report the N best gene predictions for a sequence. This is useful if there are several equally likely gene structures and may even indicate alternative splicing.

Glimmer (Gene Locator and Interpolated Markov Modeler) is a system for finding genes in microbial DNA. It is especially useful for the genomes of bacteria and archaea. It is available at **http://www.tigr.org/software/glimmer/** and uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from non-coding DNA. The IMM approach uses a combination of Markov models from 1st through 8th-order, weighting each model according to its predictive power.

The Veil system (**http://www.tigr.org/~salzberg/veil.html**) stands for "the Vanet Exon-Intron Locator" and uses a custom-designed HMM to find genes in eukaryotic DNA. Veil is not actively maintained now.

GENSCAN (**http://genes.mit.edu/GENSCAN.html**) is a program to predict complete gene structures. GENSCAN can be used to identify introns, exons, promoter sites, polyA signals, etc. GENSCAN uses a general three-periodic fifth-order Markov model of coding regions. The algorithm can assign a probability as to the chance that a given sequence represents an exon or an intron.

GENSCAN uses three types of signal models to model different functional units. It uses the weight matrix model (WMM) for modeling polyadenylation signals, translation initiation signal, translation termination signal and promoters. A modified version of the weighted array model (WAM) is used for acceptor splice sites.

For training GENSCAN, the training set is divided into four categories depending on the C + G content of the sequence. The categories are:

1. (< 43% C + G)
2. (43 – 51% C + G)
3. (51 – 57% C + G)
4. (> 57% C + G)

For each of these categories, separate initial state probabilities are computed by estimating the relative frequencies of various functional units in these categories.

GENSCAN was given a sequence of 20 kbp. The outputs are given in Figure 11.8 and Figure 11.9.

Genie (**http://www.fruitfly.org/seq_tools/genie.html**) is another tool that uses generalized HMMs and neural networks.

Searching for CG islands

As discussed earlier in the Chapter 4, CG islands are regions in DNA sequences where the dimer CG repeatedly occurs. An HMM can be used to train a program on appropriately identified examples of CG islands and non-islands and learn to recognize them. The HMM can be used to find if given a short sequence, the sequence come from a CpG island or not. The HMM can also be trained to find the CpG islands in a long sequence.

Homology-based approaches

Eukaryotic genes are more difficult to identify than those of the prokaryotes. Searching for known homologues is the most widely used method for identifying genes. Homology search depends only on evolutionary relatedness, and so are widely applicable. A major advantage of finding homologous product is that some of the information about the gene may be already known.

Search for genes can include matches to one of the following:

- Known proteins
- Protein motifs (e.g. zinc finger, ATP and GTP-binding motifs, etc.)
- ESTs (Expressed Sequence Tags) and ACRs (Ancient Conserved Regions)

Homology-based gene prediction systems can find similarities to previously identified coding regions. Alternatively, a different homology-based approach is to identify totally unknown genes to compare two whole genomes and look for conserved regions on the theory that sequence is only conserved if it is important.

Procrustes (**http://hto-13.usc.edu/software/procrustes/**) is a homology-based program (currently not fully functional), which accepts as input one genomic DNA sequence, and one or several protein sequences. The proteins (targets) are assumed to be similar to the protein encoded in the genomic fragment. Procrustes finds the chain of exons with the best fit to the target proteins; if several targets are specified, it makes one gene prediction per target. Procrustes also outputs the amino acid sequence of the predicted protein and the alignment between the predicted and target proteins.

Finding coding regions can also be done by similarity searching using TBLASTX for finding exons. The approach to this problem is to translate the sequence in all six reading frames (3 forward and three reverse) and do a similarity search against the protein databanks. TBLASTX translates a DNA query sequence and performs a similarity search against protein databanks. If a protein sequence matches, get its DNA sequence and align it with your unknown sequence. The start and stop codons would get identified. If the query sequence were genomic, then the introns would also be identified.

Statistical and HMM approaches

In Chapter 10, you have learned that one of the important applications of HMMs is in gene identification. One of the programs discussed there was GeneMark (**http://www.ebi.ac.uk/genemark/**) that used HMMs for gene identification. There are several others and some can now be discussed here.

HMMs for gene prediction can be developed by using simplified gene grammar rules like start-codon, end-codon, length is divisible by 3 and no stop-codons in the reading frame. The language may also consider dinucleotide preferences (e.g.) AG is more common than AC and that nucleotides are not necessarily independent.

HMMgene (**http://www.cbs.dtu.dk/services/HMMgene/**) is a program for prediction of genes in anonymous DNA. The program predicts whole genes, and so the predicted exons always splice correctly. It can predict several whole or partial genes in one sequence, and can be used on whole cosmids or even longer sequences. HMMgene can also be used to predict

Grail II helps in analyzing protein-coding regions, poly (A) sites, and promoters, enables to construct gene models, predicts encoded protein sequences, and provides database searching capabilities. A list of most likely exon candidates is first established, and these are evaluated further using a neural network approach. The algorithm makes its final prediction by selecting the best candidates. A DP approach is then used to define the most probable gene models.

FindPatterns (**http://www.accelrys.com/products/gcg_wisconsin_package/programs_list.html#FindPatterns**) can be used for scanning for ORF patterns.

Frames (**http://www.accelrys.com/products/gcg_wisconsin_package/**) can show ORFs for the six translation frames of a DNA sequence. Frames can superimpose the pattern of rare codon choices if you provide it with a codon frequency table.

MacVector 6.5 (**http://www.sxst.it/sxm__mcv.htm**) does the ORF detection based on Fickett's statistical method, or on the designation of sequence ends as start and stop codons. ORFs can be found in 3 or 6 frames.

Sequencher (**http://www.genecodes.com/index.html**) can also be used for ORF analysis. Sequencher can also be used for contig assembly, restriction enzyme mapping, heterozygote detection, cDNA to Genomic DNA large gap alignment, motif, and SNP analysis.

ORFs are easy to find with automated tools, however there are two major problems faced in their identification:

Small proteins. The issue is to decide the "cutoff" to be used for a minimum size protein. A cutoff of 100 amino acids is often used. However, in so doing, some true small proteins containing fewer than 100 amino acids are not annotated and some ORFs containing more than 100 amino acids are annotated even though they do not encode a protein.

Small exons. Exons smaller than about 30 nucleotides cannot be reliably predicted by normal computational methods. Missing a small exon can result in prediction of a protein sequence that has an internal "frame shift", (i.e.) the protein coding frame has shifted. Such a shift changes all the amino acids after the frame shift position, resulting in major errors in prediction of the protein sequence.

There are various tests to verify that a predicted ORF is in fact likely to encode a protein. Some of these are as follows:

1. The method is based on an unusual type of sequence variation that is found in ORFs—every third base tends to be the same one much more often than by chance alone. This property is due to non-random use of codons in ORFs and is true in any ORF, independent of the species. The program TestCode (**http://www.accelrys.com/products/gcg_wisconsin_package/**) provides a plot of the non-randomness of every third base in the sequence.
2. This method is based on the analysis to determine whether the codons in the ORF correspond to those used in other genes of the same organism. For this specific information on codon use for an organism is necessary, averaged over all genes.
3. The ORF may be translated into an amino acid sequence and the resulting sequence then compared to the databases of existing sequences. If one or more sequences of significant similarity are found, there will be much more confidence in the predicted ORFs.

tionary conservation of the test gene and can be predictive of function. Zoo blots are to determine similarity in DNA sequences, and hence possible homology, to the new sequence in one or more of these organisms.

nuclease mapping

g S1 nuclease mapping and Primer Extension, you can identify the 5′-end of the gene Figure 11.5).

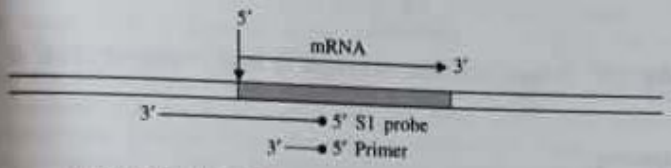

FIGURE 11.5 Identification of 5′-end of the gene.

S1 nuclease is an endonuclease that degrades single-stranded RNA or DNA to nonucleotides. S1 nuclease is useful for the removal of unpaired regions following bridization and is used in S1 nuclease mapping—a technique for the removal of single-anded regions of DNA and RNA from double-stranded hybrids of DNA and RNA. In S1 clease mapping, a DNA probe labeled at its 5′-end overlaps the gene 5′-end or the 5′-end of exon is hybridized to the gene DNA. The resulting 5′-labeled DNA probe is then "sized" ing a Southern blot. Its size pinpoints the 5′-end of the gene or exon. (Figure 11.6).

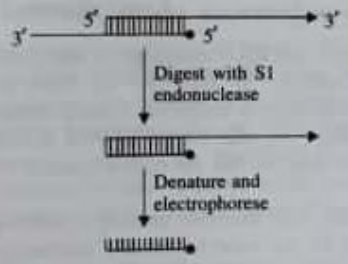

FIGURE 11.6 S1 nuclease mapping to identify the 5′-end of the gene.

Primer extension

Primer extension is the technique used to map the 5′ ends of DNA or RNA fragments. It involves annealing a specific oligonucleotide primer to a position downstream of 5′ end. The primer is labeled, usually at its 5′ end, with ^{32}P. This is extended with reverse transcriptase, which can copy either RNA or a DNA template, making a fragment that ends at the 5′ end of the template molecule. DNA polymerase can also be used with DNA templates. (Figure 11.7)

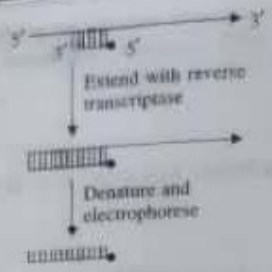

FIGURE 11.7 Primer extension technique to map 5′ ends of DNA or RNA.

Exon trapping

Exon trapping or exon amplification is a method to find expressed DNA sequences in a genome sequence and is based on selection for functional splice sites in genomic DNA. The advantages of exon trapping are that it does not require any prior knowledge about tissue-specific gene expression and can easily be performed on complex genomes. It can identify constitutive exons as well as alternative exons but cannot be used to identify intronless genes. In exon mapping, the genomic sequence is cloned into an intron (flanked by two exons) using a specialized exon trapping vector. This construct is expressed through a strong promoter. If the genomic fragment contains an exon, it will be spliced into the resulting mRNA, changing its size and allowing its detection.

Reverse transcriptase-polymerase chain reaction (RT-PCR)

RT-PCR can be used to detect the RNA transcript of any gene. This is irrespective of the quantity of the specific mRNA. In RT-PCR, an RNA template is copied into a complementary DNA (cDNA) using a retroviral reverse transcriptase. The cDNA is then amplified exponentially by PCR. As with NPAs, RT-PCR is somewhat tolerant of degraded RNA. As long as the RNA is intact within the region spanned by the primers, the target will be amplified.

Although RT-PCR is the most sensitive method of mRNA detection available, it also has some drawbacks. It can be the most technically challenging method of detection and quantitation, often requiring substantial pre-experimental planning and design. Also, because of its extreme sensitivity, even minute amounts of contamination by genomic DNA or previously amplified PCR products can lead to aberrant results, so steps must be taken to avoid any contamination.

In situ hybridization (ISH)

ISH is a powerful and versatile tool for the localization of specific mRNAs in cells or tissues. Unlike Northern blotting and nuclease mapping assays, ISH does not require the isolation or electrophoretic separation of RNA. Hybridization of the probe takes place within

Southern blotting. Suppose, one of the cDNA clones you isolate and sequence represents a new gene. If you are interested in studying this gene further, you might want to determine the structure of the gene by identifying introns, exons, and regulatory elements.

An essential step for such analysis is Southern blotting, which is a very popular technique used for a variety of purposes. It is used to determine the size and arrangement of the genomic copy of a gene, to determine the number of genes related to a clone of interest and to investigate the evolutionary conservation of a gene. The Southern blotting technique has been used for understanding a variety of biological processes such as RNA splicing and genomic rearrangements to form antibodies and T cell receptors. This technique has also played a major role in the identification of numerous rearranged genes that are associated with a variety of human genetic disorders and cancers. With the introduction of highly resolving gel systems, it is now possible to use Southern blotting to detect gene mutations involving single base-pair changes. This has led to the early diagnosis and prevention of potentially harmful diseases.

Northern blotting. Northern blotting is a technique used to examine the size, and temporal and spatial expression pattern of specific RNAs. Usually, total cellular RNA, or poly(A) RNA is isolated and separated by size on an agarose gel. The RNA molecules in the gel are transferred to nitrocellulose paper or nylon as described above and detected using an appropriate DNA or RNA probe.

Although Southern and Northern blotting techniques exhibit a number of similarities, there are several important differences also. The major difference is in the extreme care required to isolate non-degraded RNA. Full-length mRNA isolation is an important goal in generating high quality cDNA libraries. The difficulty in RNA isolation is that most ribonucleases are very stable and active and do not require activators or cofactors to function. As a precaution, the first step in all RNA purification procedures is to lyse the cell in a solution that denatures, thus inactivating ribonucleases. Another difference in the two procedures is in the type of gel used to resolve RNA. Unlike DNA that is only found as a double-stranded version, RNA migrates as a function of hybrid length, RNA can engage in non-uniform amounts of intra-molecular base pairing. Therefore, RNA must undergo electrophoresis under denatured conditions if it is to migrate as a function of nucleotide length.

Northern blotting is useful because the size of a specific mRNA can be compared with the size of cloned DNAs, revealing whether the cloned cDNA is full-length. This technique can indicate which tissues express a particular gene or the factors that regulate its expression. As an example, if you have isolated a cDNA, which is suspected to be induced by a growth factor, you could first try to experiment to stimulate cells with the growth factor and isolate total RNA at intervals following stimulation. The RNA isolated at each point of time would be analyzed by Northern blotting using the cloned cDNA as a probe. If the results indicate that the abundance of the mRNA in question is low in untreated cells but significantly increases following stimulation, a good evidence that it provides the expression of the desired gene is indeed regulated by growth factors.

Zoo blots. Zoo blots are variants of Southern blots in which a panel of genomic DNA from different species is probed with a test gene or cDNA. The patterns obtained reveal the

If you consider the G + C composition of a region, you would note that there are many organisms that have skewed composition of G + C bases that lead to a biased codon composition. It is hypothesized that in many such cases, non-coding regions tend to reflect the overall G + C composition of the organism. The coding regions due to natural selection and restriction of nucleotides for coding lead to codon bias.

Streptomyces coelicolor is a model representative of a group of soil-dwelling organisms with a complex lifecycle involving mycelia growth and formation of spore. In *Streptomyces coelicolor*, the chromosome is 8,667,507 bp long with a G + C content of 72.1%, and is predicted to contain 7825 protein encoding genes. The coding regions have 70% G + C at the first codon position, ~50% at the second codon position, and >95% at the third codon position. If you compare with the non-coding DNA sequence, all three positions will show the overall G + C of the organism which might be ~50%.

A coding statistics is a function that uses a DNA sequence to compute a real number related to the likelihood that the sequence is coding for a protein. Most coding statistics measure one or more of the following:

- Codon usage bias
- Base compositional bias (among various codon positions)
- Base occurrence periodicity

These biases can be used to make predictions of the likely coding potential for a region of DNA. Introns are identified as large gaps in the alignment, typically (but not necessarily) flanked by the consensus GT and AG dinucleotides at the donor and acceptor sites, respectively.

11.3 PATTERN RECOGNITION

Pattern recognition is the method of scanning a nucleic acid or a protein sequence for matches to short sequence patterns. These short patterns can be important indicators of some biological function. The presence of the matching pattern in the target nucleic acid or protein sequence is a signal of the same function for the target gene or protein sequence.

Patterns most often examined in DNA sequences are given in Table 11.3.

TABLE 11.3 Summary of Gene Features and Corresponding DNA Characteristics

Gene Feature	*DNA Characteristic*
Coding sequences (CDS)	Open reading frames (ORFs); GC-rich; CpG-content
Translational start and stop sites	Codons: Start (ATG) and Stop (TAA, TAG, TGA)
Splice sites (Exon/Intron borders)	Consensus sequences
Promoter regions	TATA, Shine-Dalgarno, Pribnow, Kozak Consensus, CpG-content
PolyA-signals	Consensus sequences (characteristic nucleotide combinations at about 10–20 nucleotides upstream of the insertion site for the polyA-tail)

Some times, exons encode functionally distinct protein domains and recombination between introns of different genes results in novel genes. Some genes are chimeras of exons derived from several other genes, providing direct evidence that new genes can be formed by recombination between intron sequences.

As discussed earlier, the process of copying the portion of the DNA containing a gene into RNA is transcription (Figure 11.2). There are signals in the DNA sequence, which trigger the transcription through enzyme RNA polymerase. RNA pol binds to specific patterns. $TTGACAN_{17}TATAAT$, a consensus sequence. (If we compare all very strong promoters we get

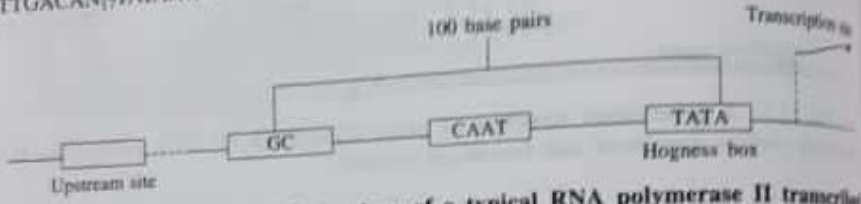

FIGURE 11.2 A generalized structure of a typical RNA polymerase II transcribed genes showing various structural and functional domains. The upstream regulatory elements include enhancers and silencers.

an average sequence for strong promoters, this is called a *consensus sequence*). This sequence is positioned 10 bp upstream (i.e. at –10) of the first nucleotide (labelled +1) of each gene. There are other binding sites upstream from the gene called promoters that signal when to express or inhibit the gene from expressing as RNA. Transcription stops when it encounters the signal to terminate—sometimes, a GC rich region which can form a hairpin loop, followed by a run of As (or Ts in the template strand). This sequence is referred to as the transcriptional terminator. The GC hairpin disrupts the binding of the mRNA to the DNA template.

Some genes have weak and others strong promoters. Strong promoters have a sequence close to the ideal consensus $TTGACAN_{17}TATAAT$. RNA pol does not overlook these promoters under normal circumstances and these genes may be transcribed at a frequency every 2 seconds. Other promoters can have slightly different sequences. For example, there may be changes in the TTGACA (–35 box) or in the TATAAT (–10 or Pribnow box) or in the spacing between the two motifs (Figure 11.3). Thus strongly and weakly expressed genes may simply have different promoters—some proteins are needed at 10,000 molecules/cell others at 10 copies/cell.

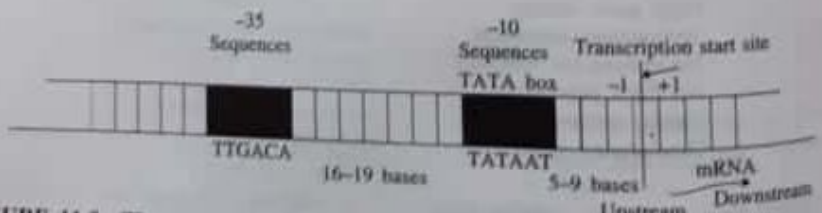

FIGURE 11.3 The prokaryotic promoter showing –10 sequence (Pribnow box) and –35 sequence separated by a distance of 16 and 18 bp. First nucleotide of the DNA template is transcribed into RNA at the transcriptional start site.

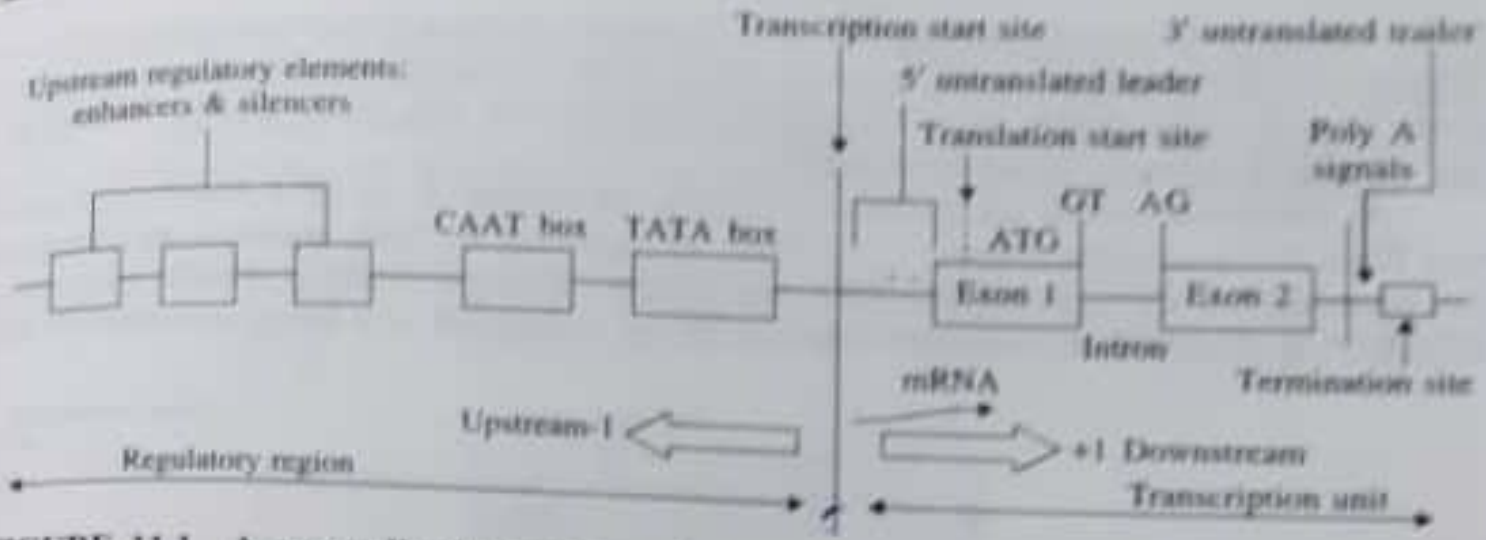

FIGURE 11.1 A generalized structure of genes transcribed by RNA polymerase II, displaying various structural and functional domains.

TABLE 11.1 Summary of Gene Structure

Upstream	*Promoter*	*First Exon*	*Intron(s)*	*Exon(s)*	*Intron(s)*	*Last Exon*	*Downstream*
Intergenic Region	For example: TATA box with consensus sequence TATA (A/T)A(A/T) and Inr with consensus sequence YYAN(T/A)YY	Transcriptional Start, 5'-UTR, Translational Start, Coding Sequences (CDS)/Open Reading Frames (ORF) and Enhancer Sites	Frequent Stop Codons	CDS/ORF and Enhancer Sites	Frequent Stop Codons	CDS/ORF and Enhancer Sites, Translational Stop, 3'-UTR, PolyA-insertion Site, Transcriptional Stop	Intergenic Region

Most prokaryotic genes