Chapter 5. Genomics, Proteomics and Bioinformatics

5.1 Genomics:

Genomics is a relatively new discipline. Although, the DNA was first isolated as early as 1869, it took more than one Century for the first genomes to be sequenced. The term genomics was introduced recently by Thomas Roderick in 1986. Genomics describe the detailed study of genome; it is structural organisation and function using various modern methods including computational biology. It involves the genome sequencing and computer aided analysis to understand its structural organisation and functions, genome mapping and related studies. The term genome represents the complete genetic material including both nuclear and cytoplasmic genes present in a cell. The Human Genome Project (HGP), sponsored in the United States by the Department of Energy and the National Institutes of Health, has created the field of genomics understanding genetic material on a large scale. The field of molecular life science is changing rapidly, because of the genomic revolution. Revolutionary improvements in the DNA sequencing techniques have given rise to a large amount of DNA sequences, which is difficult to manage, particularly for future references and analysis. Technological developments in computer and information technology have helped a lot in managing the huge data of DNA sequences in the form of computerised databases and it is access through internet.

5.1.1 Concept of genomics:

Thormas Roderich introduced the term genomics in 1986. It is scientific method of mapping, sequencing and analysing and making the use of genetic information for further use in multifarious area. Genomics can be defined as the study of molecular organisation of genomes, their information contents and the gene products they encode.

"Genomics is the study of structure and functions of a genome of an organism. It concerns with the sequencing and analysis of an organism's genome. The genome is nothing but the total DNA content that present within one cell of an organism".

5.1.2 Types of genomics:

In the last few years, some interesting findings have been recorded and several new branches have emerged. Consequently, the area of genomics has quietly widened. However, the genomics is broadly categorised into three types namely, structural genomics, functional genomics and comparative genomics.

1) Structural Genomics: The process of finding out the sequences of genome is called as structural genomics. The structural genomics deals with DNA sequencing, sequence assembly, sequence organisation and management. Structural genomics attempts to determine the structure of every protein encoded by the genome, rather than focusing on one particular protein. Basically, it is the starting stage of genome analysis i.e. construction of genetic map or sequence maps of high resolution of the organism. The complete DNA sequence of an organism is its ultimate physical map. Due to rapid advancement in DNA technology and completion of several genome sequencing projects for the last few years, the concept of structural genomics has come to a stage of transition. Now structural genomics also includes systematic and determination of 3D structure of proteins found in living cells, because proteins in every group of individuals vary and so there would also be variations in genome sequences.

2) Functional Genomics: To study and understand the function of gene is the basis of functional genomics. Based on the structural genomics the reconstruction of genome sequences is useful to find out the function that the genes do. It gives an idea of function of all gene sequence and their expression in organism. The different tools useful for structural genomics are bioinformatics sequences, DNA chips, 2D gels etc. This information lends support to design experiment to find out the functions that specific genome does. The strategy of functional genomics has widened the scope of biological investigations. This strategy is based on systematic study of single gene or protein to all genes. Therefore, the large-scale experimental methodologies characterise the functional genomics. Hence, the functional genomics provide the novel information about the genome. This eases the understanding of genes and function of proteins and protein interactions. The development of microarray technology and proteomics helped to explore the instantaneous events of all the genes expressed in a cell or tissue present at varying environmental conditions like temperature, pH, etc.

3) Comparative Genomics: The complete genome sequences of cellular organisms become available, the notable finding was recorded. It was found that one third of the genes encoded on each genome had no predictable or known function. e.g. in *E.coli* K_{12} about 40 % genes have unknown function. The level of evolutionary conservation of microbial proteins is rather uniform with about 70 % of gene products from each of sequenced genomes having homologous in distinct genomes. The function of these gene can be predicted by comparing different genomes and by transferring functional annotations of protein for better studies organisms to their orthologs (the same gene in different species that connect) as opposed to paralogs i.e., genes related by duplication within the genome from less studied organism. For better understanding of genomes, this makes comparative genomics as a powerful approach. Comparative genomics includes several aspects such as analysis of protein sets from completely sequenced genomes. General purpose databases and organisms specific databases used for comparative genomics.

5.1.3 Methods used for whole genome sequencing

The genome, of an organism (bacteria, virus, potato, human) is made up of DNA. Each organism has a unique DNA sequence which is composed of bases (A, T, C, and G). If the sequence of the bases in an organism are known, then we can identify its unique DNA fingerprint, or pattern. Determining the order of bases is called sequencing. Whole genome sequencing is a laboratory procedure that determines the order of bases in the genome of an organism in one process.

3 steps

There are several methods used for whole genome sequencing. Sequencing of genome chiefly comprises three steps: i) the cloning of the DNA to be sequenced, ii) the sequencing reactions and electrophoretic separations and iii) the analysis of ensuing data. Following are important methods of whole genome sequencing:

1) Chemical Methods:

Maxam-Gilbert sequencing

This method was developed by Maxam and Gilbert (1977). A restriction fragment of DNA is labelled with 32p at either its 5' or 3' using either of the enzymes polynucleotide kinase or terminal transferase. From a restriction map, an enzyme is selected to remove a small piece from one end of the molecule leaving just one end labelled. The DNA is then chemically cleaved at specific residues in five different reactions. These reactions are partially completed and partial digestion products are separated on a polyacrylamide gel and autoradiographed. The fragments having the labelled terminus are seen.

2) Whole Genome Shotgun Sequencing:

J. Craig Venter and H. Smith developed whole Genome shotgun sequencing and the two genome of bacteria *Haemophilus influenzae* and *Mycoplasma genitalium*. This method consists of four steps:

i) Library Construction: The chromosome is isolated from the desired cells following the methods of molecular biology. The isolated DNA is randomly fragmented into small pieces using ultrasonic waves. Then fragments are purified and attached to plasmid vectors. Plasmid with single insert is isolated. A library of plasmid clones are prepared by transforming *E. coli* strains with plasmid that lacked restriction enzymes.

ii) Random Sequencing: The DNA is purified from plasmid. Thousands of DNA fragments are sequenced using automated sequencer by using primers labelled with special dyes. Normally with universal primers, thousands of templates were used. These recognise the plasmid DNA sequences next to bacterial DNA insert. The whole genome is sequenced several times.

iii) Fragment - alignments and Gap Closure) By using special computer programme, the sequenced DNA fragments are clustered and assembled into longer stretches of sequence by comparing nucleotide sequence overlaps between fragments. Two fragments are joined to form a large stretch of DNA if the sequences at their ends overlapped and matched. This overlap comparison method resulted in a set of larger contiguous nucleotide sequence called contigs. The contigs are aligned in a proper order to form the completed genome sequence.

iv) Proof Reading: Then the proof reading of sequences is done carefully so that any ambiguities in the sequence could be resolved. The sequence is also checked for the presence of any frame shift mutation; if so, the mutation is corrected.

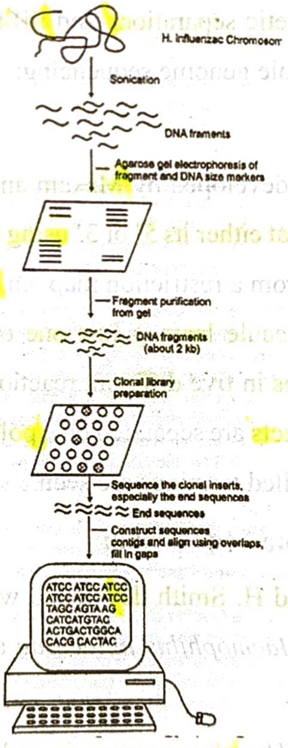

H. Influenzac Chromosom
Sonication
DNA framents
Agarose gel electrophoresis of fragment and DNA size markers
Fragment purification from gel
DNA fragments (about 2 kb)
Clonal library preparation
Sequence the clonal inserts, especially the end sequences
End sequences
Construct sequences contigs and align using overlaps, fill in gaps
ATCC ATCC ATCC
ATCC ATCC ATCC
TAGC AGTA AG
CATCATGTAC
ACTGACTGGCA
CACG CACTAC

Fig. 5.1 : Whole Genome shotgun sequencing

5.2 Proteomics:

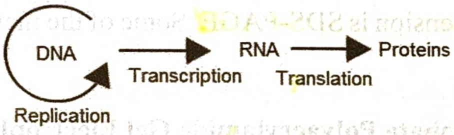
Proteomics is the study of all the proteins produced by a cell. Proteomics is the identification, analysis and large scale characterisation of proteome expressed by any cells, tissues and organs under the defined conditions. The major objectives of proteomics are: i) to

① ← Objectives of proteomics → ②

characterise post-transcriptional modifications in protein and ii) to prepare 3D map of a cell indicating the exact location of protein.

5.2.1Concept of proteomics:

The total protein component in a cell or organism is referred as the proteome. Proteomics deals with the study of proteomes. In broader term, proteomics is defined as the total protein content of a cell or that of an organism. The terms 'proteome' and 'proteomics' were coined in the early 1990 by Marc Wilkins. Proteomics helps in understanding of alteration in protein expression during different stages of life cycle or under stress condition. Likewise, Proteomics helps in understanding the structure and function of different proteins as well as protein - protein interactions of an organism. A minor defect in protein structure, it is function or alternation in expression pattern can be easily detected using proteomics studies. This is important with regards to drug development and understanding various biological processes, as proteins are the most favourable targets for various drugs. The first protein studies that can be called proteomics began in 1975 with the introduction of the two dimensional gel and mapping of the proteins from the bacterium Escherichia coli, guinea pig and mouse. Proteins are macromolecules; long chains of amino acids. This amino acid chain is constructed when the cellular machinery of the ribosome translates RNA transcripts from DNA in the cell's nucleus. The transfer of information within cells commonly follows this path from DNA to RNA to protein.



5.2.2 Types of proteomics:

The types of proteomics are as follows:

1) Structural Proteomics:

Structural proteomics deals with the study of structure and nature of protein complexes present in a particular cell organelle. It is mapping out the 3-D structure and nature of protein complexes present specifically in a particular cell organelle. The ultimate aim of structural proteomics is to build a body of structural information that will help predict the probable structure and potential function for almost any protein from knowledge of its coding sequence. Structural proteomics can also help assembling information about protein - protein interactions and about architecture of cells to explain how the expression of certain proteins contributes in cell's unique characteristics.

2) Functional Proteomics:

Functional proteomics refers to the use of proteomics techniques to analyse the characteristics of molecular protein-networks involved in a living cell. One of the recent successes of functional proteomics is the identification and analysis of molecular protein networks involved in the nuclear pore complex (NPC) in yeast. This success helps understand the translocation of molecules from nucleus to the cytoplasm and vice versa.

3) Expression Proteomics:

Expression proteomics concerned with to the quantitative study of protein expression between samples differing by some variable. The pattern of expression of the complete proteome or of its part (sub-proteome) between samples can be compared with the help of expression proteomics. The expression proteomics is quite useful in identifying disease specific proteins. For example, over expression or under-expression of proteins in cancerous cells and normal cells taken from a cancer patient and a normal individual, respectively, can be analysed using various techniques, such as two dimensional gel elec trophoresis, mass spectrometry, microarray, etc. This can help understand the development of cancer and facilitate development of drugs to treat cancer.

5.2.3 Methods used in proteome analysis:

Although new methods in proteomics are being developed, the traditional methods are; two-dimensional electrophoresis, and mass spectrometry. The first dimension uses iso-electric focusing and second dimension is SDS-PAGE. Some of the methods used in proteome (protein) analysis are as follow:

a) Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE):

Separation of some of the proteins dose not occur due to similar charge: mass ratio. Therefore, such proteins are treated first with an ionic detergent called sodium dodecyl sulphate (SDS) before to electrophoresis (PAGE). Therefore, such electrophoresis is called SDS-PAGE electrophoresis.

SDS-PAGE is a high resolution method used universally for analysing the mixture of proteins according to their respective size. SDS solubilised in soluble protein makes possible the analysis of the other insoluble mixture. Separation of the proteins doses not occur due to similar charge: mass ratio (z/m). Therefore, such proteins are treated first with an ionic detergents SDS before the start and during the course of electrophoresis.

Identical proteins are denatured by SDS resulting in their sub-units. The polypeptide chains get opened and extended. On the basis of their mass but not the charge, the molecules are separated. Electrophoretic separation is normally used for these reasons i.e. (i) gel acts as molecular sieves hence separates the molecules the molecules on the basis of their size, and (ii) gel suppresses conventional currents produced by small temperature gradient which improve the resolution. Polyacrylamide gel is used for this purpose due to its good nature (chemically inert, stable over a wide range of pH, temperature and transparent). Polyacrylamide gel is better for size fraction of proteins.

Why is electrophoretic separation used? (+) gel?

why is polyacrylamide gel used for electrophoretic separation?

(+) check resolving gel and stacking gel.

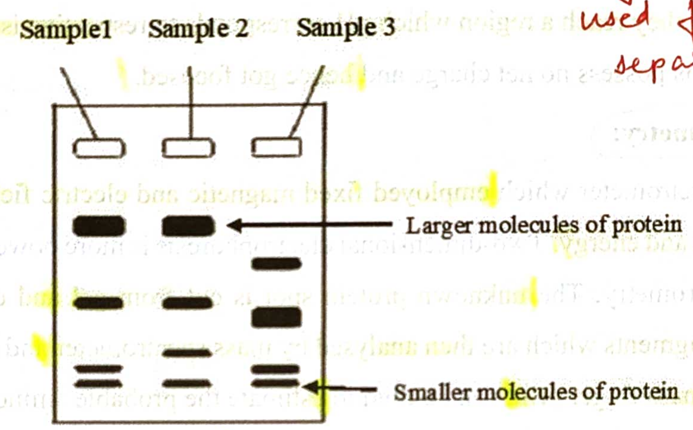


Fig. 5.2 SDS-PAGE analysis of proteins

The proteins are denatured and have negative charge with a uniform charge to mass ratio(z/m) when treated with SDS. Proteins migrates towards anode at alkaline pH through PAGE gel during electrophoresis. The smaller polypeptides moves faster followed by the larger polypeptides. Therefore, intrinsic charge on proteins is masked in SDS-PAGE. Hence separation is based on size. Molecular weight of the separated protein can be analysed by comparing the molecular weight of the standard protein and its mobility. In analysis of a complex a complex mixture of proteins the resolution is improved by the initial movement through a stacking gel. The final bands in the separating gel are sharper and focused in better way.

Two dimensional electrophoresis is very useful and effective method as it separating proteins and can resolve thousands of proteins in a mixture.

b) Iso-electric Focusing (IEF):

The biomolecule like proteins have electric charge which depends on molecule to molecule and conditions of medium (pH of buffer in which dissolved). Charged molecules can be separated by electrophoresis in gels. Due to the differences in amino acid composition proteins have net charge or iso-electric points (no charge) as a given pH of buffer.

The atmospheric substances such as proteins which differ in their isoelectric points can be separated by IEF. Isoelectric point is a pH value at which the net charges on molecules are zero. Ampholytes (i. e. complex mixture of synthetic polyamino-polycarboxylic acids) are introduced into gel to create the pH gradient (wide range from 3 to 10, or narrow range of 7 to 8). Then potential difference is applied across the gel. The molecule having difference in isoelectric points by a little as 0.01 pH unit can be separated. Proteins migrate depending on their charge until they reach a region which pH corresponds to respective iso-electric points at which pH proteins possess no net charge and hence got focused.

c) Mass spectrometry:

Mass spectrometer which employed fixed magnetic and electric field to separate ions of different mass and energy. Two-dimensional electrophoresis is more powerful when coupled with mass spectrometry. The unknown protein spot is cut from gel and cleaved by trypsin digestion into fragments which are then analysed by mass spectrometer and mass of fragments is plotted. This mass finger print can be used to estimate the probable amino acid composition of each fragment and tentatively identify the protein. The proteome and its charges can be studied very effectively by employing the two techniques together.

Mass spectrometry can also provide valuable information about covalent modification of proteins which can affect their activity. Mass spectrometry is very useful technique. It is used in identification of unknown compounds and determination of structural and chemical properties of compounds when present in small amount (10^{-4} -10^{-8} g). This technique involves: (i) the production of ions of the material in sample, (ii) their separation on the basis of their mass change (m:e), and (iii) determination of relative abundance of each ion.

Therefore, mass spectrometer consists of three components: the source of ion, an analyser, and a detector. It dose not directly measure the molecular mass but detects m:e ratio. Mass is measured in terms of Dalton (Da). One Dalton = 1/12th mass of a single atom of isotonic carbon (^{13} C).

In recent days, mass spectrometry has become an essential tool for analysis of genome and proteome in its many forms. It is capable of identifying and characterising proteins present even in picomoles (10^{-12}).

* Categories of Gene Prediction Programs –

① Ab-initio based approach –

- predicts genes based on given sequence only.
- relies on 2 major features –

(a) Existence of gene signals → include start and stop codons, introns splice signals, transcription factor binding site, ribosomal binding sites, poly-A sites, coding frame length limited to multiples of 3.

~~CHAPTER EIGHT~~

~~Gene Prediction~~

(b) Gene Content = statistical description of coding regions

↳ Nucleotide composition + statistical patterns of coding regions ≠ Non-coding regions

↳ ∴ use Markov models or HMMs.

With the rapid accumulation of genomic sequence information, there is a pressing need to use computational approaches to accurately predict gene structure. Computational gene prediction is a prerequisite for detailed functional annotation of genes and genomes. The process includes detection of the location of open reading frames (ORFs) and delineation of the structures of introns as well as exons if the genes of interest are of eukaryotic origin. The ultimate goal is to describe all the genes computationally with near 100% accuracy. The ability to accurately predict genes can significantly reduce the amount of experimental verification work required.

Defn of Computational gene prediction

However, this may still be a distant goal, particularly for eukaryotes, because many problems in computational gene prediction are still largely unsolved. Gene prediction, in fact, represents one of the most difficult problems in the field of pattern recognition. This is because coding regions normally do not have conserved motifs. Detecting coding potential of a genomic region has to rely on subtle features associated with genes that may be very difficult to detect.

Through decades of research and development, much progress has been made in prediction of prokaryotic genes. A number of gene prediction algorithms for prokaryotic genomes have been developed with varying degrees of success. Algorithms for eukarytotic gene prediction, however, are still yet to reach satisfactory results. This chapter describes a number of commonly used prediction algorithms, their theoretical basis, and limitations. Because of the significant differences in gene structures of prokaryotes and eukaryotes, gene prediction for each group of organisms is discussed separately. In addition, because of the predominance of protein coding genes in a genome (as opposed to rRNA and tRNA genes), the discussion focuses on the prediction of protein coding sequences.

① Ab-inito-based
② Homology-based
③ Consensus based

CATEGORIES OF GENE PREDICTION PROGRAMS

The current gene prediction methods can be classified into two major categories, ab initio–based and homology-based approaches. The ab initio–based approach predicts genes based on the given sequence alone. It does so by relying on two major features associated with genes. The first is the existence of gene signals, which include start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites, and polyadenylation (poly-A) sites. In addition, the triplet codon structure limits the coding frame length to multiples of three, which can be used as a condition for gene prediction. The second feature used by ab initio algorithms is gene content,

② Homology-based Method –

- predictions based on significant matches of the (query sequence) with (sequence of known genes)
- eg – if a translated DNA seq → found to be similar to known protein or protein family from db search
  ↓
  ∴ strong evidence that the region codes for a protein.

③ Consensus-based –

- Algorithms that make use of both gene-finding strategies.
- Combine prediction results from multiple individual programs to derive a consensus prediction.

* Gene Prediction in Prokaryotes –

Ⓘ Prokaryotic genome –

- relatively small genomes (size – 0.5 to 10 Mbp)
- High gene density (>90% contains coding sequence)
- Very few repetitive sequences
- composed of a single contiguous stretch of ORF coding for a single protein / RNA with no interruptions within a gene.

Ⓘⓘ Start codons –

- Start codon = ATG (AUG in mRNA) – codes for methionine
- Alternative start codons = GTG, TTG
- ∵ Multiple start codons in frame → ∴ presence of these codons not clear indication of translation initiation site.
- ∴ Other features that help locate start codon → Shine-Dalgarno sequence (ribosomal binding site) ⇄ stretch of purine-rich seq. complementary to 16S rRNA in the ribosome – located; consensus motif = AGGAGGT

which is statistical description of coding regions. It has been observed that nucleotide composition and statistical patterns of the coding regions tend to vary significantly from those of the noncoding regions. The unique features can be detected by employing probabilistic models such as Markov models or hidden Markov models (HMMs; see Chapter 6) to help distinguish coding from noncoding regions.

The homology-based method makes predictions based on significant matches of the query sequence with sequences of known genes. For instance, if a translated DNA sequence is found to be similar to a known protein or protein family from a database search, this can be strong evidence that the region codes for a protein. Alternatively, when possible exons of a genomic DNA region match a sequenced cDNA, this also provides experimental evidence for the existence of a coding region.

Some algorithms make use of both gene-finding strategies. There are also a number of programs that actually combine prediction results from multiple individual programs to derive a consensus prediction. This type of algorithms can therefore be considered as consensus based.

GENE PREDICTION IN PROKARYOTES

Prokaryotes, which include bacteria and Archaea, have relatively small genomes with sizes ranging from 0.5 to 10 Mbp (1 Mbp = 10^6 bp). The gene density in the genomes is high, with more than 90% of a genome sequence containing coding sequence. There are very few repetitive sequences. Each prokaryotic gene is composed of a single contiguous stretch of ORF coding for a single protein or RNA with no interruptions within a gene.

More detailed knowledge of the bacterial gene structure can be very useful in gene prediction. In bacteria, the majority of genes have a start codon ATG (or AUG in mRNA; because prediction is done at the DNA level, T is used in place of U), which codes for methionine. Occasionally, GTG and TTG are used as alternative start codons, but methionine is still the actual amino acid inserted at the first position. Because there may be multiple ATG, GTG, or TGT codons in a frame, the presence of these codons at the beginning of the frame does not necessarily give a clear indication of the translation initiation site. Instead, to help identify this initiation codon, other features associated with translation are used. One such feature is the ribosomal binding site, also called the *Shine-Delgarno sequence*, which is a stretch of purine-rich sequence complementary to 16S rRNA in the ribosome (Fig. 8.1). It is located immediately downstream of the transcription initiation site and slightly upstream of the translation start codon. In many bacteria, it has a consensus motif of AGGAGGT. Identification of the ribosome binding site can help locate the start codon.

At the end of the protein coding region is a stop codon that causes translation to stop. There are three possible stop codons, identification of which is straightforward. Many prokaryotic genes are transcribed together as one operon. The end of the operon is characterized by a transcription termination signal called *ρ-independent terminator*. The terminator sequence has a distinct stem-loop secondary structure

Ⓘⓘⓘ Stop codons –

- 3 possible stop codons – ∴ straightforward identification
- Operon = prokaryotic genes transcribed together as one
- ↳ characterized by – ρ-independent terminator
  = transcription termination signal
  = distinct stem-loop 2° structure followed by a string of . Ts.

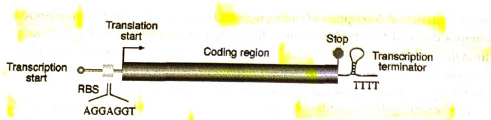


Figure 8.1: Structure of a typical prokaryotic gene structure. *Abbreviation:* RBS, ribosome binding site.

followed by a string of Ts. Identification of the terminator site, in conjunction with promoter site identification (see Chapter 9), can sometimes help in gene prediction.

Conventional Determination of Open Reading Frames

Without the use of specialized programs, prokaryotic gene identification can rely on manual determination of ORFs and major signals related to prokaryotic genes. Prokaryotic DNA is first subject to conceptual translation in all six possible frames, three frames forward and three frames reverse. Because a stop codon occurs in about every twenty codons by chance in a noncoding region, a frame longer than thirty codons without interruption by stop codons is suggestive of a gene coding region, although the threshold for an ORF is normally set even higher at fifty or sixty codons. The putative frame is further manually confirmed by the presence of other signals such as a start codon and Shine–Delgarno sequence. Furthermore, the putative ORF can be translated into a protein sequence, which is then used to search against a protein database. Detection of homologs from this search is probably the strongest indicator of a protein-coding frame.

Normal threshold of ORF set at >50 or 60 codons.

Methods for examining non-randomness of nucleotide distribution - ①

In the early stages of development of gene prediction algorithms, genes were predicted by examining the nonrandomness of nucleotide distribution. One method is based on the nucleotide composition of the third position of a codon. In a coding sequence, it has been observed that this position has a preference to use G or C over A or T. By plotting the GC composition at this position, regions with values significantly above the random level can be identified, which are indicative of the presence of ORFs (Fig. 8.2). In practice, because genes can be in any of the six frames, the statistical patterns are computed for all possible frames. In addition to codon bias, there is a similar method called TESTCODE (implemented in the commercial GCG package) that exploits the fact that the third codon nucleotides in a coding region tend to repeat themselves. By plotting the repeating patterns of the nucleotides at this position, coding and noncoding regions can be differentiated (see Fig. 8.2). The results of the two methods are often consistent. The two methods are often used in conjunction to confirm the results of each other.

②

Drawback

These statistical methods, which are based on empirical rules, examine the statistics of a single nucleotide (either G or C). They identify only typical genes and tend to miss atypical genes in which the rule of codon bias is not strictly followed. To improve the prediction accuracies, the new generation of prediction algorithms use more sophisticated statistical models.

* Markov Model -

- k = order of a Markov model
- 0-order → each base occurs independently with a given probability
  ↳ non-coding seqs.
- 1st-order → occurrence of a base depends on the base preceding it.
- 2nd order → looks at the preceding 2 bases to determine which base follows
  ↳ more characteristic of codons in a coding sequence.

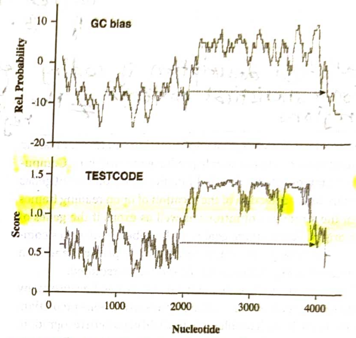


Figure 8.2: Coding frame detection of a bacterial gene using either the GC bias or the TESTCODE method. Both result in similar identification of a reading frame (*dashed arrows*).

Gene Prediction Using Markov Models and Hidden Markov Models

Markov models and HMMs can be very helpful in providing finer statistical description of a gene (see Chapter 6). A Markov model describes the probability of the distribution of nucleotides in a DNA sequence, in which the conditional probability of a particular sequence position depends on *k* previous positions. In this case, *k* is the order of a Markov model. A zero-order Markov model assumes each base occurs independently with a given probability. This is often the case for noncoding sequences. A first-order Markov model assumes that the occurrence of a base depends on the base preceding it. A second-order model looks at the preceding two bases to determine which base follows, which is more characteristic of codons in a coding sequence.

The use of Markov models in gene finding exploits the fact that oligonucleotide distributions in the coding regions are different from those for the noncoding regions. These can be represented with various orders of Markov models. Since a fixed-order Markov chain describes the probability of a particular nucleotide that depends on previous *k* nucleotides, the longer the oligomer unit, the more nonrandomness can be described for the coding region. Therefore, the higher the order of a Markov model, the more accurately it can predict a gene.

Because a protein-encoding gene is composed of nucleotides in triplets as codons, more effective Markov models are built in sets of three nucleotides, describing nonrandom distributions of trimers or hexamers, and so on. The parameters of a Markov model have to be trained using a set of sequences with known gene locations. Once the parameters of the model are established, it can be used to compute the nonrandom

steps ① ②

✱ Longer the oligomeric unit → ↑ Non-randomness

✱ ∴ ↑ order of Markov model → More accuracy in gene prediction.

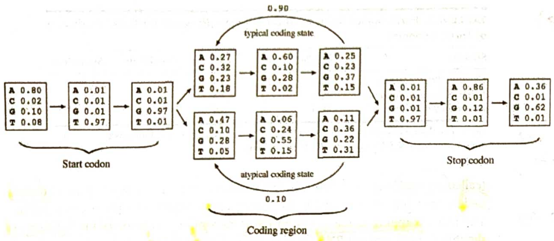


Figure 8.3: A simplified second-order HMM for prokaryotic gene prediction that includes a statistical model for start codons, stop codons, and the rest of the codons in a gene sequence represented by a typical model and an atypical model.

distributions of trimers or hexamers in a new sequence to find regions that are compatible with the statistical profiles in the learning set.

Statistical analyses have shown that pairs of codons (or amino acids at the protein level) tend to correlate. The frequency of six unique nucleotides appearing together in a coding region is much higher than by random chance. Therefore, a fifth-order Markov model, which calculates the probability of hexamer bases, can detect nucleotide correlations found in coding regions more accurately and is in fact most often used.

A potential problem of using a fifth-order Markov chain is that if there are not enough hexamers, which happens in short gene sequences, the method's efficacy may be limited. To cope with this limitation, a variable-length Markov model, called an *interpolated Markov model* (IMM), has been developed. The IMM method samples the largest number of sequence patterns with k ranging from 1 to 8 (dimers to ninemers) and uses a weighting scheme, placing less weight on rare k-mers and more weight on more frequent k-mers. The probability of the final model is the sum of probabilities of all weighted k-mers. In other words, this method has more flexibility in using Markov models depending on the amount of data available. Higher-order models are used when there is a sufficient amount of data and lower-order models are used when the amount of data is smaller.

Drawback of 5th order Markov model

∴ IMM developed

Advantage

It has been shown that the gene content and length distribution of prokaryotic genes can be either typical or atypical. Typical genes are in the range of 100 to 500 amino acids with a nucleotide distribution typical of the organism. Atypical genes are shorter or longer with different nucleotide statistics. These genes tend to escape detection using the typical gene model. This means that, to make the algorithm capable of fully describing all genes in a genome, more than one Markov model is needed. To combine different Markov models that represent typical and atypical nucleotide distributions creates an HMM prediction algorithm. A simplified HMM for gene finding is shown in Fig. 8.3.

Atypical genes tend to escape detection using typical gene model

The following describes a number of HMM/IMM-based gene finding programs for prokaryotic organisms.

GeneMark (http://opal.biology.gatech.edu/GeneMark/) is a suite of gene prediction programs based on the fifth-order HMMs. The main program – GeneMark.hmm – is trained on a number of complete microbial genomes. If the sequence to be predicted is from a nonlisted organism, the most closely related organism can be chosen as the basis for computation. Another option for predicting genes from a new organism is to use a self-trained program GeneMarkS as long as the user can provide at least 100 kbp of sequence on which to train the model. If the query sequence is shorter than 100 kbp, a GeneMark heuristic program can be used with some loss of accuracy. In addition to predicting prokaryotic genes, GeneMark also has a variant for eukaryotic gene prediction using HMM.

Glimmer (Gene Locator and Interpolated Markov Modeler, www.tigr.org/softlab/glimmer/glimmer.html) is a UNIX program from TIGR that uses the IMM algorithm to predict potential coding regions. The computation consists of two steps, namely model building and gene prediction. The model building involves training by the input sequence, which optimizes the parameters of the model. In an actual gene prediction, the overlapping frames are "flagged" to alert the user for further inspection. Glimmer also has a variant, GlimmerM, for eukaryotic gene prediction.

FGENESB (www.softberry.com/berry.phtml?topic=gfindb) is a web-based program that is also based on fifth-order HMMs for detecting coding regions. The program is specifically trained for bacterial sequences. It uses the Vertibi algorithm (see Chapter 6) to find an optimal match for the query sequence with the intrinsic model. A linear discriminant analysis (LDA) is used to further distinguish coding signals from noncoding signals.

These programs have been shown to be reasonably successful in finding genes in a genome. The common problem is imprecise prediction of translation initiation sites because of inefficient identification of ribosomal binding sites. This problem can be remedied by identifying the ribosomal binding site associated with a start codon. A number of algorithms have been developed solely for this purpose. RBSfinder is one such algorithm.

RBSfinder (ftp://ftp.tigr.org/pub/software/RBSfinder/) is a UNIX program that uses the prediction output from Glimmer and searches for the Shine–Delgarno sequences in the vicinity of predicted start sites. If a high-scoring site is found by the intrinsic probabilistic model, a start codon is confirmed; otherwise the program moves to other putative translation start sites and repeats the process.

Performance Evaluation

The accuracy of a prediction program can be evaluated using parameters such as sensitivity and specificity. To describe the concept of sensitivity and specificity accurately, four features are used: true positive (TP), which is a correctly predicted feature; false positive (FP), which is an incorrectly predicted feature; false negative (FN), which is a missed feature; and true negative (TN), which is the correctly predicted absence of

TP (True +ve) → correctly predicted feature
FP (False +ve) → incorrectly predicted feature
TN (True −ve) → correctly predicted absence of a feature
FN (False −ve) → missed feature.

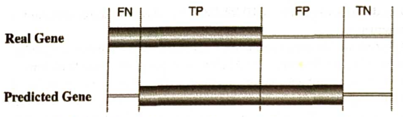


Figure 8.4: Definition of four basic measures of gene prediction accuracy at the nucleotide level. *Abbreviations:* FN, false negative; TP, true positive; FP, false positive; TN, true negative.

a feature (Fig. 8.4). Using these four terms, sensitivity (Sn) and specificity (Sp) can be described by the following formulas:

$$Sn = TP/(TP + FN) \qquad \text{(Eq. 8.1)}$$

$$Sp = TP/(TP + FP) \qquad \text{(Eq. 8.2)}$$

According to these formulas, *sensitivity* is the proportion of true signals predicted among all possible true signals. It can be considered as the ability to include correct predictions. In contrast, *specificity* is the proportion of true signals among all signals that are predicted. It represents the ability to exclude incorrect predictions. A program is considered accurate if both sensitivity and specificity are simultaneously high and approach a value of 1. In a case in which sensitivity is high but specificity is low, the program is said to have a tendency to overpredict. On the other hand, if the sensitivity is low but specificity high, the program is too conservative and lacks predictive power.

Because neither sensitivity nor specificity alone can fully describe accuracy, it is desirable to use a single value to summarize both of them. In the field of gene finding, a single parameter known as the correlation coefficient (CC) is often used, which is defined by the following formula:

$$CC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FP)(TN + FN)(FP + TN)}} \qquad \text{(Eq. 8.3)}$$

The value of the CC provides an overall measure of accuracy, which ranges from −1 to +1, with +1 meaning always correct prediction and −1 meaning always incorrect prediction. Table 8.1 shows a performance analysis using the Glimmer program as an example.

GENE PREDICTION IN EUKARYOTES

Eukaryotic nuclear genomes are much larger than prokaryotic ones, with sizes ranging from 10 Mbp to 670 Gbp (1 Gbp = 10^9 bp). They tend to have a very low gene density. In humans, for instance, only 3% of the genome codes for genes, with about 1 gene per 100 kbp on average. The space between genes is often very large and rich in repetitive sequences and transposable elements.

Most importantly, eukaryotic genomes are characterized by a mosaic organization in which a gene is split into pieces (called *exons*) by intervening noncoding sequences

$$S_n = \frac{TP}{(TP + FN)} \quad \Big| \quad S_p = \frac{TP}{(TP + FP)}$$

① ↑ S_n and ↑ S_p (approaching value of 1) → program considered accurate

② ↑ S_n and ↓ S_p → tendency to overpredict

③ ↓ S_n and ↑ S_p → too conservative ∴ lacks predictive power.

∴ Correlation coefficient (CC)

$$= \frac{(TP.TN) - (FP.FN)}{\sqrt{(TP+FP)(TN+FN)(FP+TN)}}$$

↓

∴ overall measure of accuracy

range → [illegible]

−1 to +1

always incorrect prediction — always correct prediction.

TABLE 8.1. Performance Analysis of the Glimmer Program for Gene Prediction of Three Genomes

Species	GC (%)	FN	FP	Sensitivity	Specificity
Campylobacter jejuni	30.5	10	19	99.3	98.7
Haemophilus influenzae	38.2	3	54	99.8	96.1
Helicobacter pylori	38.9	6	39	99.5	97.2

Note: The data sets were from three bacterial genomes (Aggarwal and Ramaswamy, 2002). *Abbreviations:* FN, false negative; FP, false positive.

(called *introns*) (Fig. 8.5). The nascent transcript from a eukaryotic gene is modified in three different ways before becoming a mature mRNA for protein translation. The first is capping at the 5′ end of the transcript, which involves methylation at the initial residue of the RNA. The second event is splicing, which is the process of removing introns and joining exons. The molecular basis of splicing is still not completely understood. What is known currently is that the splicing process involves a large RNA-protein complex called spliceosome. The reaction requires intermolecular interactions between a pair of nucleotides at each end of an intron and the RNA component of the spliceosome. To make the matter even more complex, some eukaryotic genes can have their transcripts spliced and joined in different ways to generate more than one transcript per gene. This is the phenomenon of alternative splicing. As to be discussed in more detail in Chapter 16, alternative splicing is a major mechanism for generating functional diversity in eukaryotic cells. The third modification is polyadenylation, which is the addition of a stretch of As (~250) at the 3′ end of the RNA.

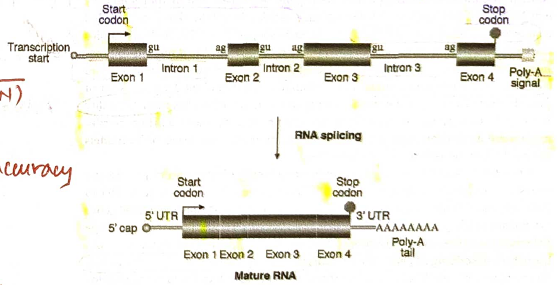


Figure 8.5: Structure of a typical eukaryotic RNA as primary transcript from genomic DNA and as mature RNA after posttranscriptional processing. *Abbreviations:* UTR, untranslated region; poly-A, polyadenylation.

3 modifications before becoming mature mRNA —
① capping at 5′ end of transcript (methylation)
② (a) splicing (b) alternative splicing
③ polyadenylation at 3′ end of RNA. (poly-A signal)
↳ consensus motif = CAATAAA (T/C)

(*) GT-AG rule – consensus motif of intron at –
① 5' splice junction → GTAAGT
② 3' —"— → $(Py)_{12}$NCAG

This process is controlled by a poly-A signal, a conserved motif slightly downstream of a coding region with a consensus CAATAAA(T/C).

Difficulty →

The main issue in prediction of eukaryotic genes is the identification of exons, introns, and splicing sites. From a computational point of view, it is a very complex and challenging problem. Because of the presence of split gene structures, alternative splicing, and very low gene densities, the difficulty of finding genes in such an environment is likened to finding a needle in a haystack. The needle to be found actually is broken into pieces and scattered in many different places. The job is to gather the pieces in the haystack and reproduce the needle in the correct order.

(*)

The good news is that there are still some conserved sequence features in eukaryotic genes that allow computational prediction. For example, the splice junctions of introns and exons follow the GT–AG rule in which an intron at the 5′ splice junction has a consensus motif of GTAAGT; and at the 3′ splice junction is a consensus motif of $(Py)_{12}$NCAG (see Fig. 8.5). Some statistical patterns useful for prokaryotic gene finding can be applied to eukaryotic systems as well. For example, nucleotide compositions and codon bias in coding regions of eukaryotes are different from those of the noncoding regions. Hexamer frequencies in coding regions are also higher than in the noncoding regions. Most vertebrate genes use ATG as the translation start codon and have a uniquely conserved flanking sequence call a *Kozak sequence* (CCGCCATGG). In addition, most of these genes have a high density of CG dinucleotides near the transcription start site. This region is referred to as a CpG island (*p* refers to the phosphodiester bond connecting the two nucleotides), which helps to identify the transcription initiation site of a eukaryotic gene. The poly-A signal can also help locate the final coding sequence.

Kozak sequence
CCGCCATGG

→ CpG island
p = phosphodiester bond connecting 2 nucleotides

Gene Prediction Programs

To date, numerous computer programs have been developed for identifying eukaryotic genes. They fall into all three categories of algorithms: ab initio based, homology based, and consensus based. Most of these programs are organism specific because training data sets for obtaining statistical parameters have to be derived from individual organisms. Some of the algorithms are able to predict the most probable exons as well as suboptimal exons providing information for possible alternative spliced transcription products.

Ab Initio–Based Programs

The goal of the ab initio gene prediction programs is to discriminate exons from noncoding sequences and subsequently join the exons together in the correct order. The main difficulty is correct identification of exons. To predict exons, the algorithms rely on two features, gene signals and gene content. Signals include gene start and stop sites and putative splice sites, recognizable consensus sequences such as poly-A sites. *Gene content* refers to coding statistics, which includes nonrandom nucleotide distribution, amino acid distribution, synonymous codon usage, and hexamer frequencies. Among these features, the hexamer frequencies appear to be most discriminative for coding potentials. To derive an assessment for this feature, HMMs can be used, which require proper training. In addition to HMMs, neural network-based algorithms are also common in the gene prediction field. This begs the question of what is a neural network algorithm. A brief introduction is given next.

} Goal

→ most discriminative for coding potentials

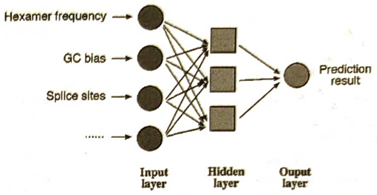


Figure 8.6: Architecture of a neural network for eukaryotic gene prediction.

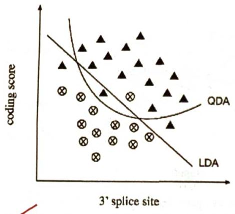
Prediction Using Neural Networks. A *neural network* (or *artificial neural network*) is a statistical model with a special architecture for pattern recognition and classification. It is composed of a network of mathematical variables that resemble the biological nervous system, with variables or nodes connected by weighted functions that are analogous to synapses (Fig. 8.6). Another aspect of the model that makes it look like a biological neural network is its ability to "learn" and then make predictions after being trained. The network is able to process information and modify parameters of the weight functions between variables during the training stage. Once it is trained, it is able to make automatic predictions about the unknown.

Steps of NN –

In gene prediction, a neural network is constructed with multiple layers; the input, output, and hidden layers. The input is the gene sequence with intron and exon signals. The output is the probability of an exon structure. Between input and output, there may be one or several hidden layers where the machine learning takes place. ① The machine learning process starts by feeding the model with a sequence of known gene structure. ② The gene structure information is separated into several classes of features such as hexamer frequencies, splice sites, and GC composition during training. ③ The weight functions in the hidden layers are adjusted during this process to recognize the nucleotide patterns and their relationship with known structures. ④ When the algorithm predicts an unknown sequence after training, it applies the same rules learned in training to look for patterns associated with the gene structures.

The frequently used ab initio programs make use of neural networks, HMMs, and discriminant analysis, which are described next.

GRAIL (Gene Recognition and Assembly Internet Link; http://compbio.ornl.gov/public/tools/) is a web-based program that is based on a neural network algorithm. The program is trained on several statistical features such as splice junctions, start



(*)

Figure 8.7: Comparison of two discriminant analysis, LDA and QDA. ▲ coding features; ⊗ noncoding features.

and stop codons, poly-A sites, promoters, and CpG islands. The program scans the query sequence with windows of variable lengths and scores for coding potentials and finally produces an output that is the result of exon candidates. The program is currently trained for human, mouse, *Arabidopsis*, *Drosophila*, and *Escherichia coli* sequences.

Prediction Using Discriminant Analysis. Some gene prediction algorithms rely on discriminant analysis, either LDA or quadratic discriminant analysis (QDA), to improve accuracy. LDA works by plotting a two-dimensional graph of coding signals versus all potential 3′ splice site positions and drawing a diagonal line that best separates coding signals from noncoding signals based on knowledge learned from training data sets of known gene structures (Fig. 8.7). QDA draws a curved line based on a quadratic function instead of drawing a straight line to separate coding and noncoding features. This strategy is designed to be more flexible and provide a more optimal separation between the data points.

LDA

QDA

FGENES (Find Genes; www.softberry.com/) is a web-based program that uses LDA to determine whether a signal is an exon. In addition to FGENES, there are many variants of the program. Some programs, such as FGENESH, make use of HMMs. There are others, such as FGENESH_C, that are similarity based. Some programs, such as FGENESH+, combine both ab initio and similarity-based approaches.

MZEF (Michael Zhang's Exon Finder; http://argon.cshl.org/genefinder/) is a web-based program that uses QDA for exon prediction. Despite the more complex mathematical functions, the expected increase in performance has not been obvious in actual gene prediction.

Prediction Using HMMs. GENSCAN (http://genes.mit.edu/GENSCAN.html) is a web-based program that makes predictions based on fifth-order HMMs. It combines hexamer frequencies with coding signals (initiation codons, TATA box, cap site, poly-A, etc.) in prediction. Putative exons are assigned a probability score (P) of being a true exon. Only predictions with $P > 0.5$ are deemed reliable. This program is trained for sequences from vertebrates, *Arabidopsis*, and maize. It has been used extensively in annotating the human genome (see Chapter 17).

HMMgene (www.cbs.dtu.dk/services/HMMgene) is also an HMM-based web program. The unique feature of the program is that it uses a criterion called the *conditional maximum likelihood* to discriminate coding from noncoding features. If a sequence already has a subregion identified as coding region, which may be based on similarity with cDNAs or proteins in a database, these regions are locked as coding regions. An HMM prediction is subsequently made with a bias toward the locked region and is extended from the locked region to predict the rest of the gene coding regions and even neighboring genes. The program is in a way a hybrid algorithm that uses both ab initio-based and homology-based criteria.

Homology-Based Programs

Homology-based programs are based on the fact that exon structures and exon sequences of related species are highly conserved. When potential coding frames in a query sequence are translated and used to align with closest protein homologs found in databases, near perfectly matched regions can be used to reveal the exon boundaries in the query. This approach assumes that the database sequences are correct. It is a reasonable assumption in light of the fact that many homologous sequences to be compared with are derived from cDNA or expressed sequence tags (ESTs) of the same species. With the support of experimental evidence, this method becomes rather efficient in finding genes in an unknown genomic DNA.

Assumption

Drawback of Homology-based programs

The drawback of this approach is its reliance on the presence of homologs in databases. If the homologs are not available in the database, the method cannot be used. Novel genes in a new species cannot be discovered without matches in the database. A number of publicly available programs that use this approach are discussed next.

GenomeScan (http://genes.mit.edu/genomescan.html) is a web-based server that combines GENSCAN prediction results with BLASTX similarity searches. The user provides genomic DNA and protein sequences from related species. The genomic DNA is translated in all six frames to cover all possible exons. The translated exons are then used to compare with the user-supplied protein sequences. Translated genomic regions having high similarity at the protein level receive higher scores. The same sequence is also predicted with a GENSCAN algorithm, which gives exons probability scores. Final exons are assigned based on combined score information from both analyses.

Advantages of Homology-based programs

EST2Genome (http://bioweb.pasteur.fr/seqanal/interfaces/est2genome.html) is a web-based program purely based on the sequence alignment approach to define intron–exon boundaries. The program compares an EST (or cDNA) sequence with a genomic DNA sequence containing the corresponding gene. The alignment is done using a dynamic programming–based algorithm. One advantage of the approach is the ability to find very small exons and alternatively spliced exons that are very difficult to predict by any ab initio–type algorithms. Another advantage is that there is no need

(1)

(2)

no need of model training
∴ more flexibility

Drawback

for model training, which provides much more flexibility for gene prediction. The limitation is that EST or cDNA sequences often contain errors or even introns if the transcripts are not completely spliced before reverse transcription.

SGP-1 (Syntenic Gene Prediction; http://195.37.47.237/sgp-1/) is a similarity-based web program that aligns two genomic DNA sequences from closely related organisms. The program translates all potential exons in each sequence and does pairwise alignment for the translated protein sequences using a dynamic programming approach. The near-perfect matches at the protein level define coding regions. Similar to EST2Genome, there is no training needed. The limitation is the need for two homologous sequences having similar genes with similar exon structures; if this condition is not met, a gene escapes detection from one sequence when there is no counterpart in another sequence.

TwinScan (http://genes.cs.wustl.edu/) is also a similarity-based gene-finding server. It is similar to GenomeScan in that it uses GenScan to predict all possible exons from the genomic sequence. The putative exons are used for BLAST searching to find closest homologs. The putative exons and homologs from BLAST searching are aligned to identify the best match. Only the closest match from a genome database is used as a template for refining the previous exon selection and exon boundaries.

Consensus-Based Programs

Working

Because different prediction programs have different levels of sensitivity and specificity, it makes sense to combine results of multiple programs based on consensus. This idea has prompted development of consensus-based algorithms. These programs work by retaining common predictions agreed by most programs and removing inconsistent predictions. Such an integrated approach may improve the specificity by correcting the false positives and the problem of overprediction. However, since this procedure punishes novel predictions, it may lead to lowered sensitivity and missed predictions. Two examples of consensus-based programs are given next.

} Drawback

GeneComber (www.bioinformatics.ubc.ca/genecomber/index.php) is a web server that combines HMMgene and GenScan prediction results. The consistency of both prediction methods is calculated. If the two predictions match, the exon score is reinforced. If not, exons are proposed based on separate threshold scores.

DIGIT (http://digit.gsc.riken.go.jp/cgi-bin/index.cgi) is another consensus-based web server. It uses prediction from three ab initio programs – FGENESH, GENSCAN, and HMMgene. It first compiles all putative exons from the three gene-finders and assigns ORFs with associated scores. It then searches a set of exons with the highest additive score under the reading frame constraints. During this process, a Bayesian procedure and HMMs are used to infer scores and search the optimal exon set which gives the final designation of gene structure.

Performance Evaluation

Because of extra layers of complexity for eukaryotic gene prediction, the sensitivity and specificity have to be defined on the levels of nucleotides, exons, and entire genes.

TABLE 8.2. Accuracy Comparisons for a Number of Ab Initio Gene Prediction Programs at Nucleotide and Exon Levels

	Nucleotide level			Exon level				
	Sn	Sp	CC	Sn	Sp	(Sn + Sp)/2	ME	WE
FGENES	0.86	0.88	0.83	0.67	0.67	0.67	0.12	0.09
GeneMark	0.87	0.89	0.83	0.53	0.54	0.54	0.13	0.11
Genie	0.91	0.90	0.88	0.71	0.70	0.71	0.19	0.11
GenScan	0.95	0.90	0.91	0.70	0.70	0.70	0.08	0.09
HMMgene	0.93	0.93	0.91	0.76	0.77	0.76	0.12	0.07
Morgan	0.75	0.74	0.74	0.46	0.41	0.43	0.20	0.28
MZEF	0.70	0.73	0.66	0.58	0.59	0.59	0.32	0.23

Note: The data sets used were single mammalian gene sequences (performed by Sanja Rogic, from www.cs.ubc.ca/~rogic/evaluation/tablesgen.html.

Abbreviations: Sn, sensitivity; Sp, specificity; CC, correlation coefficient; ME, missed exons; WE, wrongly predicted exons.

The sensitivity at the exon and gene level is the proportion of correctly predicted exons or genes among actual exons or genes. The specificity at the two levels is the proportion of correctly predicted exons or genes among all predictions made. For exons, instead of using CC, an average of sensitivity and specificity at the exon level is used instead. In addition, the proportion of missed exons and missed genes as well as wrongly predicted exons and wrong genes, which have no overlaps with true exons or genes, often have to be indicated.

★ Sn and Sp of eukaryotes @ exon and gene level

By introducing these measures, the criteria for prediction accuracy evaluation become more stringent (Table 8.2). For example, a correct exon requires all nucleotides belonging to the exon to be predicted correctly. For a correctly predicted gene, all nucleotides and all exons have to be predicted correctly. One single error at the nucleotide level can negate the entire gene prediction. Consequently, the accuracy values reported on the levels of exons and genes are much lower than those for nucleotides.

When a new gene prediction program is published, the accuracy level is usually reported. However, the reported performance should be treated with caution because the accuracy is usually estimated based on particular datasets, which may have been optimized for the program. The datasets used are also mainly composed of short genomic sequences with simple gene structures. When the programs are used in gene prediction for truly unknown eukaryotic genomic sequences, the accuracy can become much lower. Because of the lack of unbiased and realistic datasets and objective comparison for eukaryotic gene prediction, it is difficult to know the true accuracy of the current prediction tools.

At present, no single software program is able to produce consistent superior results. Some programs may perform well on certain types of exons (e.g., internal or single exons) but not others (e.g., initial and terminal exons). Some are sensitive to the G-C content of the input sequences or to the lengths of introns and exons. Most programs make overpredictions when genes contain long introns. In sum, they all suffer from the problem of generating a high number of false positives and false negatives. This is especially true for ab initio–based algorithms. For complex genomes such as the human genome, most popular programs can predict no more than 40% of the genes exactly right. Drawing consensus from results by multiple prediction programs may enhance performance to some extent.

SUMMARY

Computational prediction of genes is one of the most important steps of genome sequence analysis. For prokaryotic genomes, which are characterized by high gene density and noninterrupted genes, prediction of genes is easier than for eukaryotic genomes. Current prokaryotic gene prediction algorithms, which are based on HMMs, have achieved reasonably good accuracy. Many difficulties still persist for eukaryotic gene prediction. The difficulty mainly results from the low gene density and split gene structure of eukaryotic genomes. Current algorithms are either ab initio based, homology based, or a combination of both. For ab initio–based eukaryotic gene prediction, the HMM type of algorithm has overall better performance in differentiating intron–exon boundaries. The major limitation is the dependency on training of the statistical models, which renders the method to be organism specific. The homology-based algorithms in combination with HMMs may yield improved accuracy. The method is limited by the availability of identifiable sequence homologs in databases. The combined approach that integrates statistical and homology information may generate further improved performance by detecting more genes and more exons correctly. With rapid advances in computational techniques and understanding of the splicing mechanism, it is hoped that reliable eukaryotic gene prediction can become more feasible in the near future.

CHAPTER NINE

Promoter and Regulatory Element Prediction

(P) (RE)

Promoters

An issue related to gene prediction is promoter prediction. Promoters are DNA elements located in the vicinity of gene start sites (which should not be confused with the translation start sites) and serve as binding sites for the gene transcription machinery, consisting of RNA polymerases and transcription factors. Therefore, these DNA elements directly regulate gene expression. Promoters and regulatory elements are traditionally determined by experimental analysis. The process is extremely time consuming and laborious. Computational prediction of promoters and regulatory elements is especially promising because it has the potential to replace a great deal of extensive experimental analysis.

Drawback of Traditional experimental analysis

However, computational identification of promoters and regulatory elements is also a very difficult task, for several reasons. First, promoters and regulatory elements are not clearly defined and are highly diverse. Each gene seems to have a unique combination of sets of regulatory motifs that determine its unique temporal and spatial expression. There is currently a lack of sufficient understanding of all the necessary regulatory elements for transcription. Second, the promoters and regulatory elements cannot be translated into protein sequences to increase the sensitivity for their detection. Third, promoter and regulatory sites to be predicted are normally short (six to eight nucleotides) and can be found in essentially any sequence by random chance, thus resulting in high rates of false positives associated with theoretical predictions.

Difficulties in computational identification of promoters + RE

Current solutions for providing preliminary identification of these elements are to combine a multitude of features and use sophisticated algorithms that give either ab initio–based predictions or predictions based on evolutionary information or experimental data. These computational approaches are described in detail in this chapter following a brief introduction to the structures of promoters and regulatory elements in both prokaryotes and eukaryotes.

PROMOTER AND REGULATORY ELEMENTS IN PROKARYOTES

P

In bacteria, transcription is initiated by RNA polymerase, which is a multi-subunit enzyme. The σ subunit (e.g., σ^{70}) of the RNA polymerase is the protein that recognizes specific sequences upstream of a gene and allows the rest of the enzyme complex to bind. The upstream sequence where the σ protein binds constitutes the promoter sequence. This includes the sequence segments located 35 and 10 base pairs (bp) upstream from the transcription start site. They are also referred to as the −35 and −10 boxes. For the σ^{70} subunit in *Escherichia coli*, for example, the −35 box has a consensus sequence of TTGACA. The −10 box has a consensus of TATAAT. The promoter sequence may determine the expression of one gene or a number of linked genes downstream. In the latter case, the linked genes form an operon, which is controlled by the promoter.

E. coli → −35 box consensus → TTGACA
−10 box consensus → TATAAT

- RNA Pol I – transcription of ribosomal RNA
- RNA Pol III – transcription of tRNA
- RNA Pol II – transcription of protein-encoding genes (synthesis of mRNA).

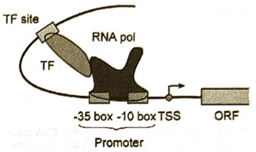


Figure 9.1: Schematic representation of elements involved in bacterial transcription initiation. RNA polymerase binds to the promoter region, which initiates transcription through interaction with transcription factors binding at different sites. *Abbreviations:* TSS, transcription start site; ORF, reading frame; pol, polymerase; TF, transcription factor (see color plate section).

P

RE

In addition to the RNA polymerase, there are also a number of DNA-binding proteins that facilitate the process of transcription. These proteins are called *transcription factors*. They bind to specific DNA sequences to either enhance or inhibit the function of the RNA polymerase. The specific DNA sequences to which the transcription factors bind are referred to as *regulatory elements*. The regulatory elements may bind in the vicinity of the promoter or bind to a site several hundred bases away from the promoter. The reason that the regulatory proteins binding at long distance can still exert their effect is because of the flexible structure of DNA, which is able to bend and and exert its effect by bringing the transcription factors in close contact with the RNA polymerase complex (Fig. 9.1).

PROMOTER AND REGULATORY ELEMENTS IN EUKARYOTES

In eukaryotes, gene expression is also regulated by a protein complex formed between transcription factors and RNA polymerase. However, eukaryotic transcription has an added layer of complexity in that there are three different types of RNA polymerase complexes, namely RNA polymerases I, II, and III. Each polymerase transcribes different sets of genes. RNA polymerases I and III are responsible for the transcription of ribosomal RNAs and tRNAs, respectively. RNA polymerase II is exclusively responsible for transcribing protein-encoding genes (or synthesis of mRNAs).

Unlike in prokaryotes, where genes often form an operon with a shared promoter, each eukaryotic gene has its own promoter. The eukaryotic transcription machinery also requires many more transcription factors than its prokaryotic counterpart to help initiate transcription. Furthermore, eukaryotic RNA polymerase II does not directly bind to the promoter, but relies on a dozen or more transcription factors to recognize and bind to the promoter in a specific order before its own binding around the promoter.

(★) Eukaryotic promoter core → TATA box
↳ consensus motif → TATA(A/T)A(A/T)

Exception → Housekeeping genes do not have TATA box in their promoters.

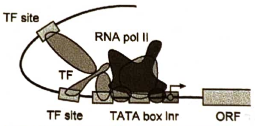


Figure 9.2: Schematic diagram of an eukaryotic promoter with transcription factors and RNA polymerase bound to the promoter. *Abbreviations:* Inr, initiator sequence; ORF, reading frame; pol, polymerase; TF, transcription factor (see color plate section).

The core of many eukaryotic promoters is a so-called TATA box located 30 bps upstream from the transcription start site, having a consensus motif TATA(A/T)A (A/T) (Fig. 9.2.). However, not all eukaryotic promoters contain the TATA box. Many genes such as housekeeping genes do not have the TATA box in their promoters. Still, the TATA box is often used as an indicator of the presence of a promoter. In addition, many genes have a unique initiator sequence (Inr), which is a pyrimidine-rich sequence with a consensus (C/T)(C/T)CA(C/T)(C/T). This site coincides with the transcription start site. Most of the transcription factor binding sites are located within 500 bp upstream of the transcription start site. Some regulatory sites can be found tens of thousands base pairs away from the gene start site. Occasionally, regulatory elements are located downstream instead of upstream of the transcription start site. Often, a cluster of transcription factor binding sites spread within a wide range to work synergistically to enhance transcription initiation.

(★) Inr (initiator sequence) unique
↳ (C/T)(C/T)CA(C/T)(C/T)

PREDICTION ALGORITHMS

Current algorithms for predicting promoters and regulatory elements can be categorized as either ab initio based, which make de novo predictions by scanning individual sequences; or similarity based, which make predictions based on alignment of homologous sequences; or expression profile based using profiles constructed from a number of coexpressed gene sequences from the same organism. The similarity type of prediction is also called phylogenetic footprinting. As mentioned, because RNA polymerase II transcribes the eukaryotic mRNA genes, most algorithms are thus focused on prediction of the RNA polymerase II promoter and associated regulatory elements. Each of the categories is discussed in detail next.

(★)

① Ab-initio based
② Similarity based (Phylogenetic footprinting)
③ Expression profiling based

Ab Initio–Based Algorithms

This type of algorithm predicts prokaryotic and eukaryotic promoters and regulatory elements based on characteristic sequences patterns for promoters and regulatory elements. Some ab initio programs are signal based, relying on characteristic promoter sequences such as the TATA box, whereas others rely on content information such as hexamer frequencies. The advantage of the ab initio method is that the sequence can be applied as such without having to obtain experimental information. The limitation is the need for training, which makes the prediction programs species specific. In addition, this type of method has a difficulty in discovering new, unknown motifs.

(a) Signal based
- characteristic promoter sequences (TATA box) ← rely on → - hexamer frequencies
(b) content information

Advantage of Ab-initio → sequence applied without obtaining experimental information
Disadvantage of Ab-initio → need for training to make prediction programs - species specific + generate ↑ rate of FP

The conventional approach to detecting a promoter or regulatory site is through matching a consensus sequence pattern represented by regular expressions (see Chapter 7) or matching a position-specific scoring matrix (PSSM; see Chapter 6) constructed from well-characterized binding sites. In either case, the consensus sequences or the matrices are relatively short, covering 6 to 10 bases. As described in Chapter 7, to determine whether a query sequence matches a weight matrix, the sequence is scanned through the matrix. Scores of matches and mismatches at all matrix positions are summed up to give a log odds score, which is then evaluated for statistical significance. This simple approach, however, often has difficulty differentiating true promoters from random sequence matches and generates high rates of false positives as a result.

To better discriminate true motifs from background noise, a new generation of algorithms has been developed that take into account the higher order correlation of multiple subtle features by using discriminant functions, neural networks, or hidden Markov models (HMMs) that are capable of incorporating more neighboring sequence information. To further improve the specificity of prediction, some algorithms selectively exclude coding regions and focus on the upstream regions (0.5 to 2.0 kb) only, which are most likely to contain promoters. In that sense, promoter prediction and gene prediction are coupled. (★)

Prediction for Prokaryotes

One of the unique aspects in prokaryotic promoter prediction is the determination of operon structures, because genes within an operon share a common promoter located upstream of the first gene of the operon. Thus, operon prediction is the key in prokaryotic promoter prediction. Once an operon structure is known, only the first gene is predicted for the presence of a promoter and regulatory elements, whereas other genes in the operon do not possess such DNA elements.

There are a number of methods available for prokaryotic operon prediction. The most accurate is a set of simple rules developed by Wang et al. (2004). This method relies on two kinds of information: gene orientation and intergenic distances of a pair of genes of interest and conserved linkage of the genes based on comparative genomic analysis. More about gene linkage patterns across genomes is introduced in Chapters 16 and 18. A scoring scheme is developed to assign operons with different levels of confidence (Fig. 9.3). This method is claimed to produce accurate identification of an operon structure, which in turn facilitates the promoter prediction.

This newly developed scoring approach is, however, not yet available as a computer program. The prediction can be done manually using the rules, however. The few dedicated programs for prokaryotic promoter prediction do not apply the Wang et al. rule for historical reasons. The most frequently used program is BPROM.

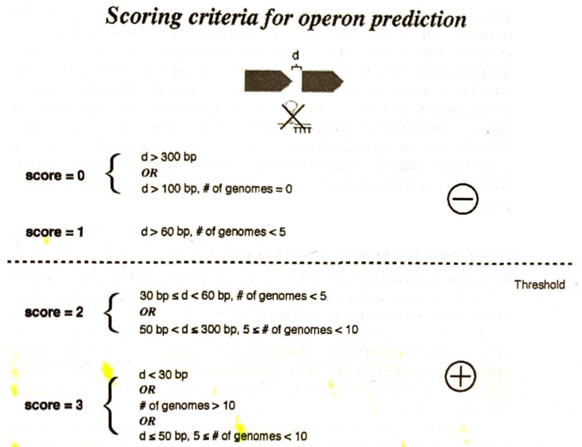

Scoring criteria for operon prediction
d
score = 0
d > 300 bp
OR
d > 100 bp, # of genomes = 0
⊖
score = 1
d > 60 bp, # of genomes < 5
Threshold
score = 2
30 bp ≤ d < 60 bp, # of genomes < 5
OR
50 bp < d ≤ 300 bp, 5 ≤ # of genomes < 10
⊕
score = 3
d < 30 bp
OR
# of genomes > 10
OR
d ≤ 50 bp, 5 ≤ # of genomes < 10

Figure 9.3: Prediction of operons in prokaryotes based on a scoring scheme developed by Wang et al. (2004). This method states that, for two adjacent genes transcribed in the same orientation and without a ρ-independent transcription termination signal in between, the score is assigned 0 if the intergenic distance is larger than 300 bp regardless of the gene linkage pattern or if the distance is larger than 100 bp with the linkage not observed in other genomes. The score is assigned 1 if the intergenic distance is larger than 60 bp with the linkage shared in less than five genomes. The score is assigned 2 if the distance of the two genes is between 30 and 60 bp with the linkage shared in less than five genomes or if the distance is between 50 and 300 bp with the linkage shared in between five to ten genomes. The score is assigned 3 if the intergenic distance is less than 30 bp regardless of the conserved linkage pattern or if the linkage is conserved in more than ten genomes regardless of the intergenic distance or if the distance is less than 50 bp with the linkage shared in between five to ten genomes. A minimum score of 2 is considered the threshold for assigning the two genes in one operon.

BPROM (www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb) is a web-based program for prediction of bacterial promoters. It uses a linear discriminant function (see Chapter 8) combined with signal and content information such as consensus promoter sequence and oligonucleotide composition of the promoter sites. This program first predicts a given sequence for bacterial operon structures by using an intergenic distance of 100 bp as basis for distinguishing genes to be in an operon. This rule is more arbitrary than the Wang et al. rule, leading to high rates of false positives. Once the operons are assigned, the program is able to predict putative promoter sequences. Because most bacterial promoters are located within 200 bp of the protein coding region, the program is most effectively used when about 200 bp of upstream sequence of the first gene of an operon is supplied as input to increase specificity.

FindTerm (http://sun1.softberry.com/berry.phtml?topic=findterm&group=programs&subgroup=gfindb) is a program for searching bacterial ρ-independent termination signals located at the end of operons. It is available from the same site as FGENES and BPROM. The predictions are made based on matching of known profiles of the termination signals combined with energy calculations for the derived RNA secondary structures for the putative hairpin-loop structure (see Chapter 16). The sequence region that scores best in features and energy terms is chosen as the prediction. The information can sometimes be useful in defining an operon.

Prediction for Eukaryotes

⊛ Disadvantage
↓
tendency for both FP & FN

The ab initio method for predicting eukaryotic promoters and regulatory elements also relies on searching the input sequences for matching of consensus patterns of known promoters and regulatory elements. The consensus patterns are derived from experimentally determined DNA binding sites which are compiled into profiles and stored in a database for scanning an unknown sequence to find similar conserved patterns. However, this approach tends to generate very high rate of false positives owing to nonspecific matches with the short sequence patterns. Furthermore, because of the high variability of transcription factor binding sites, the simple sequence matching often misses true promoter sites, creating false negatives.

→ Disadvantage

To increase the specificity of prediction, a unique feature of eukaryotic promoter is employed, which is the presence of CpG islands. It is known that many vertebrate genes are characterized by a high density of CG dinucleotides near the promoter region overlapping the transcription start site (see Chapter 8). By identifying the CpG islands, promoters can be traced on the immediate upstream region from the islands. ⊛ By combining CpG islands and other promoter signals, the accuracy of prediction can be improved. Several programs have been developed based on the combined features to predict the transcription start sites in particular.

⊛ Cooperativity

As discussed, the eukaryotic transcription initiation requires cooperation of a large number of transcription factors. *Cooperativity* means that the promoter regions tend to contain a high density of protein-binding sites. Thus, finding a cluster of transcription factor binding sites often enhances the probability of individual binding site prediction.

A number of representatives of ab initio promoter prediction algorithms that incorporate the unique properties of eukaryotic promoters are introduced next.

CpGProD (http://pbil.univ-lyon1.fr/software/cpgprod.html) is a web-based program that predicts promoters containing a high density of CpG islands in mammalian genomic sequences. It calculates moving averages of GC% and CpG ratios (observed/expected) over a window of a certain size (usually 200 bp). When the values are above a certain threshold, the region is identified as a CpG island.

Eponine (http://servlet.sanger.ac.uk:8080/eponine/) is a web-based program that predicts transcription start sites based on a series of preconstructed PSSMs of several regulatory sites, such as the TATA box, the CCAAT box, and CpG islands. The query sequence from a mammalian source is scanned through the PSSMs. The sequence stretches with high-score matching to all the PSSMs, as well as matching of the spacing between the elements, are declared transcription start sites. A Bayesian method is also used in decision making.

Cluster-Buster (http://zlab.bu.edu/cluster-buster/cbust.html) is an HMM-based, web-based program designed to find clusters of regulatory binding sites. It works by detecting a region of high concentration of known transcription factor binding sites and regulatory motifs. A query sequence is scanned with a window size of 1 kb for putative regulatory motifs using motif HMMs. If multiple motifs are detected within a window, a positive score is assigned to each motif found. The total score of the window is the sum of each motif score subtracting a gap penalty, which is proportional to the distances between motifs. If the score of a certain region is above a certain threshold, it is predicted to contain a regulatory cluster.

FirstEF (First Exon Finder; http://rulai.cshl.org/tools/FirstEF/) is a web-based program that predicts promoters for human DNA. It integrates gene prediction with promoter prediction. It uses quadratic discriminant functions (see Chapter 8) to calculate the probabilities of the first exon of a gene and its boundary sites. A segment of DNA (15 kb) upstream of the first exon is subsequently extracted for promoter prediction on the basis of scores for CpG islands.

McPromoter (http://genes.mit.edu/McPromoter.html) is a web-based program that uses a neural network to make promoter predictions. It has a unique promoter model containing six scoring segments. The program scans a window of 300 bases for the likelihoods of being in each of the coding, noncoding, and promoter regions. The input for the neural network includes parameters for sequence physical properties, such as DNA bendability, plus signals such as the TATA box, initiator box, and CpG islands. The hidden layer combines all the features to derive an overall likelihood for a site being a promoter. Another unique feature is that McPromoter does not require that certain patterns must be present, but instead the combination of all features is important. For instance, even if the TATA box score is very low, a promoter prediction can still be made if the other features score highly. The program is currently trained for *Drosophila* and human sequences.

★ TSSW

TSSW (www.softberry.com/berry.phtml?topic=promoter) is a web program that distinguishes promoter sequences from non-promoter sequences based on a combination of unique content information such as hexamer/trimer frequencies and signal information such the TATA box in the promoter region. The values are fed to a linear discriminant function (see Chapter 8) to separate true motifs from background noise.

CONPRO (http://stl.bioinformatics.med.umich.edu/conpro) is a web-based program that uses a consensus method to identify promoter elements for human DNA.

★ Phylogenetic footprinting

- identification of conserved non-coding DNA elements that serve crucial functional roles.

# "Phylogenetic footprints"

- applicable to both prokaryotic + eukaryotic sequences.
- Caveats / Considerations -

① If organisms too closely related (human, chimpanzee)
↳ seq. difference not sufficient

② If Evolutionary distance too long (human, fish)
↳ may render P + RE undetectable

③ To extract non-coding sequences upstream of corresponding genes and focus the comparison to this region only. → ∴ helps prevent FP

To use the program, a user supplies the transcript sequence of a gene (cDNA). The program uses the information to search the human genome database for the position of the gene. It then uses the GENSCAN program to predict 5′ untranslated exons in the upstream region. Once the 5′-most exon is located, a further upstream region (1.5 kb) is used for promoter prediction, which relies on a combination of five promoter prediction programs, TSSG, TSSW, NNPP, PROSCAN, and PromFD. For each program, the highest score prediction is taken as the promoter in the region. If three predictions fall within a 100-bp region, this is considered a consensus prediction. If no three-way consensus is achieved, TSSG and PromFD predictions are taken. Because no coding sequence is used in prediction, specificity is improved relative to each individual program.

Phylogenetic Footprinting–Based Method / Similarity-based method

It has been observed that promoter and regulatory elements from closely related organisms such as human and mouse are highly conserved. The conservation is both at the sequence level and at the level of organization of the elements. Therefore, it is possible to obtain such promoter sequences for a particular gene through comparative analysis. The identification of conserved noncoding DNA elements that serve crucial functional roles is referred to as *phylogenetic footprinting*; the elements are called *phylogenetic footprints*. This type of method can apply to both prokaryotic and eukaryotic sequences.

The selection of organisms for comparison is an important consideration in this type of analysis. If the pair of organisms selected are too closely related, such as human and chimpanzee, the sequence difference between them may not be sufficient to filter out functional elements. On the other hand, if the organisms' evolutionary distances are too long, such as between human and fish, long evolutionary divergence may render promoter and other elements undetectable. One example of appropriate selection of species is the use of human and mouse sequences, which often yields informative results. ★

Another caveat of phylogenetic footprinting is to extract noncoding sequences upstream of corresponding genes and focus the comparison to this region only, which helps to prevent false positives. The predictive value of this method also depends on the quality of the subsequent sequence alignments. The advanced alignment programs introduced in Chapter 5 can be used. Even more sophisticated expectation maximization (EM) and Gibbs sampling algorithms can be used in detecting weakly conserved motifs.

There are software programs specifically designed to take advantage of the presence of phylogenetic footprints to make comparisons among a number of related species to identify putative transcription factor binding sites. The advantage in implementing the algorithms is that no training of the probabilistic models is required; ① hence, it is more broadly applicable. ② There is also a potential to discover new regulatory ③

3 Advantages

④ predictive value ∝ quality of subsequent sequence alignments

1

motifs shared among organisms. The obvious limitation is the constraint on the evolutionary distances among the orthologous sequences.

ConSite (http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite) is a web server that finds putative promoter elements by comparing two orthologous sequences. The user provides two individual sequences which are aligned by ConSite using a global alignment algorithm. Alternatively, the program accepts precomputed alignment. Conserved regions are identified by calculating identity scores, which are then used to compare against a motif database of regulatory sites (TRANSFAC). High-scoring sequence segments upstream of genes are returned as putative regulatory elements.

rVISTA (http://rvista.dcode.org/) is a similar cross-species comparison tool for promoter recognition. The program uses two orthologous sequences as input and first identifies all putative regulatory motifs based on TRANSFAC matches. It then aligns the two sequences using a local alignment strategy. The motifs that have the highest percent identity in the pairwise comparison are presented graphically as regulatory elements.

PromH(W) (www.softberry.com/berry.phtml?topic=promhw&group=programs&subgroup=promoter) is a web-based program that predicts regulatory sites by pairwise sequence comparison. The user supplies two orthologous sequences, which are aligned by the program to identify conserved regions. These regions are subsequently predicted for RNA polymerase II promoter motifs in both sequences using the TSSW program. Only the conserved regions having high scored promoter motifs are returned as results.

Bayes aligner (www.bioinfo.rpi.edu/applications/bayesian/bayes/bayes_align12.pl) is a web-based footprinting program. It aligns two sequences using a Bayesian algorithm which is a unique sequence alignment method. Instead of returning a single best alignment, the method generates a distribution of a large number of alignments using a full range of scoring matrices and gap penalties. Posterior probability values, which are considered estimates of the true alignment, are calculated for each alignment. By studying the distribution, the alignment that has the highest likelihood score, which is in the extreme margin of the distribution, is chosen. Based on this unique alignment searching algorithm, weakly conserved motifs can be identified with high probability scores.

FootPrinter (http://abstract.cs.washington.edu/~blanchem/FootPrinterWeb/FootPrinterInput2.pl) is a web-based program for phylogenetic footprinting using multiple input sequences. The user also needs to provide a phylogenetic tree that defines the evolutionary relationship of the input sequences. (One may obtain the tree information from the "Tree of Life" web site [http://tolweb.org/tree/phylogeny.html], which archives known phylogenetic trees using ribosomal RNAs as gene markers.) The program performs multiple alignment of the input sequences to identify conserved motifs. The motifs from organisms spanning over the widest evolutionary distances are identified as promoter or regulatory motifs. In other words, it identifies unusually well-conserved motifs across a set of orthologous sequences.

★ EM and Gibbs motif sampling -

- advanced alignment - independent profile construction method

EM = motif extraction algorithm
↳ finds motifs by repeatedly optimizing a PSSM through comparison with single sequences.

Gibbs sampling = uses a similar matrix optimizing approach
↳ samples motifs with a more flexible strategy
∴ may have ↑ likelihood of finding the optimal pattern.

Expression Profiling–Based Method

Recent advances in high throughput transcription profiling analysis, such as DNA microarray analysis (see Chapter 18) have allowed simultaneous monitoring of expression of hundreds or thousands of genes. Genes with similar expression profiles are considered coexpressed, which can be identified through a clustering approach (see Chapter 18). The basis for coexpression is thought to be due to common promoters and regulatory elements. If this assumption is valid, the upstream sequences of the coexpressed genes can be aligned together to reveal the common regulatory elements recognizable by specific transcription factors. ★

This approach is essentially experimentally based and appears to be robust for finding transcription factor binding sites. The problem is that the regulatory elements of coexpressed genes are usually short and weak. Their patterns are difficult to discern using simple multiple sequence alignment approaches. Therefore, an advanced alignment-independent profile construction method such as EM and Gibbs motif sampling (see Chapter 7) is often used in finding the subtle sequence motifs. As a reminder, EM is a motif extraction algorithm that finds motifs by repeatedly optimizing a PSSM through comparison with single sequences. Gibbs sampling uses a similar matrix optimization approach but samples motifs with a more flexible strategy and may have a higher likelihood of finding the optimal pattern. ★ Through matrix optimization, subtly conserved motifs can be detected from the background noise.

One of the drawbacks of this approach is that determination of the set of coexpressed genes depends on the clustering approaches, which are known to be error prone. That means that the quality of the input data may be questionable when functionally unrelated genes are often clustered together. In addition, the assumption that coexpressed genes have common regulatory elements is not always valid. Many coexpressed genes have been found to belong to parallel signaling pathways that are under the control of distinct regulatory mechanisms. Therefore, caution should always be exercised when using this method.

& 3 Drawbacks

The following lists a small selection of motif finders using the EM or Gibbs sampling approach.

MEME (http://meme.sdsc.edu/meme/website/meme-intro.html) is the EM-based program introduced in Chapter 7 for protein motif discovery but can also be used in DNA motif finding. The use is similar to that for protein sequences.

AlignACE (http://atlas.med.harvard.edu/cgi-bin/alignace.pl) is a web-based program using the Gibbs sampling algorithm to find common motifs. The program is optimized for DNA sequence motif extraction. It automatically determines the optimal number and lengths of motifs from the input sequences.

Melina (Motif Elucidator In Nucleotide sequence Assembly; http://melina.hgc.jp/) is a web-based program that runs four individual motif-finding algorithms – MEME, GIBBS sampling, CONSENSUS, and Coresearch – simultaneously. The user compares the results to determine the consensus of motifs predicted by all four prediction methods.

6/8/24

PAGE:
DATE: / /

Notes

Q] Unit 1 Elective

I] Gene, Introns & Exons

→ 1] Genes are functional unit of heredity as they are made up of DNA, the chromosome is made up of DNA containing many genes.

2] Every gene comprises of the particular set of instructions for a particular function or protein coding 3) There are about 30,000 gene in each cell of human body. Gene comprise of alternating pattern of introns & exons, promoter, open reading frame & splice sites. which collectively contributes for protein building

a] Introns:-

i) An Introns ~~are~~ is a region that resides within a gene but does not remain in the final mature mRNA molecule following transcription of that gene & thus not code for any aminoacid that make up the protein. ~~encoded~~

ii) The protein coding sequences contains both exons & introns wherein introns are non coding sequences whereas exons are coding sequences.

iii) Introns are removed during the process called splicing so only exons are included in the mature mRNA.

iv) Introns are much longer than exons.

v) Introns may contain sequences that regulate how genes are expressed ~~or~~ transcribed & how mRNA is processed.

Function:-

Transcription terminator

Genome organization

Transcription initiation

b] Exons:-

i) A part of gene encode for final mature RNA produced by that gene after introns have been removed by splicing

ii] exons usually include both the 5' & 3' untranslated regions of mRNA, which contain start & stop codons, in addition to any protein coding sequences

iii] There are 8.8 exons & 7.8 introns per gene

~~Function:~~

~~(1) Take Transcription process ahead~~

~~(2)~~

iv) Exons are coding sequences that code for a protein's aminoacid sequence

v) After post transcriptional alteration the exons are translated into mature mRNA

vi) There are highly conserved sequences meaning they don't change much over time

Functions of exons:-

1) Exons are the parts of a gene that code for a protein

2) exons are mRNA coding regions that code for aminoacids.

3) Various exons code for different protein domains.

4] A single exon or numerous exons spliced together can encode the domain

GeneAlign is a coding exon prediction [illegible] for Predicting Protein coding genes by measuring the homologies between a seq of a genome & related seq which have been annotated of other genome

5] When exons on sister chromosomes are switched during recombination, exon shuffling occurs.

6] - This enables creation of new genes

7] · Exons also allow for alternative splicing which allows several proteins to be translated from same gene

8] Introns are deleted from mature mRNA and exons are joined together

9] After introns have been eliminated by RNA splicing, an exon is any component of a gene that will constitute a part of the final mature RNA generated by that gene

Note:- The tool used for exon prediction is GeneAlig

Q] ORF Finder.

① In molecular genetics, an open reading frame is the part of a reading frame that has the ability to be translated. An ORF is a continuous stretch of codons that begins with a start codon and ends at a stop codon. An ATG codon within the ORF may indicate where translation starts. In other words we can say that the region of a nucleotide (can that starts from an initiation codon and ends with a stop codon is called ORF.

② The CDS (coding Sequence) is the actual region of DNA that is translated to form proteins while the ORF may contains introns as well. The CDS refers to those nucleotides (concatenated exons) that can be divided into codons which are actually translated into aminoacids by process of translation.

ORF finder

1) ORF finder is a program or graphical analysis tool available at NCBI website which searches for open reading frames (ORFs) in the DNA sequence you enter.

2) The program or tool returns the range of each ORF, along with its protein translation.

3) Use ORF finder to search newly sequenced DNA for potential protein encoding segments. This tool identifies all open reading frames using the standard or alternative genetic codes.

Importance of ORFs

① ORFs is a piece of evidence to assist in gene prediction

② Long ORF's are often used, along with other evidence to initially identify candidate protein coding regions or functional RNA coding regions in a DNA sequence.

Q Geneprediction.

1) Gene prediction by computational methods for finding the location of protein coding regions is one of the essential issues in bioinformatics

2) The gene sequencing of a gene is productive only when it is analysed & predicted correctly.

3) Gene prediction is carried out to identify the structure of genes in order to differentiate protein coding genes from non coding regions, and to identify promoters & other regulatory elements

4) Gene prediction basically means locating genes along genome. Also called genefinding. It refers to the process of identifying the regions of genomic DNA that encode genes.

5) This includes protein coding genes, RNA genes & other functional elements such as regulatory genes.

Importance of Gene Prediction

1) Helps to annotate large, contiguous sequences

2) It provides information on the evolution of genes, speciation & evolution of species.

3) It gives an understanding of the structure & function of genomes of different organism

4) Distinguish between coding & noncoding regions of a genome.

Types of approaches

-The current gene prediction method can be classified.

1) AB initio based ·(prediction based on given sequence only

a) The first feature is the existence of gene signals which include start & stop codons, introns, splice signals, transcription factor binding sites, ribosomal binding site.

b) Second feature used by abinitio algorithms is gene content which is statistical description of coding regions. It has been observed that nucleotide complementary & statistical pattern of coding region tend to vary significantly from non coding region. thus unique features can be detected by applying probabilistic models such as Markov model.

2] Homology based:

Predicts based on significant matches of the query sequence with sequence of known genes.

Eg:- If a translated DNA is found to be similar to a known protein family from a database search, this can be strong evidence that the region codes for a protein. Also if possible exons of a genomic DNA region match a sequenced cDNA. this also provides experimental evidence of a coding region

Promoters:

1) Promoter sequence are DNA sequences that define where transcription of a gene by RNA polymerase begins.

2) Promoter sequences are typically located directly upstream or at the 5' end of the transcription initiation site.

3) RNA polymerase and the necessary transcription factors bind: to the promoter sequence & initiate transcription

4) Promoter sequences define the direction of transcription and indicate which DNA strand will be transcribed, this strand is known as sense strand.

5) Many eukaryotic genes have a conserved promoter sequence called the TATA box, located 25 to 35 base pairs upstream of the transcription start site. Transcription factors bind to the TATA box and initiate the formation of the RNA polymerase transcription complex, which promotes transcription.

Splice sites

A genetic alteration in the DNA sequence that occurs at the boundary of an exon & intron known as splice site.

Most commonly RNA sequence that is removed

DATE: / /

begins with the dinucleotide GU nucleotide at its 5' end & ends with AG at its 3' end. The GU-AG rule originally called GT-AG rule interms of DNA seq.

Why splice sites important

Mutation in these sequence may lead to retention of large segments of intronic DNA by the mRNA or the entire exons being spliced out of the mRNA. these changes could result in production of a nonfunctional protein

a Prediction of splice sites were accurate localization of splicesites can substantially help to explore the structure of genes

Regulatory site

1) Regulator gene encodes for a DNA binding protein that act as a repressor.

2) A variety of DNA regulatory elements are involved in the regulation of gene expression & rely on the biochemical interactions involving DNA, the cellular proteins that make up chrom- & transcription factors. Promoters & enhancers are the primary genomic regulatory components of gene expression

Role of regulatory gene

Q Microarray technology

1) Microarray technology is a general laboratory approach that involves binding an array of thousands to million of known nucleic acid fragments to a solid surface, referred to as a "chip".

2) The chip is then bathed with DNA or RNA isolated from a study sample (such as cells or tis complementary base pairing between the sample and the chip immobilized fragments produces light through fluorescence that can be detected using a specialized machine.

3) Microarray technology can be used for a variety of purposes in research and clinical studies, such as measuring gene expression & detecting specific DNA sequences (eg single nucleotide polymorphisms, or SNPs).

4) Microarrays were revolutionary. They really allow genomic analysis without sequencing, which tremendously reduced the cost of doing large studies across a wide area of biology & biomedicine.

Goals:-

1) Gene expression or the amount of gene product, RNA, from any given gene that you found in a cell.

2) Single nucleotide polymorphisms or SNPs which were useful for genome wide association studies, or GWASS.

Application

1] Microarray is one of the most recent advances being used for cancer research it provides assistance in pharmacological approach to treat various diseases including oral lesions.

2] Microarray helps in analyzing large amount of samples which have either been recorded previously or new samples; it even helps to ~~the~~ test the incidence of a particular marker in tumor

3] Microarray provides a basis ~~of~~ to genotype thousands of different loci at time, which is useful for association & lineage studies to isolate chromosal region related to a particular disease

4) Genome microarrays have been used for comparative genomic hybridization. In this technique, genomic DNA is fluorescently labeled & used to determine the presence of gene loss or amplification

DATE / /

Note Continue

Genome Reference Sequence

① A human genome reference sequence is an accepted representation of the human genome sequence that is used by researchers as a standard for comparison to DNA sequences generated in their assembling and updating. Such reference sequences aim to provide the highest quality, best possible consensus representations of the sequence and structural diversity found in the human genome among populations.

② The genome reference sequence provides general framework and is not the DNA sequence of a single person.

③ Reference genome assemblies looked after by the Genome reference consortium.

Example

① Zebrafish genome reference assembly is a high quality clone based assembly that has gone through decade of continuous improvements. These improvements were based on a huge amount & variety of generated sequencing & mapping data & applied both through automated edits & manual interventions.

Tool used for Genome reference sequence

① Refseq - The reference sequence (Refseq) collection provides a comprehensive, integrated, non redundant, well annotated set of sequences, including genomic DNA, transcripts & proteins.

RefSeq sequences form a foundation for medical, functional & diversity studies. They provide a stable reference for genome annotation, gene identification & characterization mutations & polymorphism analysis.

Advantage

① The ability to characterize particular genes, or gene families, that are relevant to species specific conservation.

② The reference genome sequencing used for exploratory analysis of gene families involved in key biological process of threatened species such as immunity, reproduction & behaviour.

Integrated Genomic Maps (Intro).

① Genetic maps provide the outline & physical maps provide the details. It is easy to understand why both types of genome mapping techniques are important.

② Information obtained from each technique is used in combination to study the genome. Genome mapping is being used with different model organisms that are used for research.

③ Genome mapping is still an ongoing process and as more advanced techniques are developed, more advances are expected.

④ Mapping information generated in laboratories all over the world is entered into central databases, such as GenBank at the NCBI.

1) Physical maps.
2) Genetic maps

Gene Expression profiling

① The central dogma of biology describes the method by which information is taken from genes & used to create proteins. DNA transcription produces RNA, then RNA translation makes proteins. This process is known as gene expression and all life forms use it to create the building blocks of life from genetic information.

② A variety of DNA regulatory elements are involved in the regulation of gene expression and rely on the biochemical interactions involving DNA, the cellular proteins that make up chromatin & transcription factors. Promoters & Enhancers are the primary genomic regulatory components of gene expression.

③ Gene expression profiling measures which genes are being expressed in a cell at any given moment. This method can measure thousands of genes at a time, some experiments can measure the entire genome at once.

④ Gene expression profiling measures mRNA levels, showing the pattern of genes expressed by a cell at the transcription level. This often means measuring relative mRNA amounts in two or more experimental conditions, then assessing which conditions resulted in specific genes being expressed.

⑤ Different techniques are used to determine gene expression. These include DNA microarrays & sequencing technologies.

⑥ The former measures the activity of specific genes of interest and the latter enables researchers to determine all active genes in a cell.

⑦ A gene expression profile tells us how a cell is functioning at a specific time. This is because cell gene expression is influenced by external & internal stimuli including whether the cell is dividing, what factors are present in cell's environment, the signals it is receiving from other cells, & even the time of day.

Application

① Gene expression profiling used by a variety of biomedical researchers, from molecular biologists to environmental toxicologists. This technology can provide accurate information on gene expression, towards countless experimental goals.

② Gene expression profiling enables you to investigate the effects of different conditions on gene expression by altering the environment to which the cell is exposed & determing which genes are expressed.

③ Gene expression profiling is often used in hypothesis generation. If very little is known about when & why a gene will be expressed, expression profiling under

different conditions can help to design a hy
to test in future experiments.

④ Gene profiling can also investigate the effect of drug like molecules on cellular response. You could ~~determine~~ identify the gene markers of drug metabolism, or determine whether cells express genes known to be involved in responce to toxic environments when exposed to drug

⑤ Gene profiling can also be used as a diagnostic tool. If cancerous cells express higher levels of certain genes, and these genes code for protein receptor, this receptor may be involved in the cancer, and targeting it with a drug might treate the disease. Gene expression profiling might then be a key diagnostic tool for people with this cancer.

Q Orphan GPCR

① The superfamily of G-protein coupled receptors (GPCRs) include at least 800 seven transmembrane receptors that participate in diverse physiological & pathological functions

② GPCR represent the largest superfamily & most diverse grp of mammalian transmembrane proteins

③ GPCR play major role in numerous physiological & pathological roles in transducing extracellular signals into intracellular effector pathways through the activation of heterotrimeric G protein by binding to a broad range of ligands

(4) Human GPCRs can be divided into five main familia on the basis of phylogenetic criteria: Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2 & Secretin.

(5) The first GPCR to be identified was rhodopsin in 1878. It was later proven that rhodopsin consists of the GPCR protein opsin and a reversibly covalently bound cofactor, retinal.

(6) After completion of the human genome sequence in 2004, the number of human GPCRs increased to about 800 based on the screening approaches, such as low stringency hybridization, PCR derived methods & bioinformatic analyses.

(7) Besides the olfactory receptor family more than 140 GPCR's have not yet been linked to endogenous ligands. These are the so called orphan GPCRs

(8) These orphan GPCR represent vast opportunities for discovering new therapy for disease that has been intractable that targeting the well known GPCR and other protein family.

(9) Approach for identification of OGPCR

a) Screening of putative small molecule & peptide ligands

b) Reverse pharmacology

c) use of bioinformatics to predict candidate ligands

Q) Reverse Pharmacology

(1) Reverse Pharmacology is a science of integral documented experimental hits, into leads by transdisciplinary exploratory studies & further developing into drug candidates by experimental research.

(2) Reverse Pharmacology enhances the condition, communication & collaboration between modern science & technology with traditional medicine & modern biomedicine.

Drug discovery can be divided into two process (a) classical drug discovery (b) Reverse pharmacology

(3) Classical drug discovery process is an expensive & time consuming process whereas RP is an economical time & sparing & has least bottlenecks. It allows understanding the mechanism of drug action at multiple levels & helps in optimizing the safety, efficacy & acceptability of the leads from natural products

(4) RP utilizes traditional knowledge of medicines to discover drugs and is also called as a path of pharmacology from the bedside to bench experiments.

(5) The aim of forward pharmacology is to enhance the desired physiological effect of a compound'

(6) Forward pharmacology (phenotypic based screening) involves first identifying the functional (phenotype) activity of a compound through cellular or animal models.

(8) once knowing the physiological affect of the certain compound, only then the compound's ligand & its derivatives are

identified , purified & synthesized respectively & their binding capabilities with a target receptor are determined through biological assays / screenings.

⑦ The most potent & selective ligands was identified as the new possible drug & further research is done with ligand.

⑧ The ~~aimof~~ reverse pharmacology is to screen the ~~natural~~ natural products are screened against receptors / targets of known physiological function in order to determine functional activity.

⑨ The aim of reverse pharmacology (target based screening) began with the growth of molecular biology and caused a paradigm shift in drug discovery worldwide.

⑩ First potential ligands are screened through binding assay where the highly selective ligand that binds with the molecular target is identified. This is known as ligand fishing Then this potential ligand (compound) undergoes functional studies (animal models) to significantly show the desired physiological effect

There are three methods in Reverse Pharmacology

① library based approach - Used in pharmaceutic & some large labs have

② Tissue based extract - Used in labscale basis

③ Information based
- Identifies prospective ligands by database screening & testing them.

Deorphanization:-

① GPCRs are the most imporiant & prominent family of pharmacological targets in biomedicine

② The deorphanization of orphan GPCRs is the important mission in orphan GPCR research

③ It is a process of ~~identifying~~ indentification of ligands that are highly selective for orphan GPCR.

④ In general the standard assay are radio-ligand binding, calcium flux, GTPx binding & modulation of CAMP levels 92, 93, 94, 95, 96, 97 & 98.

⑤ with the development of molecular biology technologies serveral lines of approach have been used for deorphanization. i.e one of them is according to the sequence & functional similarity, ligand of the identified receptors & are used to examine GPCR with identical sequence or domain

Q) Synteny & Geneorder

1) In genetics the term synteny refers to two related concepts

a) In classical genetics synteny describes the physical co-localization of genetic loci on the same chromosome with an individual or species

b) In current biology synteny more commonly refers to collinearity. ie. conservation of blocks of order within two sets of chromosomes that are being compared with each other. These blocks are referred to as syntenic blocks.

Uses :-

1) Provides a framework in which conservation of homologous genes & geneorder is identified between genome of different species. The availability of human & mouse genomes paired the way of algorithm development in large scale based on synteny mapping

2) Comparing two genomes reveal homologous sequences that reflect their evolutionary origin & subsequent conservation

Gene order is much less conserved with gene sequence. Therefore syntenic relationships are normally carried out between relatively close lineages

Gene order are the permutations of genome arrangement. A fair amount of research has been done trying to determine whether gene order evolve according to a molecular clock hypothesis or in jumps (punctuated equilibrium).

Some research on gene order in animals mitochondria genome reveals that the mutation rate of gene order is not a constant in some degree

Q] Edman's degradation.

(1) Edman's degradation is the process of purifying protein by sequentially removing one residue at a time from the amino acid end of peptide.

(a) To solve the problem of damaging the protein by hydrolysing condition.

(2) Pehr Edman created a new way of labelling & cleaving the peptide, where phenyl isothiocynate was added this compound creates a phenyl thiocarbonyl derivate with N-terminal.

(3) The N-terminal is then cleaved under less harsh conditions creating a cyclic compound of phenthiohydantoin, PTH aminoacid. This does not damage the ~~pep~~ protein & leaves two constituents of the peptide.

(4) Edman sequencing is done but if composition of aa is known

Advantages

(1) the whole sequencing of the protein can be done without damaging the protein.

(2) Allows sequencing of protein in less time.

Q) Shotgun proteomics:

1) Shotgun proteomics also known as 'Bottom-up proteomics' is a widely used & mature technology for protein identification & characterization of their aminoacid sequence along with posttranslational modifications (PTMs).

2) This technique requires the proteolytic digestion of proteins prior to mass spectrometry analysis.

3) Shotgun proteomics has been demonstrated to be a valuable tool for the identification of novel large or small proteomes & protein complexes, enabling the discovery of previously unknown protein-protein interactions.

4) The most distinctive feature of shotgun proteomics is that it enables identify a wide range of proteins at the same time with minimal protein separation needed.

5) Shotgun proteomics refers to the use of bottom up proteomics techniques in identifying protein in complex mixtures using a combination of high performance liquid chromatography combined with mass spectrometry.

(6) The name is derived from shotgun sequencing of DNA which is itself named after rapid, expanding, quasi random firing pattern of shotgun.

(7) Targeted proteomics using SRM & data independent acquistion method are often considered alternative to shotgun proteomics in the field of bottomup proteomics

Advantages

1) Shotgun proteomics allows global protein identification as well as the ability to systematically profile dynamic proteome.

2) It also avoids the modest separation efficiency & poor mass spectral sensitivity associated with intact protein analysis.

Disadvantages

(1) ~~Most~~ The dynamic exclusion filtering that is often used in shotgun proteomics maximizes the number of identified proteins at the expense of random sampling.

(2) This problem may be generated by the undersampling inherent in shotgun proteomics

Applications

(1) Shotgun proteomics can be used for functional classification or comparative analysis of their protein products

(2) It can be used in projects ranging from large scale whole proteome to focusing on a single protein family. It can be done in research labs or commercially

Q] Protein identification with antibody

→① Antibodies are proteins synthesized by a animal in response to the presence of a foreign substance known as antigen

② This antibody have specific affinity for a particular region of antigen

③ This region is termed as Epitope.

④ The antibody epitope interaction can be utilized for highly specific & sensitive detection of a protein that has been immobilized on a membrane; in a process termed as immunodetection.

⑤ The antibody that binds to the protein of interest is termed as the primary antibody

⑥ The primary is applied to the membran & it is allowed to bind to the target protein in order to locate the primary antibody & the protein of interest, a secondary antibody is required.

⑦ The secondary antibody recognize & binds to all IgG antibodies because IgG antibody react specifically with the introduced protein & can be harvested from animal serum.

⑧ It is important that secondary antibody used in a experiment is directed against IgG from the species of origin of the primary antibody

Antibody based Protein detection techniques

① ELISA-

Enzyme linked Immunosorbent Assay is a method that is analogous to Immuno detection of proteins on a membrane & it is used for quantitative assay of proteins in a sample.

② In ELISA ; proteins are immobilized onto a solid support known, as well plate this step is known as fixation.

③ A wash is given to the plates containing the proteins so that to remove the non specifically bound material.

④ A secondary antibody which will be specific for this protein will be added.

⑤ This secondary antibody ~~which will be~~ ~~or~~ is usually conjugated to an enzyme, that allows its detection by the chromogenic or chemiluminescent methods.

Advantages

① High sensitive & specific can detect antigens at the picogram level in a very specific manner due to the use of antibody

② High throughput

③ Accessibility to test various sample types.

Disadvantage

① Temporary readout - Detection is based on enzyme substrate reaction & therefore readout must be obtained in a short span of time,

② limited antigen information

(1) Westurn blotting

1) It is a laboratory technique used to detect a specific protein in a blood or tissue sample.

2) The method involves the use of gel electrophoresis to separate the sample protein

3) The separated proteins are transferred out of the gel to the surface of the membrane

4) The membrane is exposed to an antibody specific to the target protein

5] Binding of the antibody is detected using a radioactive or chemical

Principle :- The sample is separated by using electrophoresis. gel
The proteins are then resolved & transferred onto a membrane of special paper. The membrane is then probed with an antibody specific protein of interest. Because the antibody is labeled with a molecule that we can visualize
We can all whether the protein of interest is expressed in this sample or not & can also know the abundancy of the protein

3784–3788 Nucleic Acids Research, 2003, Vol. 31, No. 13
DOI: 10.1093/nar/gkg563

ExPASy: the proteomics server for in-depth protein knowledge and analysis

Elisabeth Gasteiger*, Alexandre Gattiker, Christine Hoogland, Ivan Ivanyi, Ron D. Appel and Amos Bairoch

Swiss Institute of Bioinformatics, Centre Médical Universitaire, 1 Rue Michel Servet, 1211 Geneva 4, Switzerland

Received February 18, 2003; Revised and Accepted April 2, 2003

ABSTRACT

The ExPASy (the Expert Protein Analysis System) World Wide Web server (http://www.expasy.org), is provided as a service to the life science community by a multidisciplinary team at the Swiss Institute of Bioinformatics (SIB). It provides access to a variety of databases and analytical tools dedicated to proteins and proteomics. ExPASy databases include SWISS-PROT and TrEMBL, SWISS-2DPAGE, PROSITE, ENZYME and the SWISS-MODEL repository. Analysis tools are available for specific tasks relevant to proteomics, similarity searches, pattern and profile searches, post-translational modification prediction, topology prediction, primary, secondary and tertiary structure analysis and sequence alignment. These databases and tools are tightly interlinked: a special emphasis is placed on integration of database entries with related resources developed at the SIB and elsewhere, and the proteomics tools have been designed to read the annotations in SWISS-PROT in order to enhance their predictions. ExPASy started to operate in 1993, as the first WWW server in the field of life sciences. In addition to the main site in Switzerland, seven mirror sites in different continents currently serve the user community.

INTRODUCTION

The Swiss Institute of Bioinformatics (SIB, http://www.isb-sib.ch) is an academic not-for-profit foundation whose mission is to promote research, the development of databanks and computer technologies, teaching and service activities in the field of bioinformatics. One of the SIB's windows to the world is the ExPASy server, which focuses on proteins and proteomics, and provides access to a variety of databases and analysis tools. One of the major assets of ExPASy is the high degree of integration and interconnectivity that it establishes between all the available databases and services. Rather than just making each service accessible in an isolated manner, we put at the disposal of the users different expert views of the complex world of biological data and knowledge.

DATABASES

ExPASy (1,2) is the main host for the following databases that are partially or completely developed at the SIB in Geneva:

- The SWISS-PROT knowledgebase (3,4) (http://www.expasy.org/sprot/) is a curated protein sequence database, which strives to provide high quality annotations (such as the description of the function of a protein, its domain structure, post-translational modifications and variants), a minimal level of redundancy and a high level of integration with other databases. SWISS-PROT is supplemented by TrEMBL, which contains computer-annotated entries for all sequences not yet integrated in SWISS-PROT. SWISS-PROT and TrEMBL are maintained collaboratively by the SIB and the European Bioinformatics Institute (EBI).
- SWISS-2DPAGE (5) (http://www.expasy.org/ch2d/) is a database of proteins identified on two-dimensional polyacrylamide gel electrophoresis (2D PAGE). SWISS-2DPAGE contains data from a variety of human and mouse biological samples as well as from *Arabidopsis thaliana*, *Escherichia coli*, *Saccharomyces cerevisiae* and *Dictyostelium discoideum*.
- PROSITE (6,7) (http://www.expasy.org/prosite/) is a database of protein domains and families. PROSITE contains biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs.
- ENZYME (8) (http://www.expasy.org/enzyme/) is a repository of information relative to the nomenclature of enzymes.
- SWISS-MODEL Repository (9) (http://www.expasy.org/swissmod/smrep.html) is a database of automatically generated structural protein models.

Cross-references

All the databases available on ExPASy are extensively cross-referenced to other molecular biology databases or resources all over the world. SWISS-PROT for example is

*To whom correspondence should be addressed. Tel: +41 22 379 5050; Fax: +41 22 379 5858; Email: elisabeth.gasteiger@isb-sib.ch explicitly cross-referenced (10) to ~50 different databases specializing in protein and nucleic acid sequences, 3D-structure, organism-specific and genomic information, domain and family signatures, post-translational modifications or proteomics data. Examples for databases currently linked to SWISS-PROT in that manner are EMBL/GenBank/DDBJ, PDB, FlyBase, MGD, MIM, MypuList, SGD, SubtiList, TubercuList, WormPep, ZFIN, InterPro, Pfam, PRINTS, ProDom, PROSITE, SMART, TIGRFAMs, SWISS-2DPAGE, HSSP, MEROPS and REBASE. On average, a SWISS-PROT entry contains 7.8 explicit cross-references to other databases (release 40.43 of 12 February 2003). Literature references for the above-mentioned databases are listed in the SWISS-PROT user manual, (http://www.expasy.org/sprot/userman.html#DR_line).

Complementing these explicit cross-references, so-called 'implicit links' to ~25 additional resources are created on-the-fly by the NiceProt view of SWISS-PROT and TrEMBL entries (see below). This concept is targeted at data collections that do not have their own system of unique identifiers, but can be referenced via identifiers such as SWISS-PROT or EMBL accession numbers, gene names, etc. Examples for databases linked to SWISS-PROT via implicit links are those that are based on SWISS-PROT and provide a specific analytical view of each entry (e.g. ProDom—automatically derived domain views or ProtoMap—a hierarchical classification of all SWISS-PROT entries) and those databases that share some identifier with SWISS-PROT (e.g. GeneCards—information on human genes, accessible by the HUGO approved gene name). Implicit links are a specific feature of ExPASy and are not available on other web servers, or in the SWISS-PROT/TrEMBL data files that can be downloaded by ftp. They greatly enhance database interoperability and strengthen the role of SWISS-PROT as a central hub for the interconnection of biomolecular resources.

Update frequency and download options

SWISS-PROT, PROSITE, ENZYME and SWISS-2DPAGE are updated at a frequency of ~1–2 weeks.

For all the ExPASy databases, data and associated documentation files can be copied locally by anonymous FTP (ftp.expasy.org). In particular, the different download options for the SWISS-PROT and TrEMBL databases, including the different available subsections, release frequencies and data formats, are documented at http://www.expasy.org/sprot/download.html. Among others, we distribute the files to assemble a non-redundant and complete protein sequence database (ftp://ftp.expasy.org/databases/sp_tr_nrdb/) consisting of three components: SWISS-PROT, TrEMBL and new entries to be later integrated into TrEMBL (known as TrEMBLnew). These files are supplemented by a compilation of sequences for splice variants, reconstructed from the annotations in SWISS-PROT and TrEMBL feature tables. All these files are completely rebuilt every time SWISS-PROT is updated.

A large variety of documents (user manual, release notes, indices, nomenclature documents, etc.) are available with SWISS-PROT; these documents can all be browsed from ExPASy (http://www.expasy.org/sprot/sp-docu.html) and are enhanced by a variety of hyperlinks.

No fees for academic users

The use of all ExPASy databases is free for academic users. However, we implemented in September 1998 a system of annual subscription fee for commercial users of the SWISS-PROT, PROSITE and SWISS-2DPAGE databases. The funds raised are used to bring these databases up-to-date, to keep them up-to-date and to further enhance their quality. Further information on this funding scheme is available at http://www.expasy.org/announce/.

SOFTWARE TOOLS

We have developed, over the years, an extensive collection of software tools, most of which are either targeted toward the access and display of the databases mentioned above, or can be used to analyze protein sequences and proteomics data originating from 2D-PAGE and mass spectrometry experiments. These latter tools can all be accessed from ExPASy (http://www.expasy.org/tools/).

Database query, display and navigation

A variety of query options are available from the home pages of each of the ExPASy databases. These options allow the users to display and retrieve specified subsets of the database. For example, from the home page of SWISS-PROT and TrEMBL, different query forms allow searching by description, accession number, author, citation or by full text search. To complement these options, we have also implemented an SRS (11) server that allows complex searches on any fields of the combination of SWISS-PROT and TrEMBL databases. PROSITE, ENZYME and SWISS-2DPAGE can also be queried using SRS.

The original flat file format of all ExPASy databases is based on different line types, where a two-letter line code defines the information contained on the rest of that line (e.g. for SWISS-PROT: see the user manual, http://www.expasy.org/sprot/userman.html). This format is easy to parse by computer programs, but not necessarily easy to read for human users. In order to provide a more verbose and user-friendly view of the database entries, we provide for each database, on ExPASy, a 'nice' hypertext view, e.g. NiceProt for SWISS-PROT and TrEMBL entries. An example for an entry in the NiceProt view can be seen at http://www.expasy.org/cgi-bin/niceprot.pl?P57727, or in Figure 1. The figure shows parts of that entry in order to illustrate the easy navigation between information contained in the entry itself, the corresponding documentation, remote databases, and the submission forms or results of sequence alignment or other ExPASy analysis tools. Similar views are available for PROSITE (NiceSite and NiceDoc), ENZYME (NiceZyme) and SWISS-2DPAGE (Nice2Dpage).

Swiss-Shop (http://www.expasy.org/swiss-shop/) is an automated sequence alerting system which allows users to obtain new SWISS-PROT entries relevant to their field(s) of interest. Keyword-based and sequence/pattern-based requests are possible. Every time a weekly SWISS-PROT release is performed, all new database entries matching the user-specified search keywords or patterns or the entries showing sequence similarities to the user-specified sequence are automatically sent to the user by email.

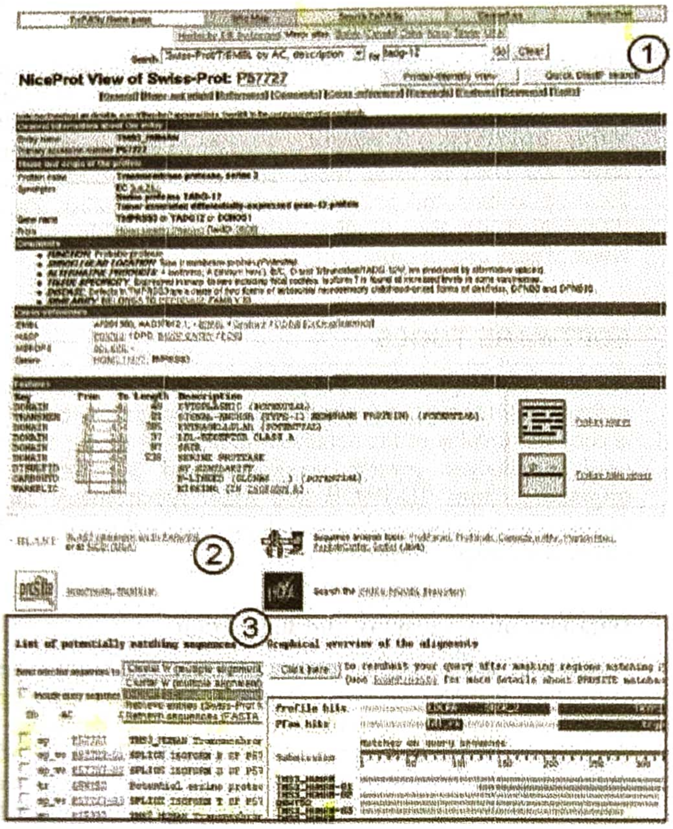

NiceProt View of Swiss-Prot: P57727
1
2
3
List of potentially matching sequences
Graphical overview of the alignments
Profile hits
Pfam hits

Figure 1. The NiceProt view of a SWISS-PROT entry presents its contents in a user-friendly view. Links are provided to >70 databases, a user manual and other documents. NiceProt is also integrated with tools provided on ExPASy and other servers. Excerpts of the view of a sample entry are presented in this figure. A BLASTP similarity search against SWISS-PROT/TrEMBL/TrEMBLnew can be performed with a single click (button 1) on a very fast server (median request time: 6 s). BLAST parameters can also be adjusted by accessing the BLAST page (link 2), which provides a choice of BLASTP or TBLASTN over a choice of databases and subsections. The result page (inset 3) combines a BLASTP search with a motif search in the PROSITE profiles and Pfam HMM domain databases and displays a graphical overview of the matching regions both on the query and on the hit. From there, it is possible to browse matching entries, including splice variants; to rerun a BLAST search after masking regions that match PROSITE or Pfam domains in order to find weaker similarities in other regions and to perform multiple alignments of selected hit sequences.

Sequence analysis tools

- BLAST (12) provides very fast similarity searches of a protein sequence against a protein or nucleotide database. The ExPASy BLAST service is maintained in collaboration with the Swiss EMBnet node on dedicated hardware. The native output of BLAST is extended with several original features (Fig. 1).
- ScanProsite (13) scans a sequence against all the patterns, profiles and rules in PROSITE or scans a pattern, profile or rule against all sequences in SWISS-PROT, TrEMBL and/or PDB.
- SWISS-MODEL (14,15) is an automated knowledge-based protein modelling server. It is able to build models for the 3D structure of proteins whose sequence is closely related to that of proteins with known 3D structure.
- ProtParam calculates physico-chemical parameters of a protein sequence such as the amino acid composition, the pI, the atomic composition, the extinction coefficient, etc.
- ProtScale computes and represents the profile produced by any amino acid scale on a selected protein. Some 50 predefined scales are available, such as the Doolittle and Kyte hydrophobicity scale.
- RandSeq generates a random protein sequence, based on a user-specified amino acid composition and sequence length.
- Sulfinator (16) predicts tyrosine sulfation sites within protein sequences.
- Translate translates a nucleotide sequence into a protein in six reading frames.

Proteomics tools

- AACompIdent (17) identifies a protein by its amino acid composition.
- AACompSim (17) finds for a given SWISS-PROT entry, the database entries which have the most similar amino acid composition.
- Compute pI/MW (18) computes the theoretical isoelectric point (pI) and molecular weight (MW) from a SWISS-PROT or TrEMBL entry or for a user sequence.
- FindMod (19) predicts potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified SWISS-PROT entry or from a user-entered sequence. Mass differences are used to better characterize the protein of interest.
- FindPept (20) identifies peptides resulting from unspecific cleavage of proteins by their experimental masses, taking into account artefactual chemical modifications, post-translational modifications and protease autolytic cleavage.
- GlycanMass calculates the mass of an oligosaccharide structure.
- GlycoMod (21) predicts possible oligosaccharide structures that occur on proteins from their experimentally determined masses. This is done by comparing the mass of a potential glycan to a list of pre-computed masses of glycan compositions.
- PeptideCutter predicts potential protease cleavage sites and sites cleaved by chemicals in a given protein sequence.
- PeptideMass (22) calculates the theoretical masses of peptides generated by the chemical or enzymatic cleavage of proteins so as to assist in the interpretation of peptide mass fingerprinting.
- PeptIdent, TagIdent, MultiIdent (23–25), these three related programs identify proteins using a variety of experimental information such as the pI, the MW, the amino acid composition, partial sequence tags and peptide mass fingerprinting data.

A very important feature of the ExPASy proteomics tools (such as PeptIdent, TagIdent, MultiIdent, PeptideMass, FindPept or FindMod) is that, when performing their computations and predictions, they use the annotations relevant to post-translational modifications and processing, as well as splice variants documented in the SWISS-PROT feature tables.

These tools are all listed on a page on ExPASy (http://www.expasy.org/tools/) that also offers links to many other useful programs for the analysis of protein sequences available elsewhere on the web. We notably have links to the tools provided by our colleagues from the bioinformatics group at ISREC (http://www.isrec.isb-sib.ch) and the Swiss EMBnet node (http://www.ch.embnet.org) in Lausanne. They have developed a BLAST similarity search server, TMpred (to predict transmembrane regions) and interfaces to the SAPS (Statistical Analysis of Protein Sequences), COILS (prediction of coiled coil regions), Clustal and T-Coffee (multiple sequence alignment) programs.

ExPASy AS A PORTAL TO OTHER LIFE SCIENCE RESOURCES

The mass of information available to life scientists on the web has completely changed the way in which biological data is accessed and processed. It has created many opportunities, but also brought new dangers. One of the most critical problems is the difficulty for researchers to distinguish useful and up-to-date sources of information from sites that provide either 'fossilized' or low-quality data. To partially address this problem, we have developed a series of lists and tools:

- Amos' WWW links page (http://www.expasy.org/alinks.html) is a list that contains links to >1000 information resources for the life sciences. This list is updated very frequently and is organized in a number of sections that correspond to specific topics.
- WORLD-2DPAGE (http://www.expasy.org/ch2d/2d-index.html) is a list of all known 2D PAGE database WWW servers and related services.
- BioHunt (http://www.expasy.org/BioHunt/) is a service to help search the internet for molecular biology information. BioHunt is built by Marvin, a software robot which automatically roams the web to search and index life science and bioinformatics information. Currently BioHunt indexes ~35 000 documents.
- 2DHunt (http://www.expasy.org/ch2d/2DHunt/) is a specialized index for 2D PAGE-related sites.
- ExPASy tools page (http://www.expasy.org/tools/), in addition to hosting the above-mentioned tools provided and maintained by the Swiss Institute of Bioinformatics, the tools page serves as a portal to useful web-accessible tools on bioinformatics servers elsewhere. Tools local to the ExPASy server are marked by the ExPASy logo.
- List of conferences and events (http://www.expasy.org/conf.html) is a list of conferences and meetings relevant to proteomics, bioinformatics and other domains in the life sciences.

OTHER INTERESTING ExPASy FEATURES

- Biochemical pathways (http://www.expasy.org/tools/pathways/) is an indexed, digitized and clickable version of the Boehringer Mannheim's 'Biochemical Pathways' poster and is available on the server. It allows the user to navigate through the graphical representation of metabolic pathways and is linked to the ENZYME database.
- DeepView (SWISS-PdbViewer) (15) (http://www.expasy.org/spdbv/) is an application running on the Microsoft Windows, Mac, SGI and Linux platforms, offering a wide range of options to visualize and manipulate protein structures. It can also be used as a WWW helper application for the display of PDB formatted entries. Swiss-PdbViewer can be downloaded from ExPASy and complements the aforementioned SWISS-MODEL homology-modeling tool.
- LALNVIEW (26) (http://www.expasy.org/tools/lalnview.html) is an application that runs on the Microsoft Windows, Mac and Unix platforms. LALNVIEW is a graphical viewer for pairwise sequence alignments. It can be used to display the results of a pairwise alignment carried out with the SIM (27) software also installed on ExPASy (http://www.expasy.org/tools/sim-prot.html).
- 2D PAGE: a wide variety of information concerning 2D PAGE is available from ExPASy. This includes the full description of experimental protocols as well as an overview of the Melanie 3 2D PAGE analysis software package. A 2D gel viewer is also available for download.
- Protein Spotlight (http://www.expasy.org/spotlight/) is a periodical review centered on a specific protein or group of proteins.
- Recreational. One must not forget that science can also have a lighter side. So we hope that users will take the time to take a small pause from the hectic pace of modern research and visit Swiss-Quiz (http://www.expasy.org/swiss-quiz/). With Swiss-Quiz one can have a chance to win some Swiss chocolate (real, not virtual!) after having successfully answered a quiz from the field of molecular biology.
- ExPASyBar is a useful navigation bar to the most important databases and tools on ExPASy. ExPASyBar was developed by Martin Hassman from the Institute of Chemical Technology in Prague, in collaboration with the ExPASy team. It is an add-on to the free Mozilla web browser (http://www.mozilla.org), and can be downloaded from http://expasybar.mozdev.org.

MIRROR SITES

Network congestion and resulting slow response times represent a major problem for users in certain parts of the world. To help address this issue, we decided to implement mirror sites of ExPASy in various countries. Such sites can help users to access the ExPASy databases and tools more rapidly in locations that do not have a fast connection to Switzerland. The mirror sites are computers that host exact copies of the information available from the Geneva ExPASy server. They are updated at the same frequency as the main ExPASy site in Switzerland. ExPASy mirror sites are located in academic institutions that have shown an active interest in hosting such sites. As of today, seven sites are operational. The ExPASy mirror sites are located in:

1. Australia: http://au.expasy.org/ at the Australian Proteome Analysis Facility (APAF), Sydney.
2. Bolivia: http://bo.expasy.org/ at the Universidad Católica Boliviana (UCB), Cochabamba.