Gene Identification *in silico*

Nita Parekh, IIIT Hyderabad

Presented at ***National Seminar on Bioinformatics and Functional Genomics***, at Bioinformatics centre, Pondicherry University, Feb 15 – 17, 2006.

Introduction

The ultimate goal of molecular cell biology is to understand the physiology of living cells in terms of the information that is encoded in the genome of the cell – and we would like to address the question how computational biology can help in achieving this goal. Genes coding for proteins is a very important pattern recognition problem in bioinformatics and this lecture focuses on some computational issues in gene identification.

DNA is the blueprint of the cell and it contains the complete information for the working and the reproduction of the cell. It is composed of four basic units called nucleotides. Each nucleotide contains sugar, phosphate and one of the four bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). For all computational purposes one may consider a DNA sequence just a string of four alphabets A, T, G, C and try to find grammar rules followed by these alphabets in genomes, similar to Wren and Martin for English grammar rules to form meaningful English sentences.

General Features of a Genome:

The order of occurrence of four alphabets in a DNA sequence is not completely random (else the percentage of occurrence of each alphabet would be 0.25).

Different regions of the genome exhibit different patterns of these alphabets, A, T, G, C, e.g., protein coding regions, regulatory regions, which govern the production of proteins and enzymes, regulatory regions, repeat regions, intron/exon boundaries, etc.

The functional role of a large part of the DNA sequence is unknown, such as intragenic regions, repeat sequence regions, etc.

The analysis of DNA sequences involves identifying the various patterns and understanding their functional roles.

Pattern recognition in DNA Sequences:

The basic underlying assumption for pattern recognition in biological sequences is that - strings carrying information will be different from random strings, which have no information. So if a hidden pattern can be identified in a string, it must be carrying information

This task needs to be automated because of the large sizes of the genome. To have an estimate of the size of the DNA strings for computational analysis, it may be noted that the species with the smallest genome, Mycoplasma genitalium – a parasite genome, originally isolated from urethral specimens of patients is about 580Kb. The human genome is about $3x10^9$ bps long.

Some important examples of pattern recognition in DNA sequences are:

- Identifying Genes - regions that code for proteins, Exons
- Identifying Signals - Promoters, Enhancers, Start & Stop Codons, Donor & Acceptor Sites, Motifs, CpG islands

Gene Prediction:

What is Computational Gene Finding? Given an uncharacterized DNA sequence, we would like to find out:

Which region codes for a protein?
Which DNA strand is used to encode the gene?
Which reading frame is used in that strand?
Where does the gene starts and ends?
Where are the exon-intron boundaries in eukaryotic genes?
Where are the regulatory sequences for that gene?

The search space is the whole genome (~ 100 – 1000 Mbp), and only 2-5% of it codes for proteins in eukaryotes.

Importance:

- It is the first step towards getting at the function of a protein.
- It also helps accelerate the annotation of genomes.

The various approaches used in gene prediction are:

- Finding Open Reading Frames (ORFs)
- Homology search (involves pair-wise alignment)
- Content-based methods: Ab initio methods based on statistics, nucleotide distribution, periodicity in base occurrence, their dependencies on the characters preceding it (i.e., how often an A is followed by a C, etc.), frequency of occurrence of codons (triplets), di-codons (hexamers), amino acids, etc.
- Signal-based methods: look for signals in the vicinity of coding region, viz., CpG islands, promoter sequences, translational signals, poly-A signal, splice sites, etc.
- Integration of these methods

<u>Finding genes in Prokaryotes</u>

Gene discovery in prokaryotic genomes is quite different problem from that encountered in eukaryotic sequences, owing to the higher gene density typical of prokaryotes and the absence of introns in their protein coding genes. These properties imply that most open reading frames (ORFs) encountered in a prokaryotic sequence that are longer than some

reasonable threshold will likely correspond to genes. The primary difficulties arising in this simple approach are

- Frameshift errors, a single insertion/deletion can lead to completely different aminoacid
- very small genes will be missed
- the occurrence of overlapping long ORFs on opposite DNA strands (genes and 'shadow genes') often leads to ambiguities.

Finding genes in Eukaryotes:

Homology based:

One can do a homology search to identify genes, by looking for similar genes in the public databases. In gene finding, sequence similarity can be used in at least six different ways:

1. A direct comparison of a genomic sequence with databases of expressed sequence tags (ESTs), using programs such as BLASTN and AAT
2. Comparison of a genomic sequence that is translated in all six reading frames with protein sequence databases, using programs such as BLASTX.
3. 'Spliced alignment' of a genomic sequence containing a single complete gene with a homologous protein sequence, using PROCRUSTES, may enable reconstruction of the exon and intron organization of the gene.
4. Comparison of predicted peptides, derived from programs such as GENSCAN or FGENEH, with protein sequence databases can be used to confirm predictions and/or to assign putative function to predicted proteins.
5. A comparison of a translated genomic sequence with a translated genomic or cDNA sequence using TBLASTX can identify similarities among coding regions.
6. Comparison of a genomic sequences with homologous genomic sequences from closely related organisms (e.g., human vs mouse), using BLAST and multiple alignment programs such as CLUSTAL W, to identify conserved regions, which often correspond to coding exons or important transcriptional or splicing signals.

Each of these methods can provide useful information about gene locations, as well as clues to gene function, although similarity-based methods are able to identify only about half of all human genes, and this proportion is increasing rather slowly. It should always be kept in mind that similarity-based methods are only as reliable as the databases that are searched, and apparent homology can be misleading at times.

Ab initio methods

The gene identification methods can be classified into two categories: content-based and signal-based methods.

Content-based methods - Content-based methods calculate statistics that distinguish genes from non-coding DNA. Gene sequences have constraints because of which the

succession of nucleotides cannot be random. This gets reflected in their base composition, frequency of different combinations of nucleotides and correlations between nucleotides. Coding measure therefore is a number or a list of numbers (called vector) associated with a sequence defining attributes correlated with protein coding function. Example of some coding measures used are: codon usage vector which is the frequency of usage of each of the possible 64 codons, ORF measure involves looking for longest stretch of sense codons, composition measure is the frequency of each nucleotide in each of the codon positions, Fourier measure looks at correlations between letters.

Signal-based methods

The most natural way to find genes computationally would be to mimic as closely as possible the processes of transcription and RNA processing (e.g., splicing and polyadenylation) that define genes biologically. Although this direct approach to gene finding is not yet feasible, a number of important signals related to transcription, translation and splicing are now sufficiently well characterized as to be useful in computer predictions of the location and intro-exon organization of genes.

Signal-based methods look for short sequences that are almost invariably found in and around protein coding region. These signals represent binding sites of molecules involved in gene transcription process, in post-transcriptional modifications, etc. This is perhaps the way the gene expression machinery of the cell recognizes genes. Examples of signals include promoter sequences, representing the binding sites for DNA polymerase, which include a TATA box, cap-signal, CCAAT box and GC box upstream of the gene. Translational signals, initiation codon ATG, the 'Kozak signal' located immediately upstream of initial ATG, the three stop codons, poly-A signal (AATAAA) downstream of genes, and splice site consensus (introns begin with GT and end with AG) by looking at the most frequently occurring nucleotide at each position of the motif, and look for the consensus sequence. A more sensitive search method is to define the frequency of nucleotides at each position, and use this frequency of distribution to define a position weight matrix. The position weight matrix is used to assess the potential of sequence to be a signal.

The problem of gene identification is complicated in case of eukaryotes by the vast variation that is found in the structure of genes. In eukaryotes, the coding region is usually discontinuous and composed of alternating stretches of exons and introns. Exons are the regions that code the information required for protein synthesis, while introns are the non-coding regions. On an average, a vertebrate gene is 30Kb long. Of this, the coding region is only about 1Kb. The coding region typically consists of 6 exons, each about 150bp long. These are average statistics. Huge variations from the average are observed. For e.g., a gene called *dystrophin* is 2.4Mb long. *Blood coagulation-factor VIII* has 26 exons with sizes varying from 69 bp to 3106 bp. The total coding region consists of around 186Kb. Its introns are up to 32.4Kb long. Intron number 22 produces 2 transcripts unrelated to this gene, one from each strand. An average 5' UTR is 750bp long, but it can be longer and span several exons (for e.g., in MAGE family). On an

average, the 3' UTR is about 450bp long, but example exist where its length exceeds 4Kb (e.g., the gene for Kallman's syndrome).

In higher eukaryotes the gene finding becomes far more difficult because it is now necessary to combine multiple ORFs to obtain a spliced coding region. Alternative splicing is not uncommon, exons can be very short, and introns can be very long. Furthermore, sequence features are less conserved and more spread out, reflecting the complexity of regulatory mechanisms and the diversity of interacting molecules. Given the nature of genomic sequence in humans, where large introns are known to exist, we recognize the need for highly specific gene finding algorithms.

Open Problems and Future Directions

Existing gene-finding programs have several important limitations.

1. Most programs only predict protein coding genes and not genes whose products function exclusively at the RNA level.
2. No current method can deal effectively with overlapping genes in eukaryotes, and prediction of multiple genes in a sequence is still difficult.
3. The problem of multiple protein products that correspond to a single gene through alternative splicing, alternative transcription, and/or alternative translation has not yet been dealt with effectively, although some current gene-finding programs are able to predict sets of alternative exons or genes. The rules governing alternative exon and intron choice are not well understood, presenting significant challenges to both experimental and computational biologists.

Some Important Resources:

National Center for Biotechnology Information (NCBI) - a national resource for molecular biology information (http://www.ncbi.nlm.nih.gov/)

The European Bioinformatics Institute (EBI) – a non-profit academic organization that forms part of the European Molecular Biology Laboratory (EMBL) (http://www.ebi.ac.uk/)

The National Human Genome Research Institute (NHGRI) - led the Human Genome Project for the National Institutes of Health, which culminated in the completion of the full human genome sequence in April 2003 (http://www.genome.gov/)