FACULTY OF ENGINEERING AND TECHNOLOGY

Department of Biotechnology

Basis of Gene Prediction

✓With the advent of whole-genome sequencing projects, there is considerable use for programs that scan genomic DNA sequences to find genes.

✓Even though there is no substitute to experimentation in determining the exact locations of genes in the genome sequence, a prior knowledge of the approximate location of genes will speed up the process to a great extent thereby saving huge amount of laboratory time and resources.

✓The simplest method of finding DNA sequences that encode proteins is to search for open reading frames, or ORFs.

✓An ORF is a length of DNA sequence that contains a contiguous set of codons, each of which specifies an amino acid[reference], and endmarked by a start codon and stop codon which mark the beginning and the end of a gene respectively.

✓However, biologically, not every ORF is a coding region.

✓Only few ORFs with a certain minimum length and with a specific composition can translate into proteins.

✓And hence, this method fails when there are a large number of flanking stop codons, resulting in many small ORFs.

✓DNA sequences that encode protein are not random chain of available codons for an amino acid, but rather an ordered list of specific codons that reflect the evolutionary origin of the gene and constraints associated with gene expression. This nonrandom property of coding sequences can be used to advantage for finding regions in DNA sequences that encode proteins (Ficket and tung 1992).

✓Each species also has a characteristic patten of use of synonymous codons (Wada et al 1992). Also, there is a strong preference for certain codon pairs within a coding region.

The various methods for finding genes maybe classified into the following categories :

Sequence similarity search:

One of the oldest methods of gene identification, based on sequence conservation due to functional constraint, is to search for regions of similarity between the sequence under study (or its conceptual translation) and the sequences of known genes (or their protein products). [Robinson et al. (1994)]. The obvious disadvantage of this method is that when no homologues to the new gene are to be found in the databases, similarity search will yield little or no useful information.

Based on statistical regularities:

The following measures have been used to encapsulate the features of genes - codon usage measure, hexamer-n measure, hexamer measure, open reading frame measure, amino acid usage measure, Diamino acid usage measure and many more. Combining several measures does improve accuracy.

Using signals:

Any portion of the DNA whose binding by another biochemical plays a key role in transcription is called a signal. The collection of all specific instances of some particular kind of signal will normally tend to be recognizably similar. Our aim in this project is to use machine learning techniques like Hidden Markov Models and Neural Networks to capture the above mentioned properties for gene recognition and prediction.

Issues in Gene Prediction

Three types of posttranscriptional events influence the translation of mRNA into protein and the accuracy of gene prediction. First, the genetic code of a given genome may vary from the universal code. [36, 37]Second, one tissue may splice a given mRNA differently from another, thus creating two similar but also partially different mRNAs encoding two related but different proteins. Third, mRNAs may be edited, changing the sequence of the mRNA and, as a result, of the encoded protein. Such changes also depend on interaction of RNA with RNA-binding proteins. Then there are issues of frame-shifts, insertions and deletions of bases, overlapping genes, genes on the complementary strand etc. Straight-forward solutions therefore do not work when we need to take all these issues into considerations.

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

MCQs

1. A
2. A
3. A
4. A
5. A
6. A
7. A
8. A
9. A
10. A