Computational gene finding

Devika Subramanian
Comp 470

Outline (3 lectures)

Lec 1
- The biological context
- Markov models and Hidden Markov models

Lec 2
- Ab-initio methods for gene finding
- Comparative methods for gene finding

Lec 3
- Evaluating gene finding programs

The biological context

- Introduction to the human genome and genes
- The central dogma: transcription and translation

Facts about the human genome

- The human genome contains 3 billion chemical nucleotide bases (A, C, T, and G).
- About 30,000 genes are estimated to be in the human genome. Chromosome 1 (the largest human chromosome) has the most genes (2968), and the Y chromosome has the fewest (231).

More facts

- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.

More facts

- Genes appear to be concentrated in random areas along the genome, with vast expanses of non-coding DNA between.
- About 2% of the genome encodes instructions for the synthesis of proteins.
- We do not know the function of more than 50% of the discovered genes.

More facts

- The human genome sequence is almost (99.9%) exactly the same in all people. There are about 3 million locations where single-base DNA differences occur in humans (Single Nucleotide Polymorphisms or SNPs).
- Over 40% of the predicted human proteins share similarity with fruit-fly or worm proteins.

A great site to learn more

http://www.dnai.org/index.htm

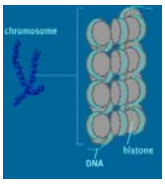
Genome sizes

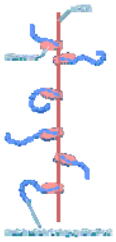
Organism	Genome Size (Bases)	Estimated Genes
Human (*Homo sapiens*)	3 billion	30,000
Laboratory mouse (*M. musculus*)	2.6 billion	30,000
Mustard weed (*A. thaliana*)	100 million	25,000
Roundworm (*C. elegans*)	97 million	19,000
Fruit fly (*D. melanogaster*)	137 million	13,000
Yeast (*S. cerevisiae*)	12.1 million	6,000
Bacterium (*E. coli*)	4.6 million	3,200
Human immunodeficiency virus (HIV)	9700	9

Codons

- 3 consecutive DNA bases code for an amino acid. There are 64 possible codons, but only 20 amino acids (some amino acids have multiple codon representations).
- Four special codons: start codon (ATG) and three stop codons (TAG, TGA, TAA). They indicate the start and end of translation regions.

The central dogma





mRNA produced by transciption

DNA → mRNA → proteins

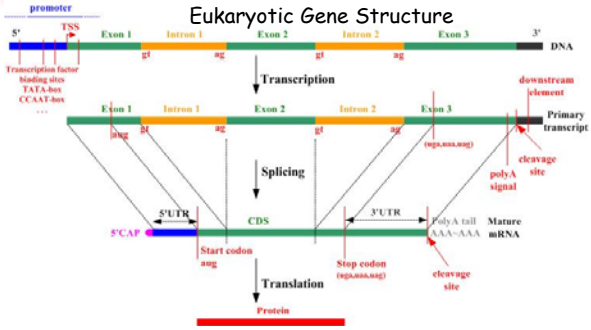
Transcription

- When a gene is "expressed" the sequence of nucleotides in the DNA is used to determine the sequence of amino acids in a protein in a two step process.
- First, the enzyme RNA polymerase uses one strand of the DNA as a template to synthesize a complementary strand of messenger RNA (mRNA) in a process called transcription. RNA is identical to DNA except that in RNA T is replaced with U (for uracil). Also, unlike DNA, RNA usually exists as a single stranded molecule.

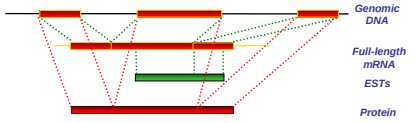
Splicing and Translation

- In eukaryotes, after a gene is transcribed the introns are removed from the mRNA and the adjacent exons are spliced together in the nucleus prior to translation outside the nucleus.
- After the mRNA for a particular gene is made it is used as a template with which ribosomes synthesize the protein in a process called translation.

The Biological Model

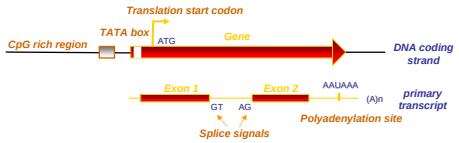


How genes are validated



- cDNA – single-stranded DNA complementary to an RNA, synthesized from it by reverse transcription
- full-length mRNAs (GenBank *RefSeq*, ~16000 human sequences)
- ESTs – Expressed Sequence Tags
  - relatively short, 500 bp long on average
  - span one or more exons
  - large data sets required (GenBank *dbEST* – 4.3 M human sequences)

Signals



- Upstream regulatory signals (TATA boxes)
- Translation start codon (ATG)
- Translation stop codon (*e.g.*, TAA)
- Polyadenylation signal (~AATAAA)
- Splice recognition signals (*e.g.*, GT-AG, branch point)

Computational gene finding

- Gene finding in prokaryotes
- Gene finding in eukaryotes
  - Ab initio
  - Comparative

Finding genes in prokaryotes

- Prokaryotes are single-celled organisms without a nucleus (e.g., bacteria).
- Few introns in prokayotic cells. Over 70% of H. influenzae genome codes for proteins.
- No introns in coding region.

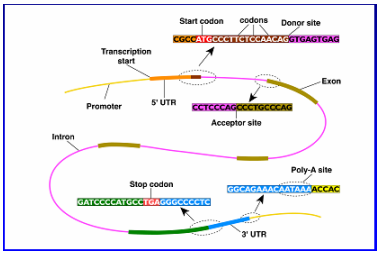
gene1 gene2 gene3

Finding genes in prokaryotes

- Main idea: if bases were drawn uniformly at random, then a stop codon is expected once every 64/3 (about 21) bases. Since coding regions are terminated by stop codons, a simple technique to find genes is to look for long stretches of bases without a stop codon. Once a stop codon is found, we work backward to find the start codon corresponding to the gene.
- Main problems: misses short genes, overlapping ORFs.

Computational gene finding

- Gene finding in eukaryotic DNA



Ab initio methods

- Use information embedded in the genomic sequence *exclusively* to predict the gene structure.
- Find structure G representing gene boundaries + internal gene structure which maximizes the probability P(G|genomic sequence).
- Hidden Markov models are the predominant generative method for modeling the problem.

Ab-initio methods

- Advantages
  - Intuitive, natural modeling
  - Prediction of 'novel' genes, *i.e.*, with no a priori known cDNA or protein evidence
- Caveats
  - Not effective in detecting alternatively spliced forms, interleaved or overlapping genes
  - Difficulties with gene boundary identification
  - Potentially large number of false positives with over-fitting