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

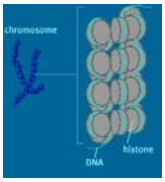
Genome sizes

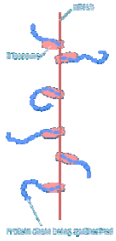
Organism	Genome Size (Bases)	Estimated Genes
Human (*Homo sapiens*)	3 billion	30,000
Laboratory mouse (*M. musculus*)	2.6 billion	30,000
Mustard weed (*A. thaliana*)	100 million	25,000
Roundworm (*C. elegans*)	97 million	19,000
Fruit fly (*D. melanogaster*)	137 million	13,000
Yeast (*S. cerevisiae*)	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

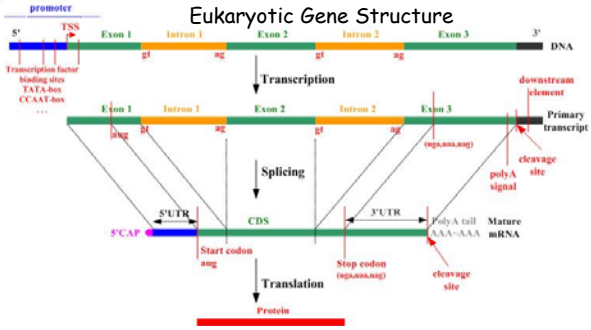
Transcription

- 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.

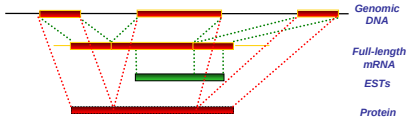
Splicing and Translation

- 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

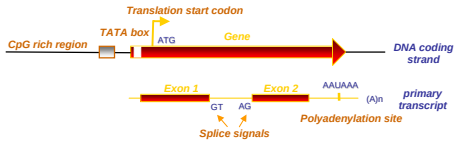


How genes are validated



- 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.

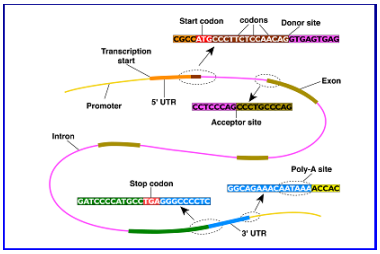
gene1 gene2 gene3

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



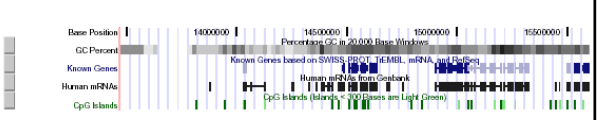
Ab initio methods

- 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

A simple example: CpG Islands

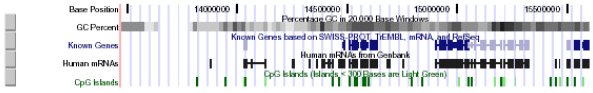


CpG nucleotides in the genome are frequently methylated. (Write CpG not to confuse with CG base pair)

$C \rightarrow$ methyl-$C \rightarrow T$

Methylation often suppressed around genes, promoters$\rightarrow$ CpG islands

Example: CpG Islands



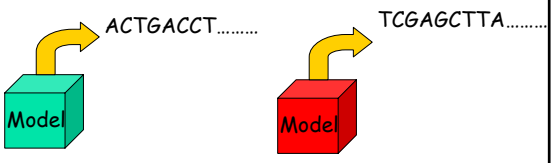
In CpG islands,

CG is more frequent than in the rest of the genome

Two problems

- Given a short DNA sequence, does it come from a CpG island or not?
  - Is this part of a CpG island or not?
- How to find the CpG islands in a long sequence?

Generative models



Models generate sequences of strings in the A,T,C,G alphabet. Model parameters are tuned to reflect characteristics of CpG and non CpG islands.

Markov processes: a quick intro

- We are interested in predicting weather, which can be either sunny or rainy.
- The weather on a given day is dependent only on the weather on the previous day.

$$P(w_t \mid w_{t-1},\ldots, w_1) = P(w_t \mid w_{t-1})$$

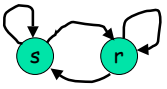
This is the Markov property.

Markov process example

- We have knowledge of the transition probabilities between the various states of the weather: P(s,s').

Rows of the transition matrix sum to 1.

	s	r
s	0.9	0.1
r	0.5	0.5



- We know the initial probabilities of s and r.

Generating weather sequences

- Let the probabilities of weather on day 1 be [0.5 0.5]. We flip a fair coin, and get heads, and obtain sunny to be our weather for day 1.
- Now we consult our transition matrix and find that P(w|s) = [0.9 0.1]. So we flip a biased coin and obtain heads again, so weather on day 2 is also summy.
- We repeat this process, flipping coins biased by the probability $P(w_t|w_{t-1})$ to get a sequence drawn from the s,r alphabet.

Prediction

- Suppose day 1 is rainy. We will represent this as a vector of probabilities over the three values.

$$\pi(1) = [0\ 1];$$

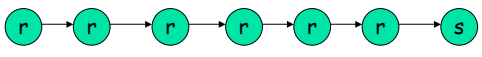
- How do we predict the weather for day 2 given pi(1) and the transition probabilities P?
- From P, we can see that the probability of day 2 being sunny is .5, and for being rainy is 0.5

$$\pi(1) * P = [0.5\ 0.5];$$

Probability of a sequence

- What is the probability of observing the sequence "rrrrrrs"?

$$P(X = rrrrrrs) = \pi(r)P(r|r)P(r|r)P(r|r)P(r|r)P(r|r)P(r|r)P(s|r)$$
$$= \pi(r)\prod_{t=2..7} P(x_t|x_{t-1}) = (0.5)^7$$



Which weather pattern is more likely?

- Given a transition model

$$\begin{array}{c} \\ s \\ r \end{array}\begin{array}{c} \begin{array}{cc} s & r \end{array} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} \end{array}$$

- And an initial state distribution: [0.5 0.5]
- And two sequences: rrrrrrs and ssssssr
- Which is more likely, given the model?

Comparing likelihoods

$$P(X = rrrrrrs \mid Model) = \pi(r)[P(r \mid r)]^5 P(s \mid r) = (0.5)^7$$

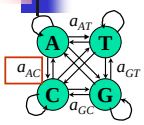
$$P(X = ssssssr \mid Model) = \pi(s)[P_{\log}(s \mid s)]^5 P(r \mid s) = 0.5*(0.9)^5*0.1$$

Markov models

- States: S = $\{s_1,\ldots,s_N\}$, N states
- Transition probability:
  - $a_{ij} = P(X_{t+1}=s_j \mid X_t=s_i)$, i,j in [1..N]
- Initial state probability
  - $pi_i = P(X_1=s_i)$, i in [1..N]

Model generates sequences of states from S, and we can compute how likely a sequence is given the model.

Markov Models for CpG islands



A state for each of the four letters A,C, G, and T in the DNA alphabet

$$a^+_{st} = \frac{c^+_{st}}{\sum_{t'} c^+_{st'}}$$

From a set of known CpG islands, and non CpG islands, estimate the transition probabilities

+	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	.274	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

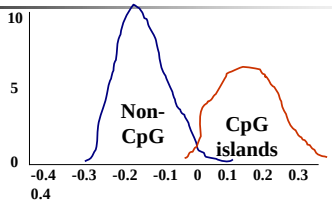
-	A	C	G	T
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

Using the model

- To use these models for discrimination, calculate the log-odds ratio.

$$S(x) = \log \frac{P(x|\text{model}+)}{P(x|\text{model}-)} = \sum_{i=1}^{L} \log \frac{a^+_{x_{i-1}x_i}}{a^-_{x_{i-1}x_i}}$$

Histogram of log-odds scores



Q1: Given a short sequence *x*, does it come from CpG island (**Yes-No** question)?

- *S(x)*

Q2: Given a long sequence *x*, how do we find CpG islands in it (**Where** question)?

- Calculate the log-odds score for a window of, say, 100 nucleotides around every nucleotide, plot it, and predict CpG islands as ones w/ positive values
- Drawbacks: Window size