CSE527, Lecture 14, Nov 13, 2006
Zach Frazier

Administration:

The homework grades will be emailed soon.

There will be a new homework soon.

Gene splicing:
This is a complex process, with lots of different RNA molecules and proteins involved.

The data given is from human, but is characteristic of most mammals. Splicing is common in most eukaryotes, but the stats vary across genera.

Gene Finding is a hard problem, like looking for a needle in a haystack.

The intron lengths can be very long, and the distribution is more skewed than other species.

There is lots of speculation about the origins of intronic DNA.

One hypothesis, the exon-shuffling hypothesis, suggests a link between protein domains or modules and exons. Suggesting that separating small functional or useful regions allows for random mutations to have a higher effect of providing functional proteins.

Scientists speculate that introns developed over one billion years ago, based on homologous exon/intron DNA segments found in both plants and humans, suggesting presence in the common ancestor.

Bacteria do not have introns, but they are under very strict evolutionary pressure, and may have had and lost them in order to gain some efficiency.

The self-splicing introns seen in Tetrahymena (and elsewhere) are believed to be precursors to modern intron splicing systems.

Tetrahymena is a very advanced species with many interesting features. It can reproduce sexually and asexually, and maintains two distinct nuclei.

GC content varies for exons and introns. G and C frequencies are often very similar on a single strand. One reason is believed to be because of flipping during mutation. Which will maintain the two pairing bases in similar frequencies.

Most of the human genome is at 38% GC. Genes most often have a higher GC. Megabase-scale regions of high GC are also known as isochores.

These are observed mostly in vertebres, and have been reported in mammals, and birds, but not fish.

One hypothesis suggests that the genome used to be AT rich, except for the ends of the chromosomes which were GC rich, and over time through chromosonal rearrangement the current pattern emerged. A piece of supporting evidence can be found in the rodent genome, which is more homogeneous then human, and is also more rapidly reproducing.

There is also debate about what evolutionary pressure would be. One theory is that processes related to DNA damage repair are biased toward converting T to C, when mismatched pairs are detected, since that is a common mutation due to methylation. The theory is that this process has over compensated.

From the figures:

> GC rich regions tend to hold more genes.
>
> intron lengths are correlated with GC content, exon lengths are not.

The GenScan program is a gene finding program.

There used to be many gene finding systems which used AI, or expert systems, but these have fallen out of favor as statistical methods proved more accurate.

The training set used was very limited, but very good for the time.

In their results the sensitivity is measured as:
> (true positives)/(actual positives)

The specificity is measured as:
> (true positives)/(predicted positives)

ME : missed exons
WE : wrong exons. (mismatched boundaries?)

They used a Generalized Hidden Markov Model (GHMM).

This differs from the standard HMMs that we have seen, in that instead of emitting single values, the states can emit arbitrary strings (do not have to be of same length).

The standard HMM algorithms such as Viterbi training can still be used, with some modifications.

An example trace through the model:

> Start at N: intergenic region. We emit a typical intergenic region of the standard length distribution we have observed.

Next we move to the promoter box. Here we might emit the TATA-box.

From there we would move to the 5' UTR, where we would emit a typical UTR, which was based on the training data.

Then we might pass through one or more exons, as well as the polyadenylation tail.

There are three exon and three intron states. These correspond to each phase of the reading frame, and ensure that when intron regions are modeled the reading frame stays in the same phase.

If the gene is on the opposite strand, then we have to identify everything backwards. We would begin with the reverse complement of the polyadenylation tail, followed by a reversed 3'-UTR, etc.

Sub-models:

The intron length distribution is approximately geometric.

The initial exon is distinct enough from the others that it is modeled separately. For the length distribution they use a smoothed distribution of observed lengths.

The 3' UTR is modeled through a 5th order Markov Model. The exon sequence is also modeled through a 5th order Markov model.

Weight matrices are used for some sub-models, such as the polyadenylation signal, where the consensus is AATAAA, but others have non-zero probability.

The translation start includes 12 base pairs, starting 6 before the start codon. This is also modelled with a WMM. Translation stop is only modeled as the stop codon and additional base pairs.

Promoters are broken down onto categories. Since 30% of human genes appeared to not have a TATA-box, there is a 30% probability the model will generate a 40 bp sequence according to the background. The other 70% corresponds to the typical promoter regions. It begins with a WMM for the TATA-box, then there is a 15 bp WMM, followed by a uniform distribution for 14 to 20 bp. Then there is another 8bp WMM which corresponds to the CAP-enzyme binding site.

Intron "Sequence Logos":

The vertical black lines represent the start and stop of the intron.

The relative heights of the letters represent their frequency, While the total height of the column represents the entropy of

that position.

The beginning of the intron for example is near several high entropy nucleotides that are almost certain to be present, as is represented by the large G and T.

At the other end we can see the polypyrimidine tract of CT.

The information present does not appear to be enough of a signal to precisely identify splicings. There are probably more signals, they just have not been recovered yet.

For the 5' UTR there was too much dependence between distant columns to use a first order model, so they use a decision tree.

In this model the splice site is at position 0. The column numbers are given relative to the splice site. However the positions 1 and 2 are not given (since they are always GT, resp.)

The \chi^2 test is used to measure the independence of two different positions. In the table high numbers are non-independent. (More precisely, high numbers are incleasingly unlikely if the positions are indepencent.) Similarly, if the \chi^2 is near 0, the two positions are assumed to be independent.

Most values in the table are interesting, in that they have a significant value according to the p-value.

The calculation was not done in an all-vs-all way for all nucleotides in all positions, because there was not enough data, instead the consensus value for each column was compared against all values in the other columns, the consensus values that were used are provided in the first column.

The sums on the right side provide a rough measure of which columns are more important.

The U1 small nuclear RNA is aligned below. This is critical in splicing and it is assumed that there must be a strong match to the reverse of this molecule. Intuitively, it seems plausible that any mismatches in one part of the match, must be made up later in the pairing, and the non-independence data in the table roughly supports this view.

These most important columns appear first in the decision tree. Since the 5th column is most important, we branch on that first. After the split based on this column, we recalculate the column importance with the given values for the column, and continue to procedure down the tree.

We stop when all values are below the \chi^2 cutoff, or when the counts are too small to continue.

The overall success of Genscan is due to its thorough and careful construction. They were careful to not cut corners, and to use the most complex model they could justify while avoiding over-fitting.

GHMM is a very powerful and generic framework for developing applications.

Issues with Training Data:

- single exon genes are over-represented.
- highly expressed genes are over-represented.

pseudo-genes:

- mouse has 5000 olfactory receptors. humans have 100, but pseudo genes can be found for 1000 more which are still recognizable after 65 million years.

There also may be 10,000 or more non-coding RNA genes which these methods will not recognize.