Libby MacKinnon
CSE 527 notes
November 19, 2007
Gene Finding

Characteristics of the Human Genome

While the coding sequences for genes are relatively short (a little over 1000 base pairs on average), their span in the genome can be very large (tens of kilobase pairs long). This is because the introns in the genes can be very long and numerous. There are a few extremely long introns that skew the mean length to three times the median length. This is true for exon length as well, but to a much lesser extent.

GC Content

GC content is different in genes compared to the whole genome. In genes, the GC content is centered around 45-50%, and it is more uniformly distributed than in the genome. The genome is mostly 38% GC with its distribution skewed to the left. As a consequence, regions of high GC content (62-68%) have higher relative gene density than regions of lower GC content.

Exon length is relatively uniform with respect to GC content, but intron length decreases dramatically in regions of high GC content. Specifically if the GC content is around 30%, the average intron length is 2300 base pairs, and if the GC content is around 65%, the average intron length is 300 base pairs. It is hard to say if GC content has an effect on the intron length or if the intron length has an effect on the GC content, but the relationship is clear.

Computational Gene Finding: GENSCAN

The composition of genes is very complex, with a lot of variation from gene to gene. The next section of the notes is dedicated to describing one way to algorithmically account for all the complexity to predict the locations of genes in genome data. GENSCAN, from Burge and Karlin's 1997 article in J. Mol. Biol. "Prediction of complete gene structures in human genomic DNA" is a gene finding program, which is very successful.

The training data was limited because this was over ten years ago. 380 known genes were used. 142 of those had only one exon and 238 of them were multi-exon. All-in-all there were 1492 exons and 1254 introns for a total of 2.5 Mega-base-pairs.

The performance was very good compared to ten other gene finding programs of the time. Performance is measured in several ways. Sensitivity and the specificity of the programs are most important. Missed exons are the total number of true exons that the programs did not find, and wrong exons were predicted exons that are not truly exons. GENSCAN did better than all ten other programs in all of these categories.

Generalized Hidden Markov Models

A generalized hidden Markov model (GHMM) is used in GENSCAN, which differs from the hidden Markov models we have seen in that the states in the model are emitting sequences which do not have to be the same length. Each state has a different submodel and there are transition probabilities between the states.

Given length L, a start state q_1 is chosen. Then the length (d_1) of the substring s_1 is chosen and the string is generated from the submodel for q_1. Then the next state is chosen and a length (d_2), and substring s_2 from the submodel for q_2. This process continues until $\sum d_i = L$. Once a sequence of these substrings are generated, an algorithm is used to calculate the probability, which involves summing over possible predecessor states and possible subsequence lengths.

The GHMM for GENSCAN is structured as follows. Start in the intergenic region, N. A typical intergenic sequence is emitted based on observations in data. The only state to move to from N is P, the promoter state. Here something close to a TATA box might be emitted. From P, the only state to move to is F, the 5' untranslated region, where a sequence would be emitted based on data.

After F, it is possible to move to E_{sngl} (meaning the gene generated will only have one exon) or E_{init} (the first exon of a multiple-exon gene). The probability of moving to these states is based on the training data. From E_{sngl}, the next state is T (3' untranslated region) and then A (polyadenylated tail) and back to N. From E_{init} there are three different intron states corresponding to how the reading frame is shifted. There are also three different exon states that can follow the intron states.

The entire model has a $+$ side and a $-$ side, so that the forward and reverse strand can be considered at the same time. The model for the $-$ side is identical with the arrows between states pointing in the opposite direction.

Submodels

The intron length distribution is approximately geometric and is GC dependent. The GC effect was accounted for by binning data according to GC percentage creating separate models for each range. 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 intron sequence is based on a 5th order Markov model based on observed data.

The 5' and 3' untranslated regions are also based on 5th order Markov models and their lengths are geometric distributed.

Weight matrices are used for some sub-models, such as the polyadenylation signal, where the consensus is AATAAA, but others have non-zero probability. They are also used for translation start (12 base pairs) and translation stop (3 base pairs). Promoter submodels use WMM's as well, broken down into two categories. One is with a TATA box (70% probability of this), with a 15 base pair TATA WMM followed by a background sequence for 14-20 base pairs followed by an 8 base pair cap signal WMM. The other is TATA-less (30% probability of this), which is a 40 base pair background sequence.

Exons are modeled by inhomogenious 3-periodic 5th order Markov models with separate models for low GC (¡43%) and high GC.

Splice Sites

Certain nucleotide positions around donor and acceptor splice sites exhibit a lot of conservation. The two nucleotides immediately following the donor splice site on the intron are almost always GT. The two nucleotides immediately preceding the acceptor splice site on the intron are almost always AG.

The donor splice site, which is the 5' end of the exon, shows dependencies between non-adjacent poisitions. The model used to describe this is basically a decision tree that uses a χ^2 test to quantitate dependence. Out of all donor splice sites, the +1 and +2 positions are always G and U based on training data, so they are not in the decision tree. 84% of the +5 position is G, so the first step in the decision tree is $+5 = G$ or $+5 \neq G$. Then there are separate statistics based on what has been observed for $+5 = G$ vs. $+5 \neq G$. If $+5 = G$ the next decision is $-1 = G$ vs. $-1 \neq G$. This continues for all sites from -3 to +5.

GENSCAN Summary

Burge and Karlin's algorithm paid a lot of attention to small details in their model, which lead to good accuracy. This is necessary because the coding DNA and signals are not random, as they are in other parts of the genome. Each region was modeled differently based on the nature of the training data. There were problems with this training data, though. Single exon genes, moderately sized genes, and highly expressed genes may be over-represented. There may be annotation errors as well, but there would likely be problems with other training sets, too.

In any gene finding method, it is difficult to account for a lot of things while keeping the program simple enough to work with. Some of these are: psuedogenes, short open reading frames, errors in sequences, ,non-coding RNA genes, overlapping genes, and alternative splicing. Also, these programs in general do not find new things. They only find what is already known in the training data.

There are other important questions to ask when predicting genes. One is if the predicted gene looks like a known protein, which would involve a database search. Also what does the same region look like in related organisms?