Nov 21, 2005 CSE527 class notes
Jeff Pincus

80 to 90% of the prokaryotic genome is coding; an open reading frame (ORF) search is sufficient to find most genes. ORF search misses short genes or multiple genes sharing portions of the same sequence. Eukaryotic genes are harder to find because they're interrupted by introns, and the intron/exon boundries are unclear.

Humans generally have 6-10 exons per gene. The median intron length is 1 kb and the mean length is 3 kb.

The splisosome assembles mRNA. It's made up of short RNAs complimentary to splice sites and about 50 known proteins. There is plausible communication between exons. Alternative splicing, which results in shorter and longer proteins, is widespread.

The Karlin paper didn't use much training data, still the GENESCAN program, which only looks at sequence, performed better than GeneID, parser3, which also use protein data.

The GENESCAN numbers look good on a per exon basis, but it's only about 50% accurate for finding whole genes. It uses a hidden Markov model with output strings rather than individual letters. The find states and the string duration are hidden.

An exon can end anywhere in a codon.

Poly-A tails are added to the end of nascent mRNA molecules, they're part of a signaling mechanism. There's a consensus sequence in the molecule for it.

Submodels are 5th order MM, used for exon sequence.

inhomogeneous: exon can start in any codon position, although there is a minor preference for frame 0.
3-periodic: another word for codon

The 3' end of the intron is the acceptor end, and the 5' end is the donor. There is a poly-pyrmidine tract at the 3' end of the intron. The U1 snRNA has a consensus sequence to hybridize to the splice site.

The intron may contain microRNA regulators.

A mismatch at one end of the splice site plausibly requires a better consensus at the other end.