CSE 527 notes, 10/22/2007 starting with the slides named “lec04”. Ryan Stephens

Slide	Notes
22, 24	Picked up in the middle of the Expectation-Maximization algorithm. • This slide is about the “E-Step”. Essentially, what you do is assume that θ is known and fixed. Now you can estimate the values for the other hidden parameters. • The E step provides a value for E(z) given a fixed value for θ, • The M step then calculates a value for θ that maximizes the likelihood
26,27	(see slides) The conclusion is that EM is not perfect, but works well in practice.
29	A gene is a DNA sequence for a single protein. DNA is transcribed into many instances of RNA, each of which is translated into multiple instances of (the same) protein.
30	Eukaryotes are more complicated than prokaryotes; euk have 3 RNA polymerases and prok have one. Genes can’t be as long in prokaryotes as in euk; e.g. a bacteria will divide before there’s enough time for a long gene to be transcribed.
31	Multiple RNA polymerases can be simultaneously transcribing a gene; each resulting mRNA can be used for multiple copies of a protein before it degrades. Alternative ribosome binding sites mean that one mRNA can produce multiple non-identical proteins.
32	“Capping” protects RNA 5’ end from degradation. Also provides a starting point for translation.
34,35	The poly-A complex cleaves the mRNA sometime after AAUAAA, the poly-A tail is added and that RNA goes on its way to being translated. Meanwhile, the RNA polymerase continues transcribing the DNA for awhile, but the resulting RNA is uncapped so it degrades quickly. Eventually the RNA polymerase falls off.
Starting on “lec05” slides	
4	The double helix has major and minor grooves. Each base pair has a distinct pattern for potential Hydrogen bonds,

4	The double helix has major and minor grooves. Each base pair has a distinct pattern for potential Hydrogen bonds, accessible within the major groove.
5-8	There are several different motifs that describe how proteins attach to DNA. (Note: the term "motif" is overloaded here to mean both structural motifs of proteins and sequence motifs of DNA. Slides 5-10 are primarily about structural motifs of proteins and DNA sequence motifs are slides 13-end.) This is determined largely by the 3-dimensional shape of the proteins and the double helix, and the alignment of chemical bonds between them. Some proteins cause the DNA to bend around them instead of vice versa.
9,10	Some interactions are well understood, but others are not. As an example, different "finger" motifs bond to DNA in different ways, e.g. TTK, Zif, GLI finger motifs. This makes it extremely difficult to predict binding sites, since they can occur in different ways.
13-15	An interesting sequence motif is the "TATA box" from E. coli, that is a transcription promoter. The TATA box occurs upstream from the gene, and indicates a starting point for transcription. There is a loose consensus on what the TATA box looks like, but it differs slightly in every organism where it is known to exist. This makes it difficult to locate programmatically.
16-23	A weight matrix makes the job easier. Using a weight matrix we can calculate the LLR that a sequence is a TATA box, compare that result to a random sequence. Neyman-Pearson says that's all we need, and it seems to work well.
25, 26	These slides show an example of an WMM.
27-34	Somewhat loosely speaking, for any 2 discrete probability distributions, the "relative entropy" quantifies how different they are. For distributions on fixed-width sequence motifs defined by a WMM versus a fixed i.i.d. background nucleotide distribution, (a) it's easy to calculate their relative entropy, which is a number between 0 and 2 bits per column, and (b) it gives the expected score difference between a sequence drawn at random according to the background model, vs one