CSC2427H, Winter 2006
Algorithms in Molecular Biology
Lecture #11

Lecturer: Michael Brudno
Scribe: Hania El Ayoubi
February 15, 2006

[**Acknowledgement**: The diagrams in the notes below were taken from the slides for Lectures 11 and 13 of the Winter 2005 session of the course *CSC 262 Computational Genomics* taught by Professor Serafim Batzoglou at Stanford University.]

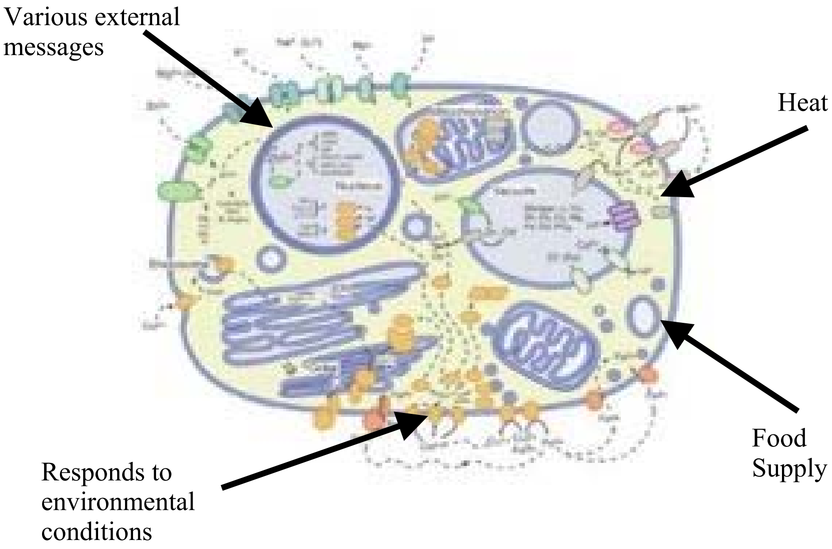
This lecture covers the following topics:

- Gene Regulation
- Gene Expression
- Gene Finding

1. Gene Regulation

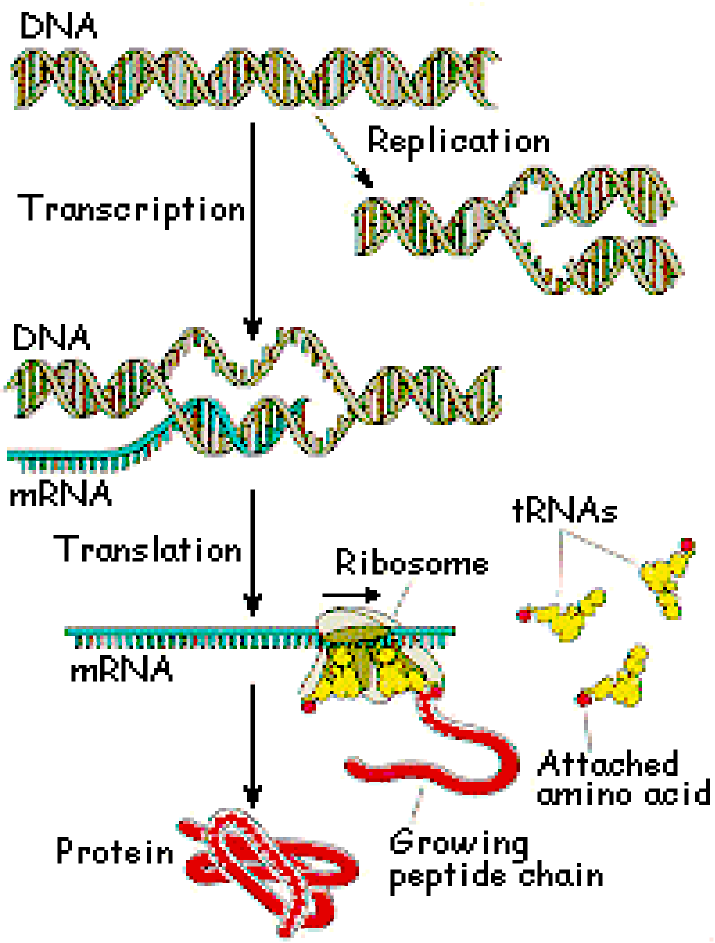
What is the structure of genes? (Not to be confused with that of proteins)

A cell is a machine responding to stimuli from its environment. For example, the sun shining on a cell could initiate the defense mechanism for preventing heat shock. Another kind of stimulus is the external messages sent to the cell. These messages affect how a cell works; the cell processes the messages and reacts to them. Examples of messages include hormones and also the (literal) transmission of messages by neuron cells. So the messages could be chemical or electrical.


Various external messages
Heat
Food Supply
Responds to environmental conditions

Note that the cell has to respond to dynamic environmental conditions. However, the genome is fixed – every cell has a copy of the same genome. Yet, if its genome is static and never changes, how does the cell respond to external stimuli, such as the defense against heat shock? Cells encode different proteins and toggle their production, allowing the cell to remain dynamic despite of its static genome. So gene regulation is responsible for the cell being dynamic. Gene expression, in which genes are toggled on or off, varies according to cell type, cell cycle, external conditions, and location.

Gene regulation takes place at many levels:



1. Opening of chromatin: DNA is not linear, and for a gene to be expressed, the chromatin has to be exposed.

2. Transcription: Regulation can happen at this level by controlling which RNA is read from the DNA sequence.

3. Translation: Even if RNA can be read, it may not code for a protein. Only a small portion of RNA will get translated.

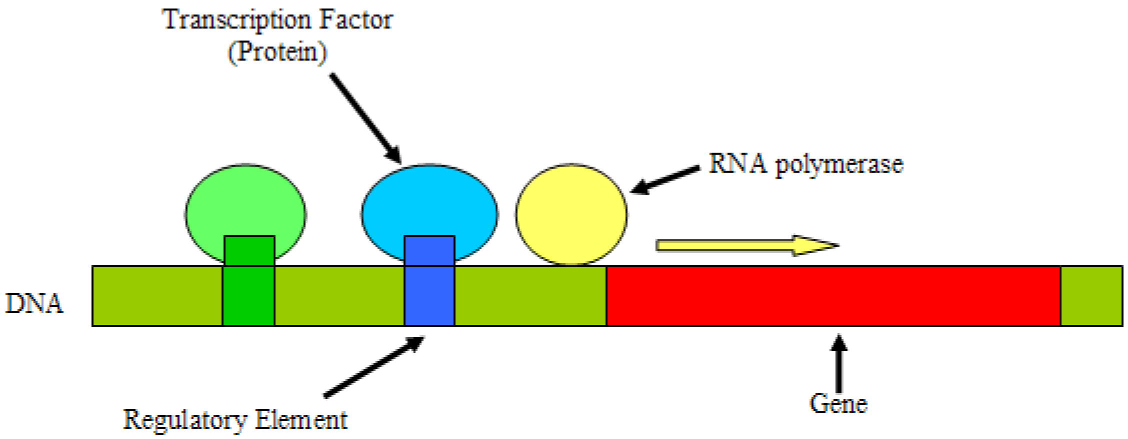
4. Protein stability: If the resultant protein does not make sense (or the cell is in a high heat condition), then the protein would fall apart.

5. Protein modifications: These modify the protein after it has been built.

Although gene expression can be regulated at each of the above levels, transcription is the main gene regulation mechanism because it conserves energy; no RNA (or protein) are produced and then broken up.

Gene regulation at the transcription level, although efficient in terms of energy, has the slowest response time. That is because, after a receptor notices a change, it has to cascade the message to the nucleus. The chromatin must then open and transcription factors must bind to the DNA. Then, RNA polymerase must be recruited to transcribe, and the resulting mRNA has to be spliced and sent to the cytoplasm. Finally, the mRNA is translated into a protein.

The transcription factors recognize DNA substrings, which are hence regulatory motifs, and bind to them. The transcription factors that bind to DNA are known as *promoters* and *enhancers* of transcription.



A promoter is necessary to start transcription and is found in front of a gene. The promoter binds proteins, called transcription factors, which recruit the RNA polymerase to come and start reading. If a promoter is absent, then that region of DNA cannot code and is known as a pseudogene. These are usually not functional. The core promoter has the TATA binding site, and other binding sites exist for transcription factors.

Side Question: Since pseudogenes are not functional, are they affected by evolution?

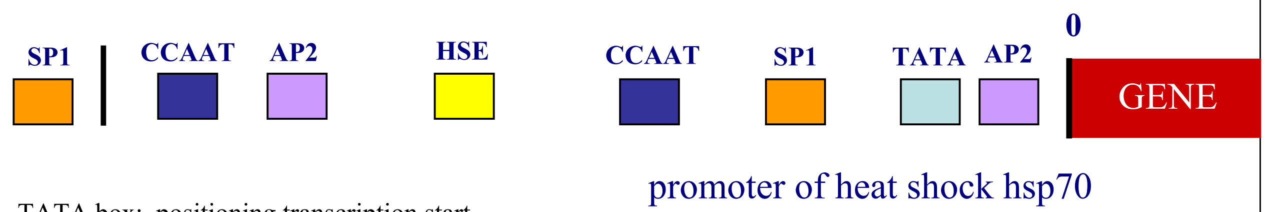
Answer: Pseudogenes may have resulted from mRNA somehow getting into the genome, and since mRNA does not contain a promoter, a pseudogene results. Over time (i.e. generations), it will start to look random.

Enhancers may be present, in front of or behind the gene. They are not really necessary but help in transcription. They can be quite far away on the linear sequence of DNA from the gene, but in 3-D they may be close to it, which means that it still has an effect on transcription despite the distance.

Promoters can have several transcription factors binding to DNA. Some transcription factors will not be sufficient to start transcription on their own and will require other transcription factors to bind as well.

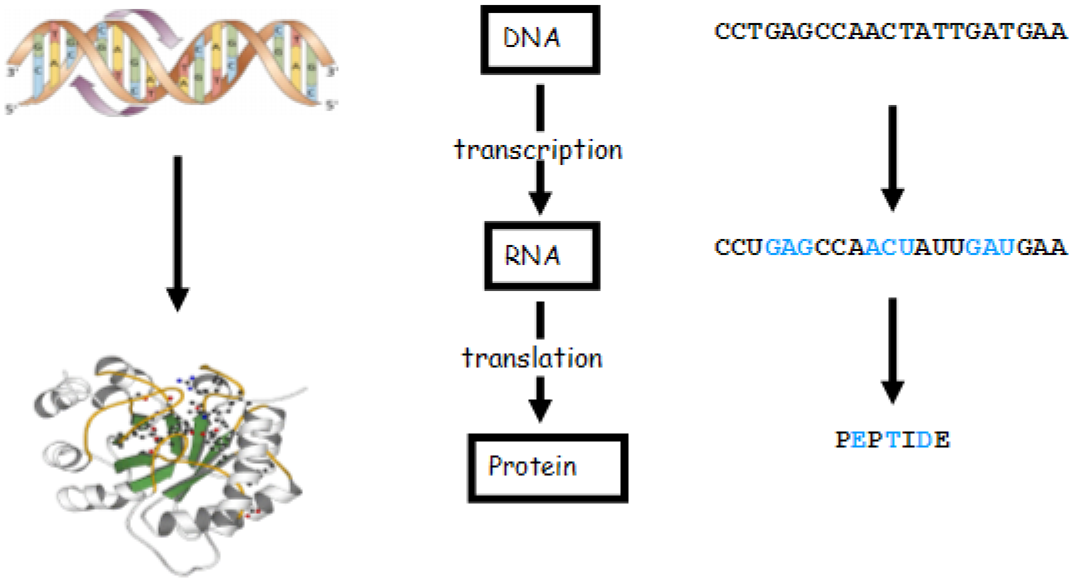
The regulatory element (shown in the above diagram) is the binding site for transcription factors. Once the transcription factors bind, the RNA polymerase bind to the transcription factors and starts reading the gene.

The diagram below shows the transcription factors in a human heat shock protein:



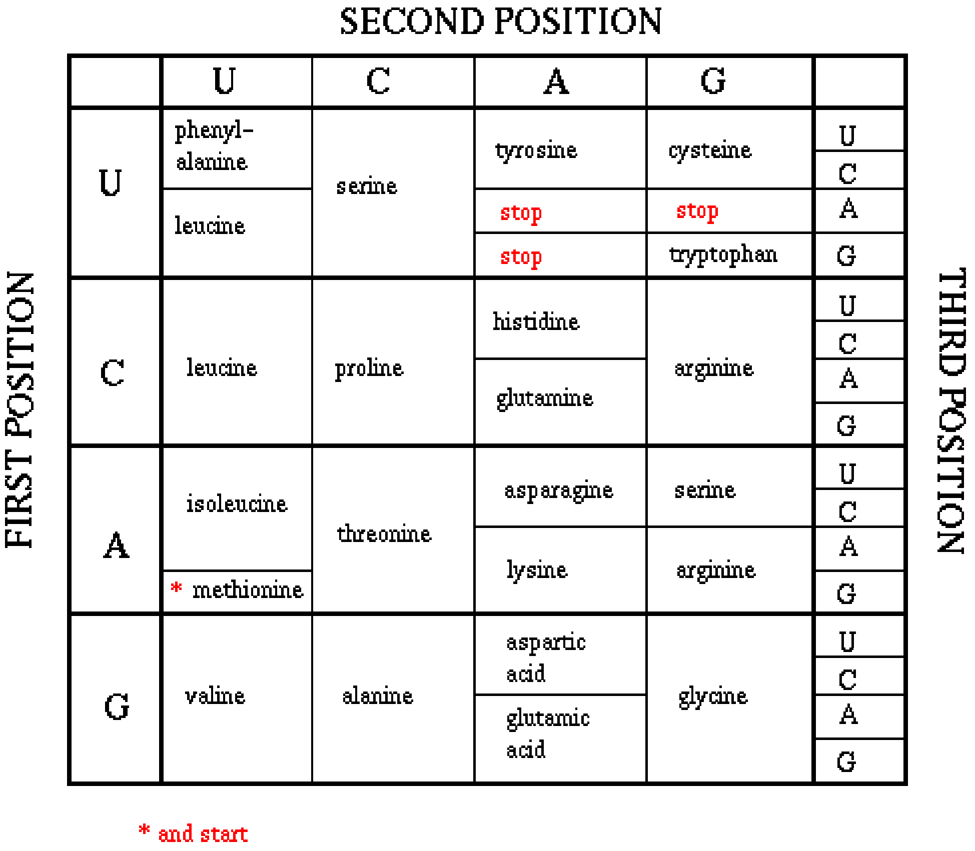
TATA box: positioning transcription start
TATA, CCAAT: constitutive transcription
GRE: glucocorticoid response
MRE:metal response
HSE:heat shock element

2. Gene Expression



Gene expression takes the actual DNA sequence coded in triplets to the peptide. Triplets of RNA code for a single peptide. Generally, the third position is wobbly without affecting structure of the protein. For example, GUA, GUC, GUU, GUG all code for Valine.

`Table 1 RNA Codon Table`
`[source: http://darwin.nmsu.edu/~molb470/fall2003/Images/codon_table.gif]`



	SECOND POSITION				
FIRST POSITION	U	C	A	G	THIRD POSITION
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
			lysine	arginine	A
	* methionine				G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

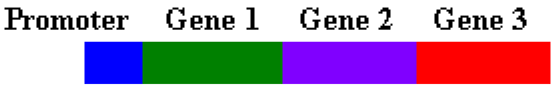
* and start

3. Gene Finding

Given a genome sequence, we would like to find the genes in it. Mechanisms differ for eukaryotes, which have a nucleus, and prokaryotes, which don't. More generally, eukaryotic cells are characterized by membrane-bound compartments, which are absent in prokaryotes.

Only 2-2.5% of the human genome is made up of genes. We do not know the function of the rest of the genome. The proportion of coding DNA in other organisms is higher. Fugu fish, for example, has a very compact genome, 10 folds shorter than that of human. Yet, Fugu genes make around 23% of its genome. In bacteria the situation is completely different, as they have no introns. For example, in Haemophilus influenza baceteria, genes make up ~85% of the genome. Bacteria have operons, which are regions having a promoter and several genes, which means all those genes are either on or off, since they have the same promoter. Open reading frames are contiguous sets of codons that start with the Methionine codon, ATG, which is the start codon and end with a stop codon. (Every single protein in eukaryotes starts with the same amino acid Methionine,
and all stop with one of three possible stop codons (TAA, TAG, TGA). The start codon gets translated into the resulting protein as Methionine, while the stop codon does not code for anything in the resulting protein.)

Several genes may have the same promoter, as shown in the diagram below. This means that they are either expressed together, or "toggled off" together.

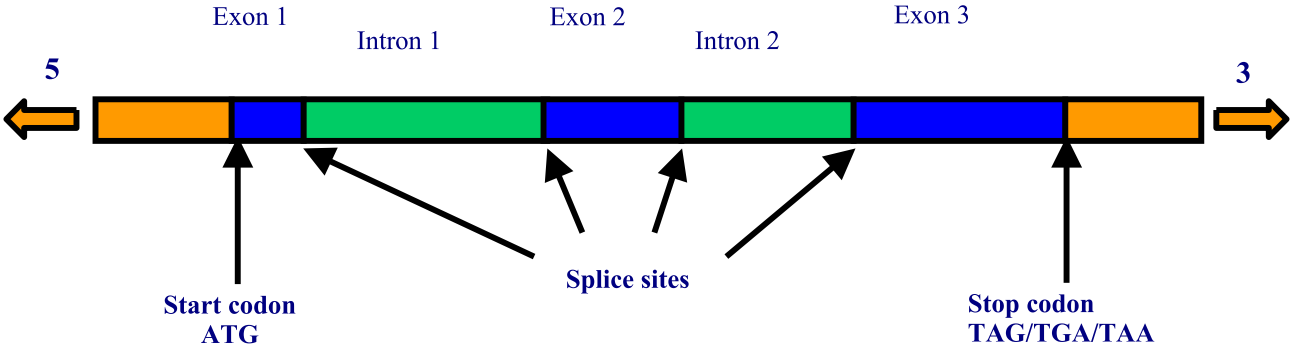


Operon

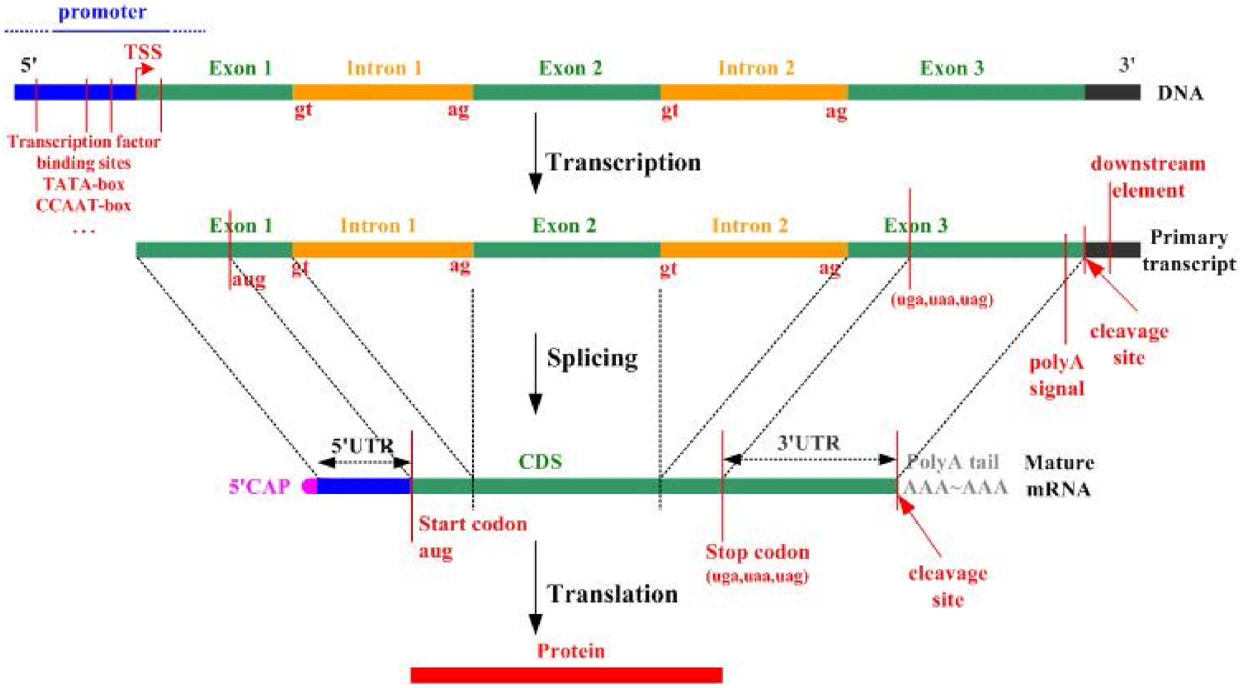
Every possible DNA sequence has 6 frames of translation. Consider a frame shift by one; you might encounter stop codons in the middle of the gene due to the shift. Open reading frames have a start and end and are of a certain length. To find genes in bacteria, look for open reading frames. Do not look for a promoter, but look for ATG, followed by a sequence of non-stop codons, typically 20 codons in length, then a stop codon. In bacteria, genes cover the entire genome (since introns do not exist), so one can easily find genes by a simple algorithm looking for open reading frames.

Eukaryotes' gene structure is much more complicated. Eukaryotes' genomes have exons separated by spaces (or introns). The first exon is preceded by a promoter box. Introns are transcribed, but after splicing, they are lost. Eukaryotes also have a 5' UTR (UnTranslated Region, which is a non-coding exon) preceding the first coding exon, and a 3' UTR following the last coding exon. UTR have functions; for instance, they are believed to encode signals for the ribosomes, for example, to indicate to the latter how aggressively the mRNA should be translated.

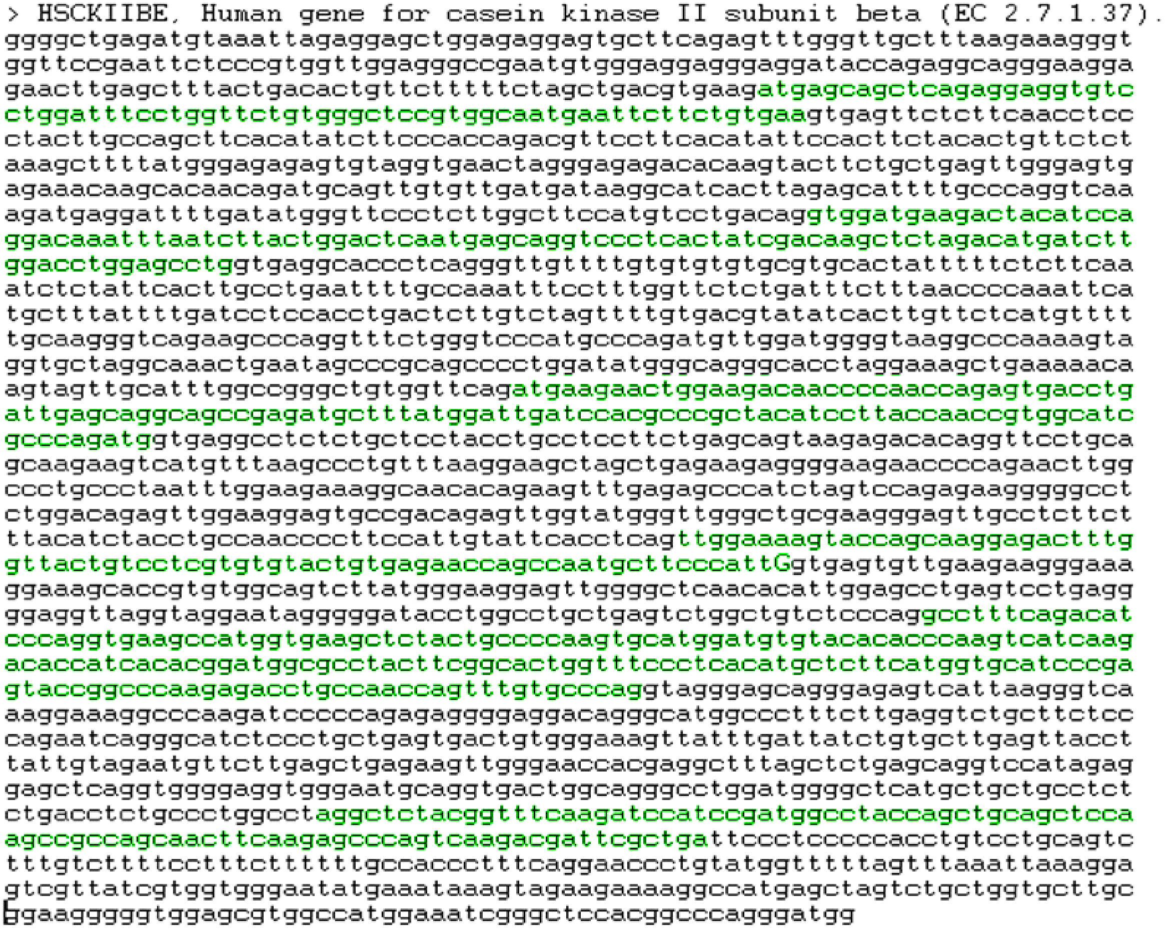
Eukaryotes' gene structure



Eukaryotes also have exons that are sometimes included and sometimes excluded, resulting in different proteins. This is referred to as *alternative* splicing. A very common feature of the skipped exons is that the number of bases in them is a multiple of 3 (or else there would be a frame shift when these exons get cut out, resulting in a major alteration of the codons on the mRNA). Introns do not have to have a multiple of 3 base pairs because they get spliced out and are not translated.



Exons' beginnings are somewhat similar (see diagram below) and their ends are also similar.



```
> HSCKIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgcttcagagtttgggttgctttaagaaagggt
ggttccgaattctcccgtggttggagggccgaatgtgggaggagggaggataccagaggcagggaagga
gaacttgagctttactgacactgttctttttctagctgacgtgaagatgagcagctcagaggaggtgtc
ctggatttcctggttctgtgggctccgtggcaatgaattcttctgtgaagtgagttctcttcaacctcc
ctacttgccagcttcacatatcttcccaccagacgttccttcacatattccacttctacactgttctct
aaagctttatgggagagagtgtaggtgaactagggagagacacaagtacttctgctgagttgggagtg
agaaacaagcacaacagatgcagttgtgttgatgataaggcatcacttagagcattttgcccaggtcaa
agatgaggatttgatatgggttccctcttggcttccatgtcctgacaggtggatgaagactacatcca
ggacaaatttaatcttactggactcaatgagcaggtccctcactatcgacaagctctagacatgatctt
ggacctggagcctgggtgaggcaccctcagggttgttttgtgtgtgtgcgtgcactatttttctcttcaa
atctctattcacttgcctgaattttgccaaatttcctttggttctctgatttctttaaccccaaattca
tgctttatttgatcctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccaggtttctgggtcccatgcccagatgttggatggggtaaggcccaaaagta
ggtgctaggcaaactgaatagcccgcagcccctggatatgggcagggcacctaggaaagctgaaaaaca
agtagttgcatttggccgggctgtggttcagatgaagaactggaagacaaccccaaccagagtgacctg
attgagcaggcagccgagatgctttatggattgatccacgcccgctacatccttaccaaccgtggcatc
gcccagatgggtgaggcctctctgctcctacctgcctccttctgagcagtaagagacacaggttcctgca
gcaagaagtcatgtttaagccctgtttaaggaagctagctgagaagaggggaagaaccccagaacttgg
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccatctagtccagagaagggggcct
ctggacagagttggaaggagtgccgacagagttggtatgggttgggctgcgaagggagttgcctcttct
ttacatctacctgccaaccccttccattgtattcacctcagttggaaaagtaccagcaaggagactttg
gttactgtcctcgtgtgtactgtgagaaccagccaatgcttcccattGgtgagtgttgaagaagggaaa
ggaaagcaccgtgtggcagtcttatgggaaggagttggggctcaacacattggagcctgagtcctgagg
ggaggttaggtaggaatagggggatacctggcctgctgagtctggctgtctcccaggcctttcagacat
cccaggtgaagccatggtgaagctctactgccccaagtgcatggatgtgtacacacccaagtcatcaag
acaccatcacacggatggcgcctacttcggcactggtttccctcacatgctcttcatggtgcatcccga
gtaccggcccaagagacctgccaaccagtttgtgcccaggtagggagcagggagagtcattaaggtca
aaggaaaggcccaagatcccccagagaggggaggacagggcatggccctttcttgaggtctgcttctcc
cagaatcagggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagttgggaaccacgaggctttagctctgagcaggtccatagag
gagctcaggtggggaggtgggaatgcaggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaagatccatccgatggcctaccagctgcagctcca
agccgccagcaacttcaagagcccagtcaagacgattcgctgattccctcccccacctgtcctgcagtc
tttgtcttttcctttctttttgccaccctttcaggaaccctgtatggttttagtttaaattaaagga
gtcgttatcgtggtgggaatatgaaataaagtagaagaaaaggccatgagctagtctgctggtgcttgc
ggaaggggtggagcgtggccatggaaatcgggctccacggcccagggatgg
```

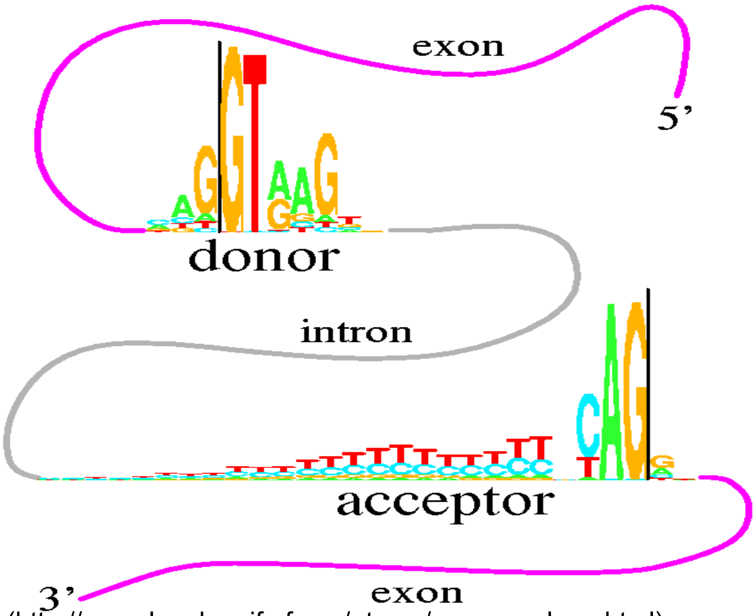
One way to find coding areas in a genome is to use EST (Expressed Sequence Tag) Technology. Biologists sequence the RNA, and so the sequence obtained is not that of the genome, but of the RNA transcribed from it. This approach is costly and produces false positive due to aberrant splicing (i.e. garbage mRNA gets sequenced, since EST sequences all RNAs, and not just those that code for proteins). EST is also difficult for rare genes, since they may be only present in certain tissues in low quantities, reducing the chance of sequencing them. Despite all its disadvantages, EST still helps with the problem of finding genes. Using EST-based gene prediction, we can design alignment algorithms for predicting genes.

Side Question: What causes cell differentiation?

Answer: There are gene regulatory networks, but those are very poorly understood. At some "magic moment", the cell knows to start differentiating.

There are also methods based on recognizing splice sites. Near the splice sites, the probability that each letter occurs deviates significantly from 25%. For example, in the sequence logo below[1], the two letters that occur right before exon 2 are (almost) always AG, and the next letter (the first in exon 2) is most often a G. Before the aforementioned AG, there is a region of Cs and Ts.

This figure shows two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAG|GT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)



(http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html)

Another source of information for finding genes is the coding bias. For instance, you know you should not encounter stop codons, and different codons appear with different frequencies. We can calculate the frequency with which each occurs in a particular gene.

[1] Note that in a sequence logo, the height of the letter is proportional to its frequency. Also note that the areas in the sequence logo with height 0 are where bases occurred with equal probability at that position. Since no information was gained for those positions, their height is 0.

Finally, genes are conserved across organisms through the course of evolution. For example, comparing genomes will show regions of similarity, which tend to correspond to genes. One can then deduce exons from the regions of high similarity among the different organisms.

The main four approaches to gene finding are:

1. Homology: Use a library of known EST, for example in the tool Procrustes.
2. Ab initio: Given a sequence, determine what the genes are, for example in the tools Genscan, Genie, and GeneId.
3. Comparative: Determine genes by comparing the genes of say, human to mouse., for example in the tools TBLASTX and Rosetta.
4. Hybrid : Takes all these approaches and puts them together to try to draw a big picture, for example in the tools GenomeScan, GenieEST, Twinscan, and SLAM.

There is a small industry that develops hidden markov models to look at different sources of information and develop proper HMMs.