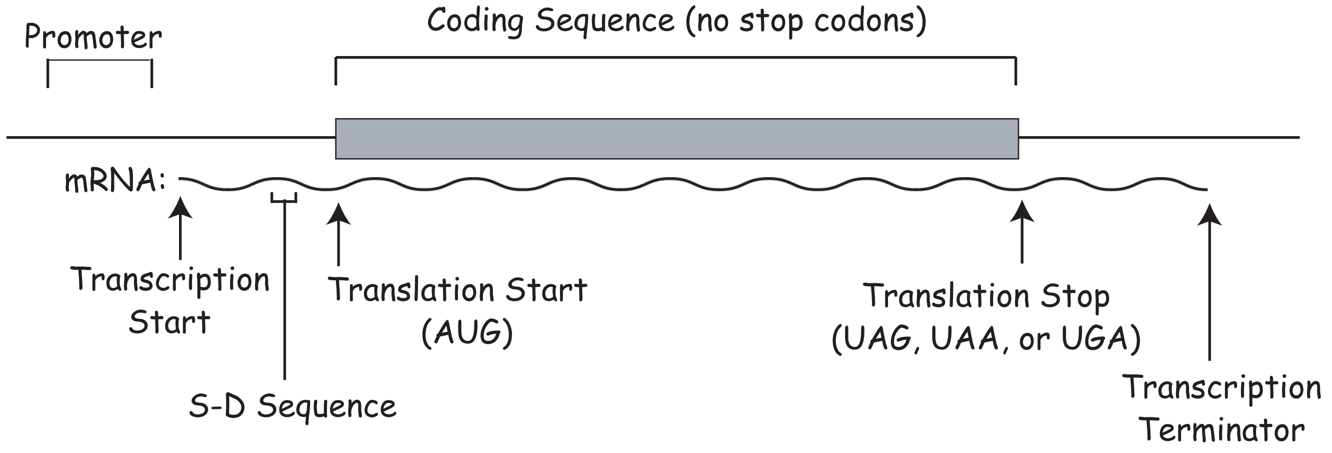
Lecture 12

Analysis of Gene Sequences

Anatomy of a bacterial gene:



Sequence Element	**Function**
Promoter	To target RNA polymerase to DNA and to start transcription of a mRNA copy of the gene sequence.
Transcription terminator	To instruct RNA polymerase to stop transcription.
Shine-Dalgarno sequence	S-D sequence in mRNA will load ribosomes to begin translation. Translation almost always begins at an AUG codon in the mRNA (an ATG in the DNA becomes an AUG in the mRNA copy). Synthesis of the protein thus begins with a methionine.
Coding Sequence	Once translation starts, the coding sequence is translated by the ribosome along with tRNAs which read three bases at a time in linear sequence. Amino acids will be incorporated into the growing polypeptide chain according to the genetic code.
Translation Stop	When one of the three stop codons [UAG (amber), UAA (ochre), or UGA is encountered during translation, the polypeptide will be released from the ribosome.

Example: A gene coding sequence that is 1,200 nucleotide base pairs in length (including the ATG but not including the stop codon) will specify the sequence of a protein 1200/3 = 400 amino acids long. Since the average molecular weight of an amino acid is 110 da, this gene encodes a protein of about 44 kd — the size of an average protein.

The Genetic Code

1st position (5′ end) ↓	2nd position				3rd position (3′ end) ↓
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Classically, genes are identified by their function. That is, the existence of a gene is recognized because of mutations in the gene that give an observable phenotypic change. Now, in the era of genomic sequencing, many genes of no known function can be detected by looking for patterns in DNA sequences.

The simplest method which works for bacterial and phage genes (but not for most eukaryotic genes as we will see later) is to look for stretches of sequence that lack stop codons. These are known as "open reading frames" or **ORF**s.

Let's see how a typical bacterial genome, which is a circular DNA molecule of about $4x10^6$ base pairs, would be analyzed in terms of the ORFs that it contains.

Each reading frame can be considered a circular string containing 1.33×10^6 triplet codons. The probability of a random stop codon for DNA with equal content of A•T and G•C base pairs will be 3/64. [This probability will be lower for DNA with less than 50% A•T bases since stop codons have are either 2/3 or all A•T bases.]

Thus each reading frame will have on average 6.25×10^4 random stop codons.

If one imagines cutting a piece of string n time this process will yield n pieces of string.

Thus each reading frame will have on average 6.25×10^4 random ORFs of different lengths.

Lets say we wanted to calculate how many fortuitous ORFs of length > 200 codons would occur at random in a given reading frame. The probability of an ORF of length > 200 is the probability of not having a stop codon 200 times in a row, which can be expressed:

$p(ORF >200) = (1 - 3/64)^{200} = (61/64)^{200} = 6.76 \times 10^{-5}$

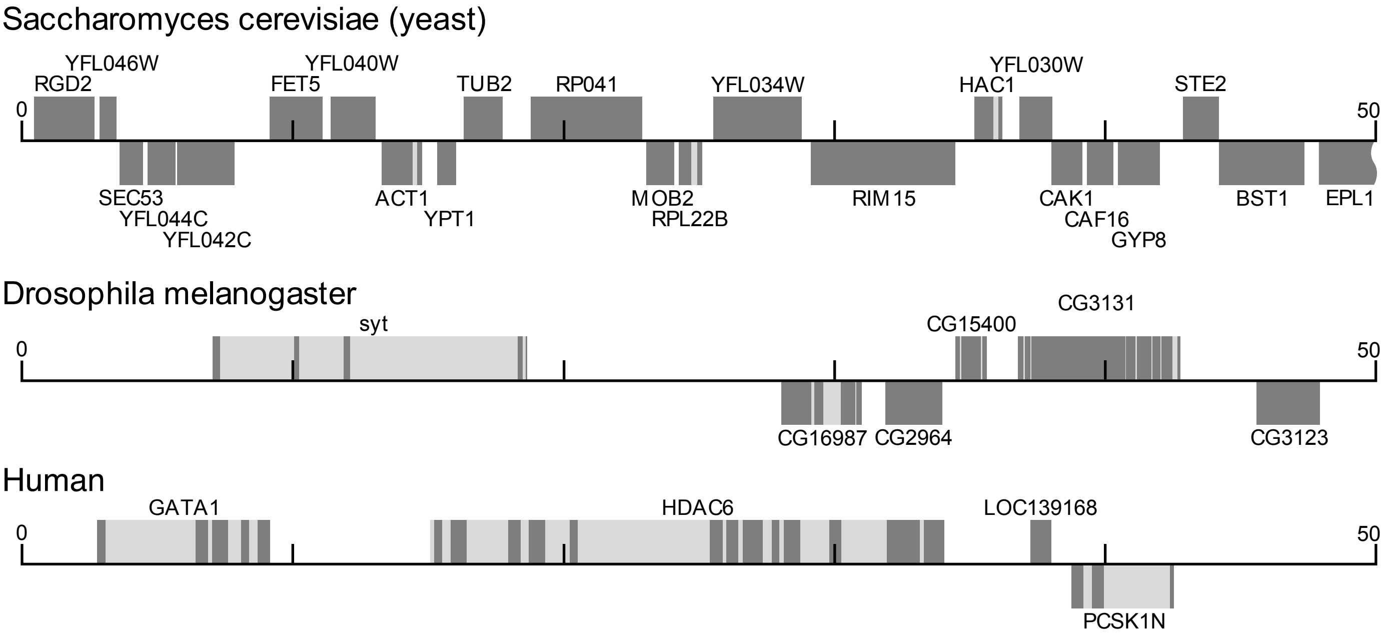
Thus for one reading frame the total number of fortuitous ORFs length > 100 will be
$= 6.76 \times 10^{-5} \times 6.25 \times 10^4 = 4.2$

If one considers reading in one direction (clockwise say) there are three distinct readig frames, because each codon is three bases long. If we represent the frames as 0, +1 and +2 then +3 = 0, +4 = +1 etc, and by the same logic -1 = +2 and -2 = +1 etc. If one considers all of the possible reading frames in both directions the total number is 6.

Thus the total number of fortuitous ORFs length > 100 = 6 x 4.2 = 25

A tyypical bacterial genome contains about 4,000 real genes (that have evolved to have an ORF of at least 200 codons). Thus if one analyzed the genome looking for ORFs of length >200 condons, one might find about 4,000 such ORFs with all but 25 representing real genes. This is an acceptable false positive rate, and many of the gene finding programs use a cutoff of length > 200 codons. As you will see later in the course, real ORFs and fortuitous ORFs can often be distinguished by comparing genome sequences of related species.

Identifying genes in DNA sequences from higher organisms is usally more difficult than in bacteria. This is because in humans, for example, gene coding sequences are separated by long sequences that do not code for proteins. Moreover, genes of higher eukaryotes are interrupted by **introns**, which are sequences that are spliced out of the RNA before translation. The presence of introns breaks up the open reading frames into short segments making them much harder to distinguish from non-coding sequences. The maps below show 50 kbp segments of DNA from yeast, Drosophila, and humans. The dark grey boxes represent coding sequences and the light grey boxes represent introns. The boxes above the line are transcribed to the right, whereas the boxes below are transcribed to the left. Names have been assigned to each of the identified genes. Although the yeast genes are much like those of bacteria (few introns and packed closely together), the Drosophila and human genes are spread apart and interrupted by many introns. Sophisticated computer algorithms were used to identify these dispersed gene sequences.


Saccharomyces cerevisiae (yeast)
YFL046W
RGD2
YFL040W
FET5
TUB2
RP041
YFL034W
YFL030W
HAC1
STE2
0
50
SEC53
YFL044C
YFL042C
ACT1
YPT1
MOB2
RPL22B
RIM15
CAK1
CAF16
GYP8
BST1
EPL1
Drosophila melanogaster
syt
CG15400
CG3131
0
50
CG16987
CG2964
CG3123
Human
GATA1
HDAC6
LOC139168
0
50
PCSK1N