DNA as a Mass-Storage Device

Robert J. Robbins

Johns Hopkins University
rrobbins@gdb.org

Goals of the Genome Project

Sequence the Genome

- equivalent to obtaining an image of a mass-storage device

Map the Genome

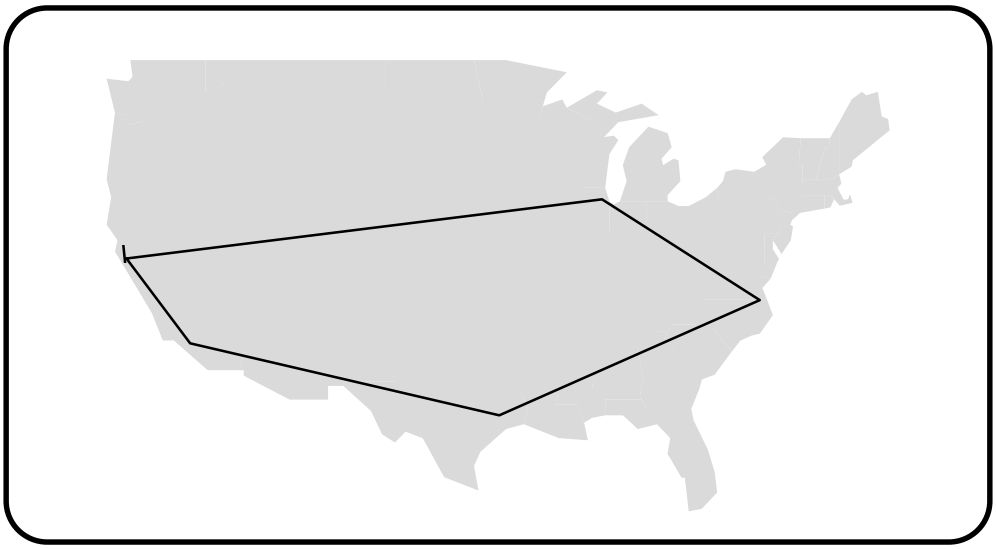
- equivalent to developing a file-allocation table for the mass-storage device

Understand the Genome

- equivalent to reverse engineering the files on the mass-storage device all the way back to design and maintenance specifications

Sequencing the Genome

Getting the Sequence



Obtaining one full human sequence will be a technical challenge. If the DNA sequence from a single human sperm cell were typed on a continuous ribbon in ten-pitch type, that ribbon could be stretched from San Francisco to Chicago to Washington to Houston to Los Angeles, and back to San Francisco, with about 60 miles of ribbon left over.

The amount of human sequence currently sequenced is equal to less than one-third of that left-over 60-mile fragment. We have a long way to go, and getting there will be expensive. Computers will play a crucial role in the entire process, from robotics to control experimental equipment to complex analytical methods for assembling sequence fragments.

year	per base cost	budget	year	cumulative	percent completed
1995	$0.50	16,000,000	10,774,411	10,774,411	0.33%
1996	$0.40	25,000,000	21,043,771	31,818,182	0.96%
1997	$0.30	35,000,000	39,281,706	71,099,888	2.15%
1998	$0.20	50,000,000	84,175,084	155,274,972	4.71%
1999	$0.15	75,000,000	168,350,168	323,625,140	9.81%
2000	$0.10	100,000,000	336,700,337	660,325,477	20.01%
2001	$0.05	100,000,000	673,400,673	1,333,726,150	40.42%
2002	$0.05	100,000,000	673,400,673	2,007,126,824	60.82%
2003	$0.05	100,000,000	673,400,673	2,680,527,497	81.23%
2004	$0.05	100,000,000	673,400,673	3,353,928,171	101.63%

Understanding the Genome

Reverse Engineering Codes

Understanding the genome involves the equivalent of reverse engineering binary codes for an unknown computer system. This is nearly impossible for a single program, but comparive analyses of similar programs can provide a start.

Suppose you had two programs, one of which caused a computer to undergo a cold boot, the other a warm boot. A comparison of these programs would give some small insights into the workings of that computer.

WARMBOOT: `BA 40 00 8E DA BB 72 00 C7 07 00 00 EA 00 00 FF FF`

COLDBOOT: `BA 40 00 8E DA BB 72 00 C7 07 34 12 EA 00 00 FF FF`

Alignments of the codes can provide insights into regions of common function:

WARMBOOT: `BA 40 00 8E DA BB 72 00 C7 07 00 00 EA 00 00 FF FF`

COLDBOOT: `BA 40 00 8E DA BB 72 00 C7 07 34 12 EA 00 00 FF FF`

Reverse Engineering Codes

Similar methods could be used to analyze programs that caused messages to be displayed on screen.

Assume that you have the sources for four short programs, each of which causes a short message to be written to the screen:

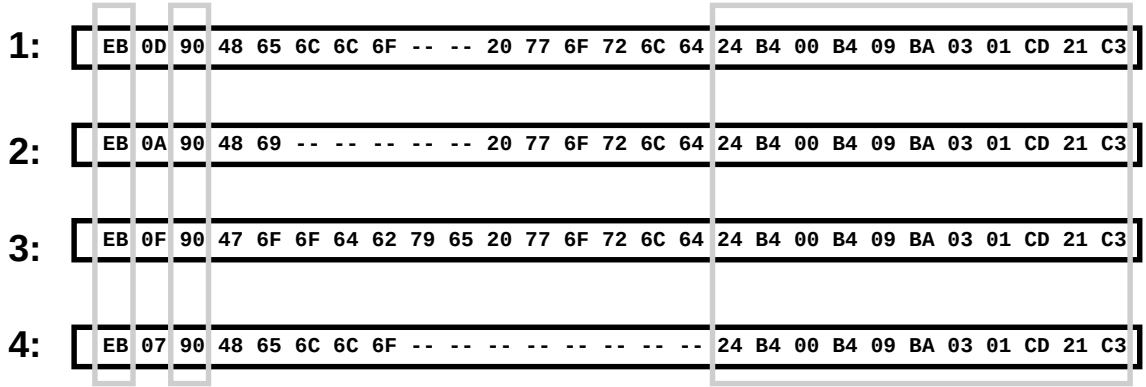
1 = Hello world

2 = Hi world

3 = Goodbye world

4 = Hello

Aligning the sequences (inserting blanks where necessary) allows the detection of common features:



Reverse Engineering Codes

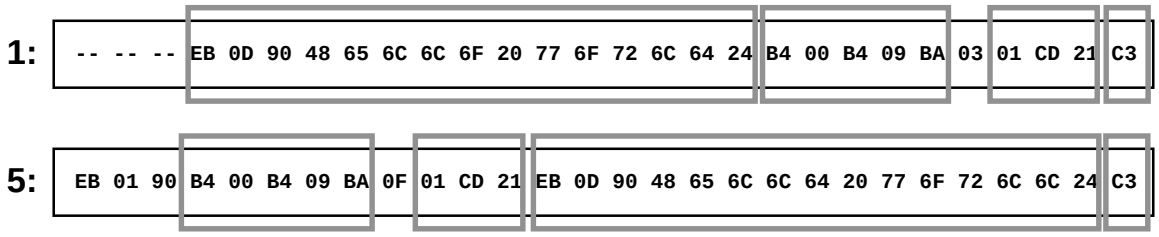
Now, suppose you get a fifth program, that writes the same "Hello world" message as the first program, but which has different binaries.

At first, the sequences appear fairly different:

1: `EB 0D 90 48 65 6C 6C 6F 20 77 6F 72 6C 64 24 B4 00 B4 09 BA 03 01 CD 21 C3`

5: `EB 01 90 B4 00 B4 09 BA 0F 01 CD 21 EB 0D 90 48 65 6C 6C 64 20 77 6F 72 6C 6C 24 C3`

Again, sequence similarities can provide clues...



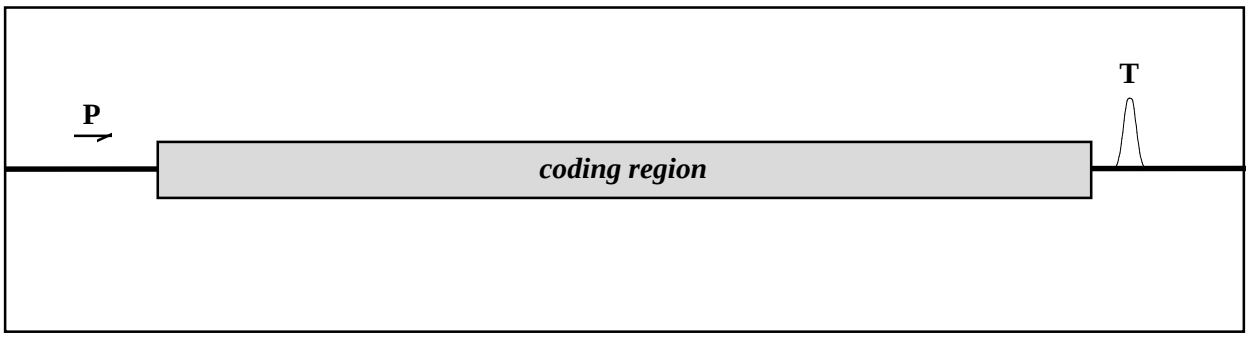
Reverse Engineering Codes

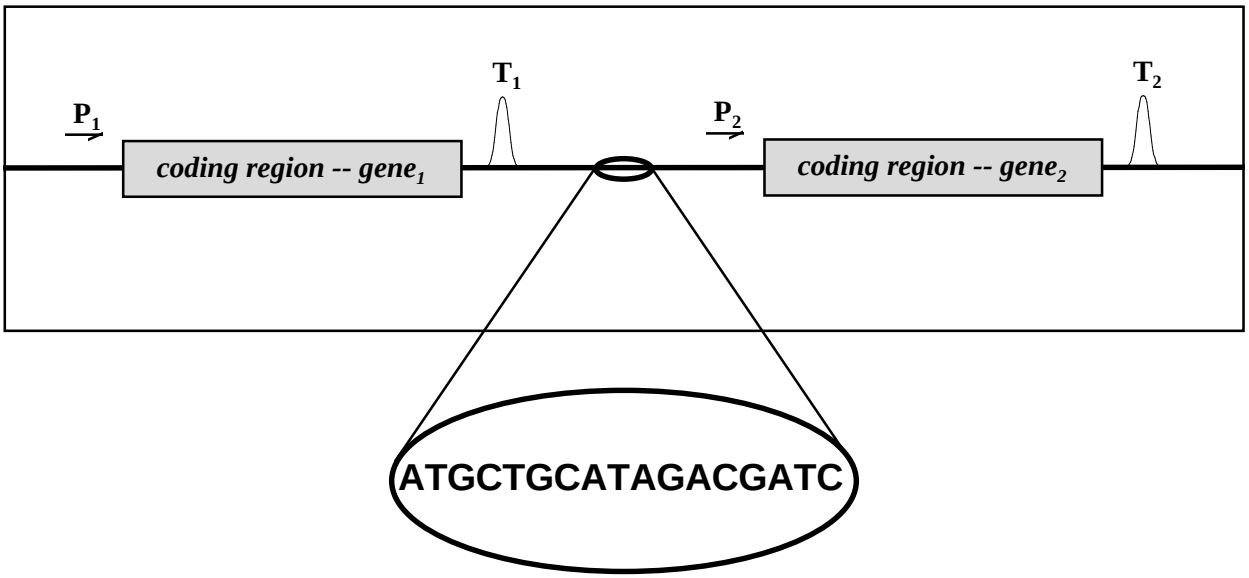
This kind of comparative technique is used in biology.

gene name	DNA sequence near transcription initiation site					
lacZ	---ccaggc	TTtACA	ctttatgcttccggctcg-	TATgtT	--gtgtgga---	
malT	--tcatcgc	TTGcat	tagaaaggtttctggcc--	gAcctT	--ataacca---	
araC	--atccatg	TgGACt	tttctgccgtgattata--	gAcAcT	tttgttacg---	
galP1	----catgt	cacACt	tttcgcatctttgttatgc	TATggT	--tatttca---	
deoP2	----gtgta	TcGAag	tgtgttgcggagtagatgt	TAgAAT	--actaaca---	
cat	gatcggcac	gtaAgA	ggttccaactttcac----	cATAAT	-gaaataag---	
tnaA	-tttcagaa	TaGACA	aaaactctgagtgtaa---	TAatgT	--agcctcg---	
araE	----ccgac	cTGACA	cctgcgtgagttgttcacg	TATttT	ttcactatg---	
consensus	---------	TTGACA	------------------	TATAAT	----------	
		- 35			- 10	

Mapping the Genome

The Simplistic View of a Gene



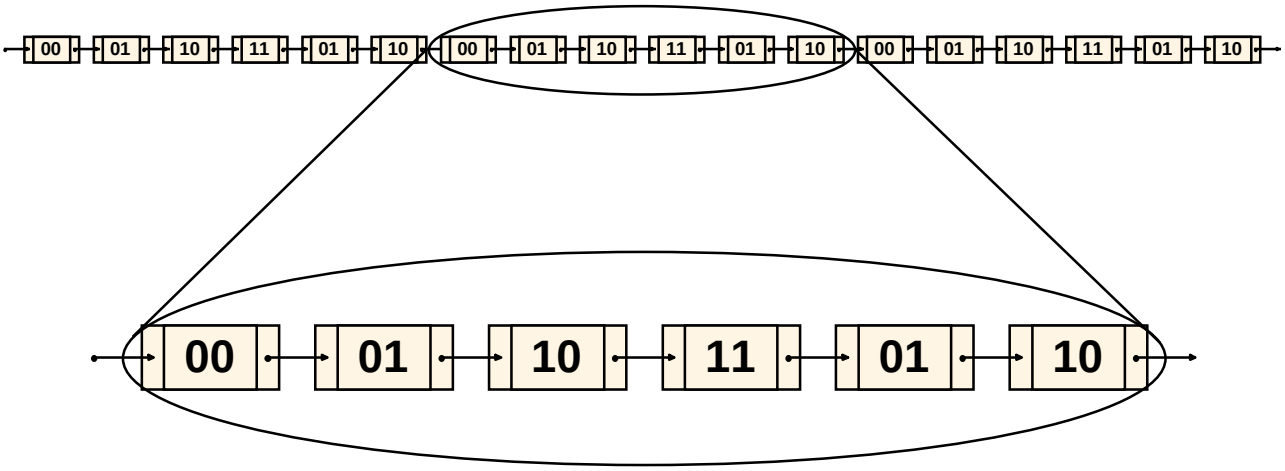


Genes are sequences of DNA. Mapping the genome involves identifying and locating specific functional regions of DNA.

DNA as a Mass-Storage Device

Mass Storage System:

- **Underlying primitive structure is linked list, not physical medium**
- **List has polarity**
- **Addressing is associative, not physical**
- **No content restriction on linear order of data**

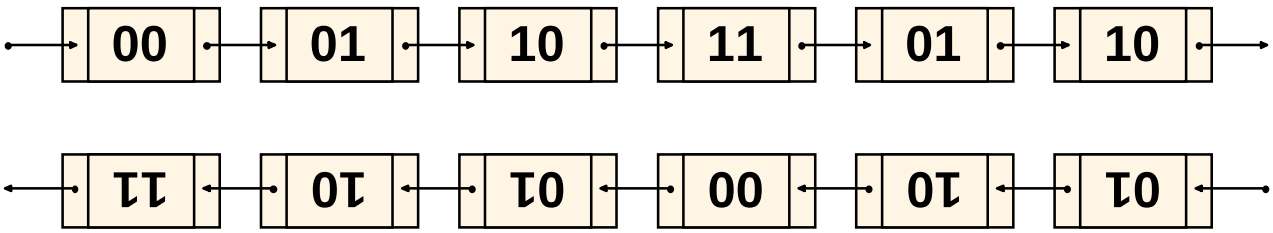


Understanding how DNA is constructed is necessary for understanding how it functions.

DNA as a Mass-Storage Device

Redundant Mass Storage System:

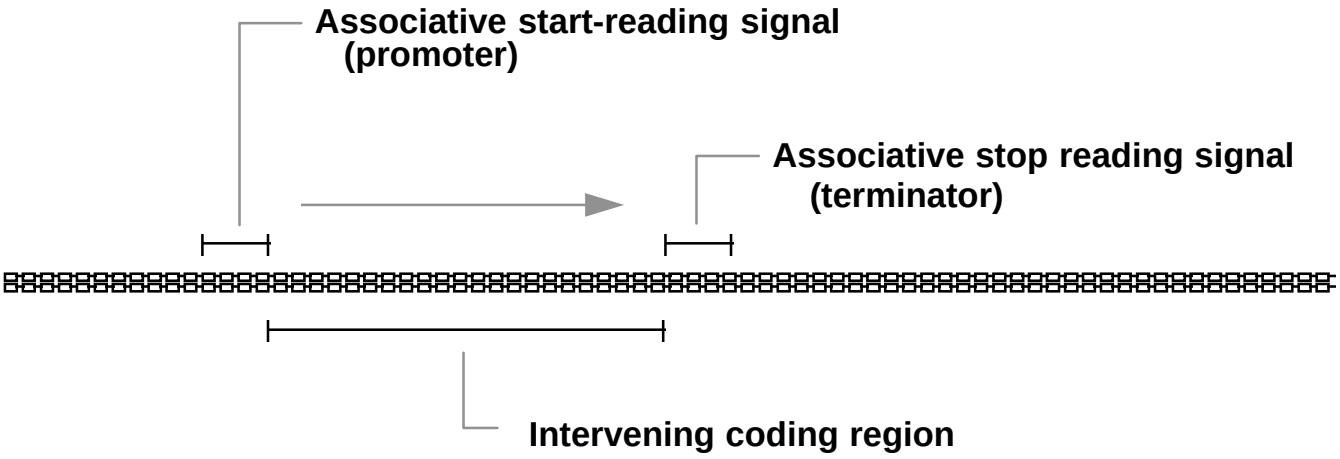
- **Full structure is dual linked list**
- **Lists have opposite polarity**
- **Total content restriction on paired data**
- **Either list can be data, the other redundant**



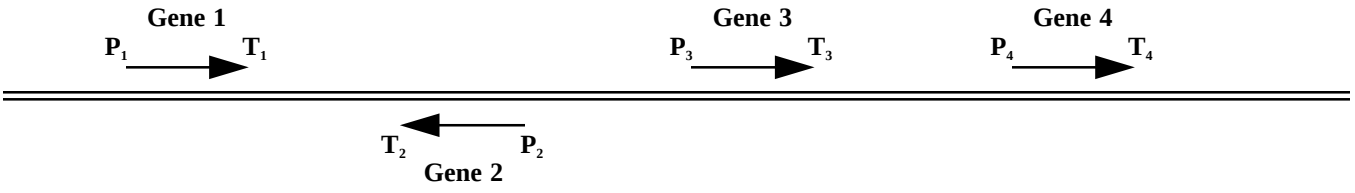
DNA has much in common with a mass-storage device, but a mass-storage device based on an underlying linked list, not a physical system with spatial addresses independent of what is being addressed.

DNA as a Mass-Storage Device

An expressed region of the genome consists of a coding region flanked by an upstream associative start signal and a downstream associative stop signal.

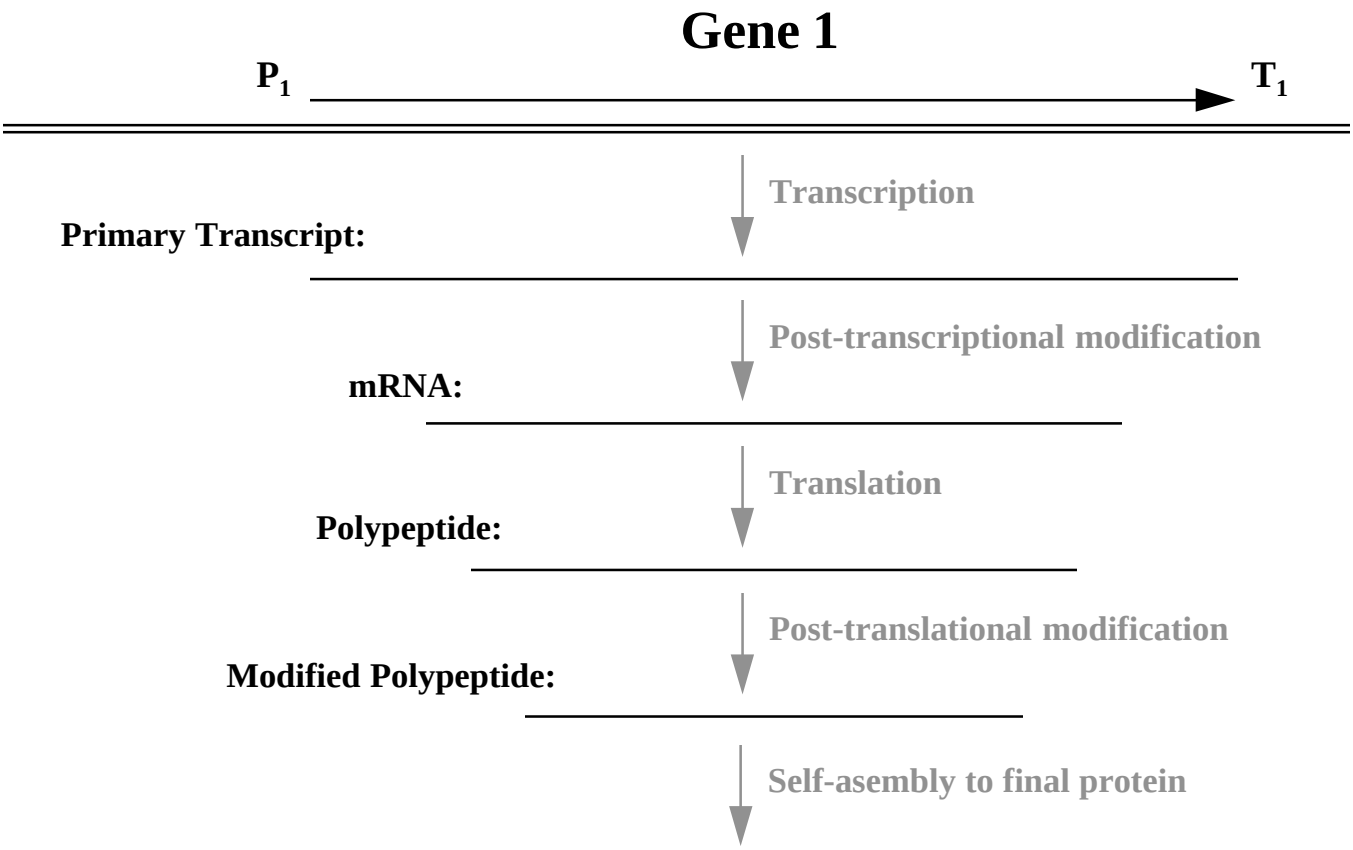


Expressed regions may occur on either strand of the DNA double helix. The two strands are transcribed in opposite directions.



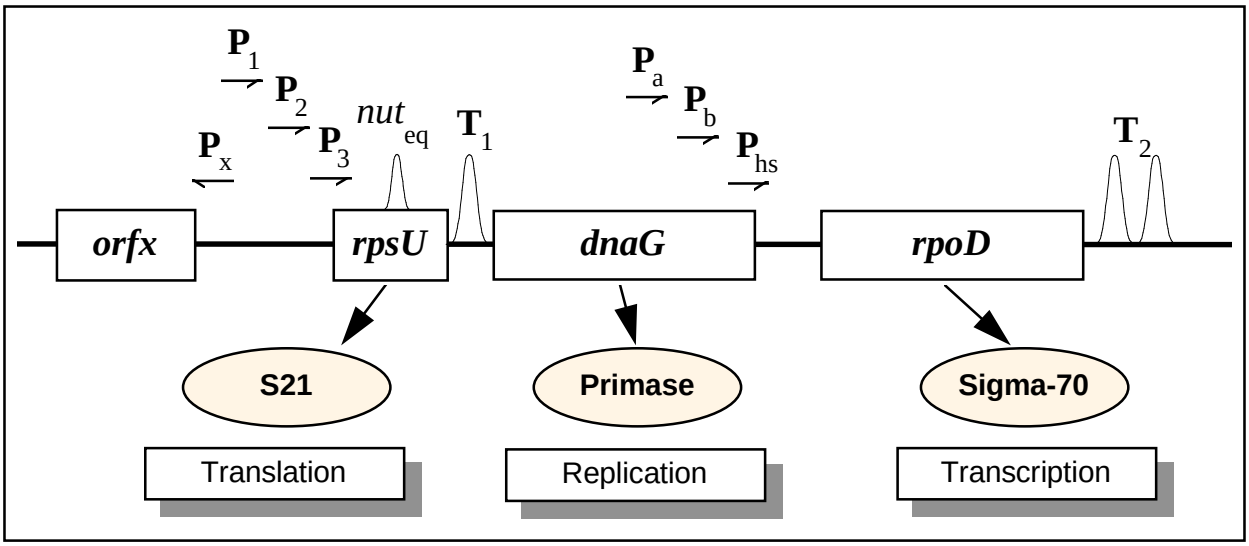
DNA as a Mass-Storage Device

The expression of an individual region occurs as a series of steps...



The MMS Operon

Escherichia coli

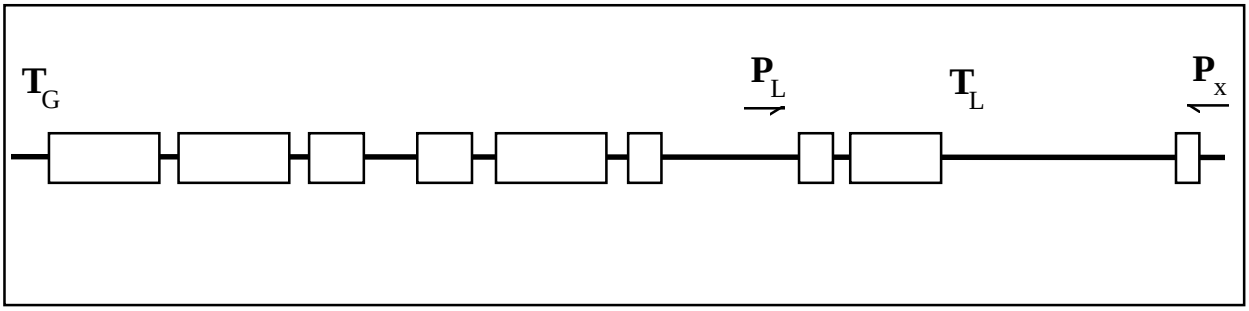


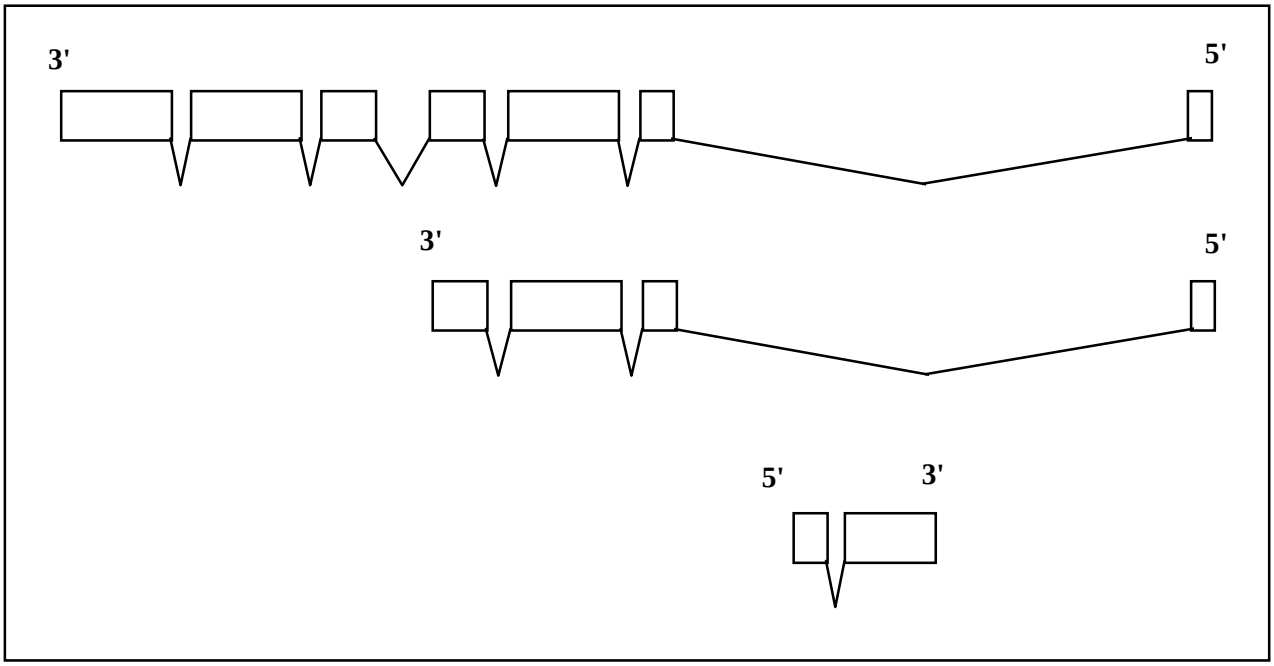
Lupski, J.R., Godson, G.N., 1989, DNA ->DNA, and DNA -> RNA->Protein: Orchestration by a single complex operon, *BioEssays*, 10:152-157.

In practice, many genomic regions exhibit considerable complexity.

The Gart/Lcp Locus

Drosophila melanogaster

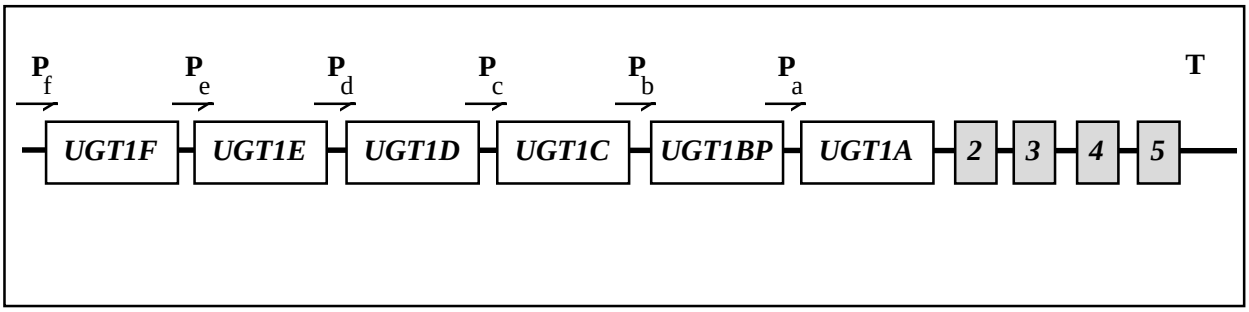


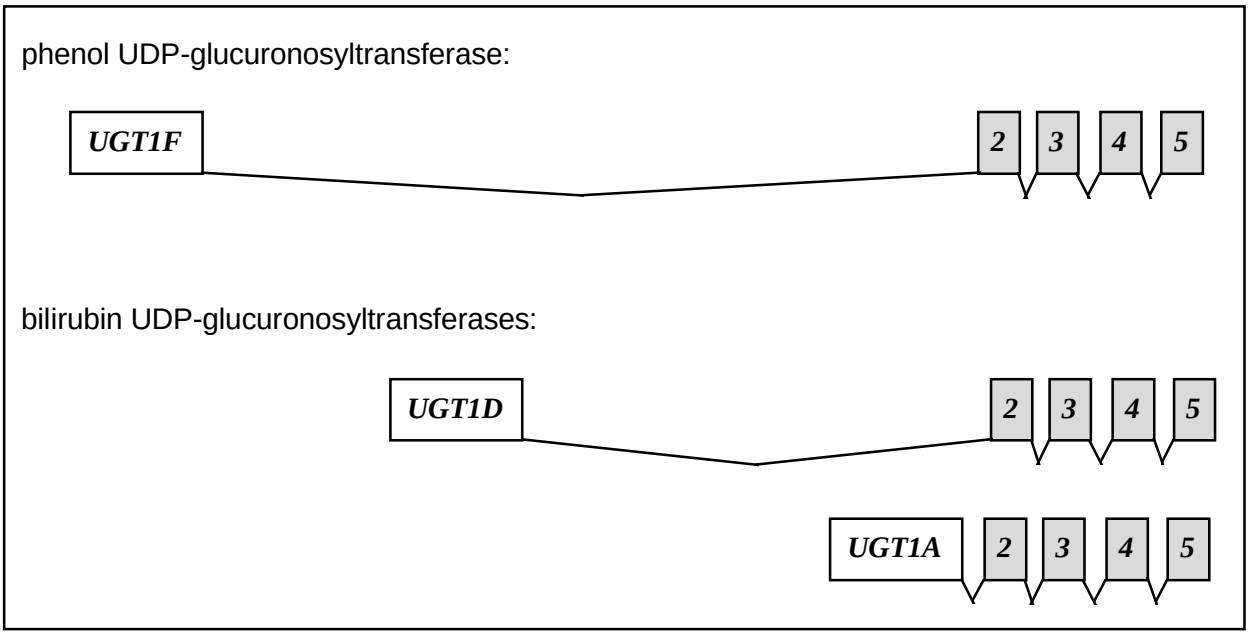


Henikoff, S., Keene, M.A., Fechtel, K., and Fristrom, J.W., 1986, Gene within a gene: Nested *Drosophila* genes encode unrelated proteins on opposite strands, *Cell* 44:33.

Genes can be broken up with non-coding sub-sequences (introns) interspersed among coding sub-sequences (exons). An entire gene may lie within an intro of another gene.

The UGT1 Locus



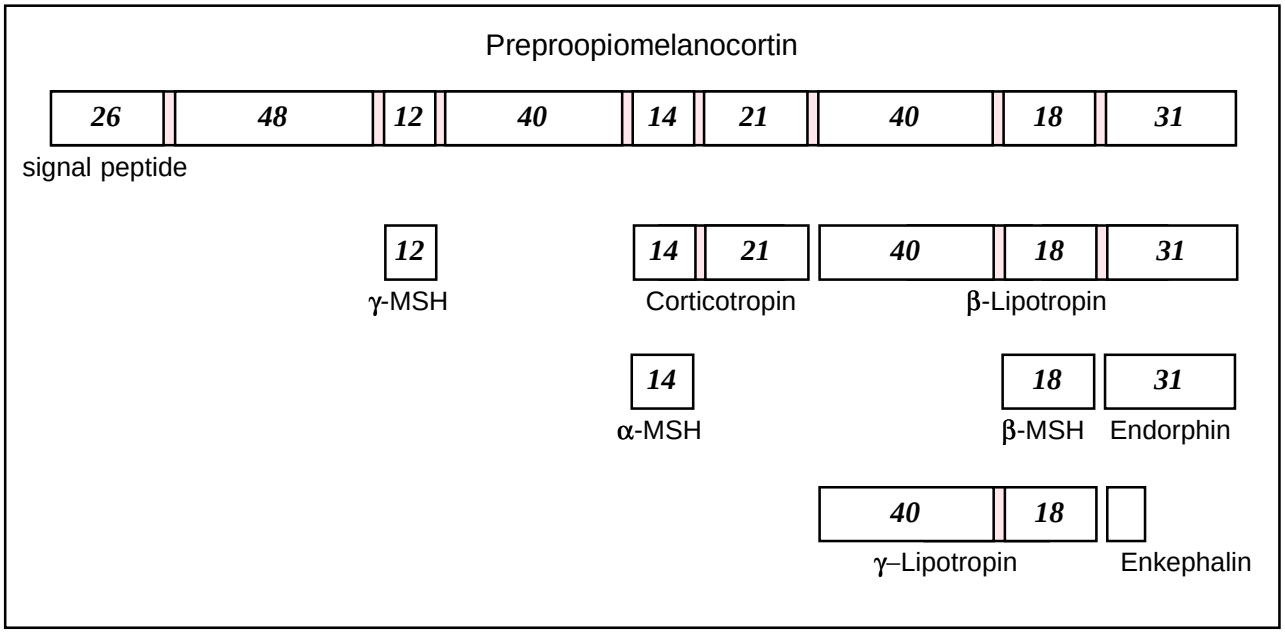


Ritter, J.K., Chen, F., et al., 1992, A novel complex locus *UGT1* encodes human bilirubin, phenol, and other UDP-glucuronosyltransferase isozymes with identical carboxyl termini, *J. Biol. Chem.* 267:3257.

The human UGT1 locus seems to be an example of a nested gene family.

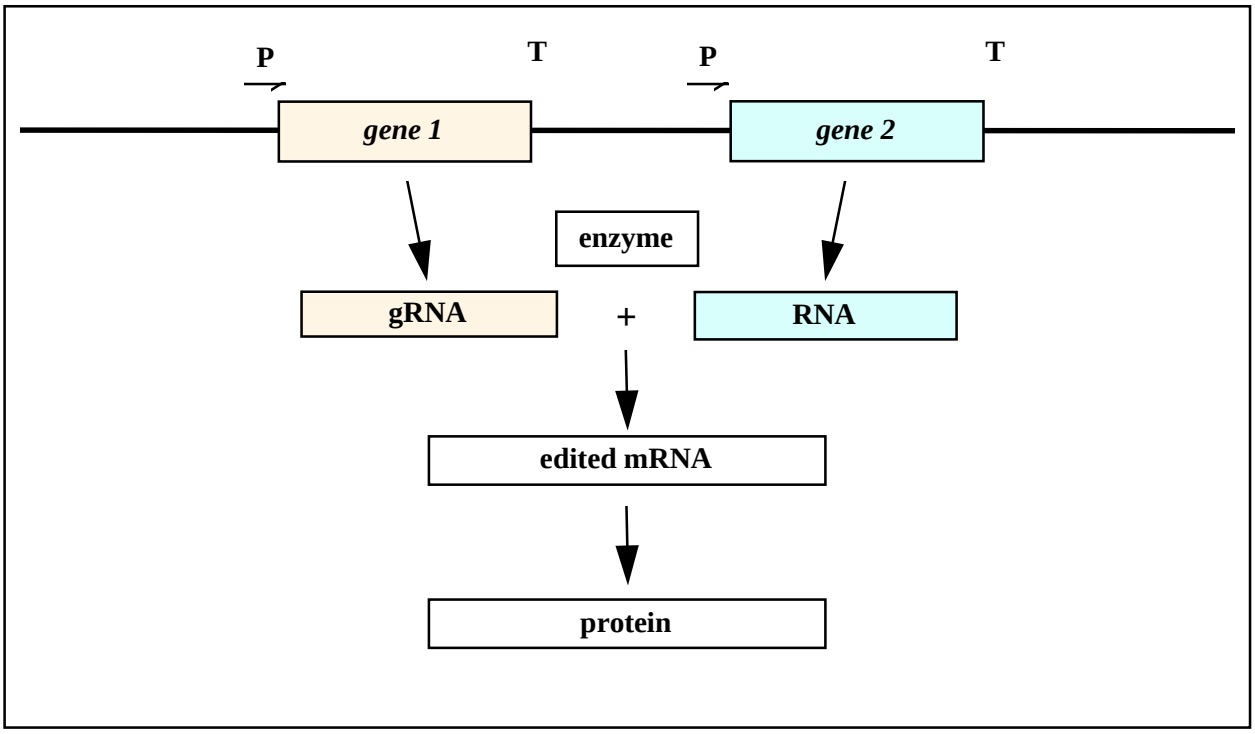
The POMC Locus Products

Homo sapiens



Some regions of the genome produce multiple proteins from a single polypeptide precursor.

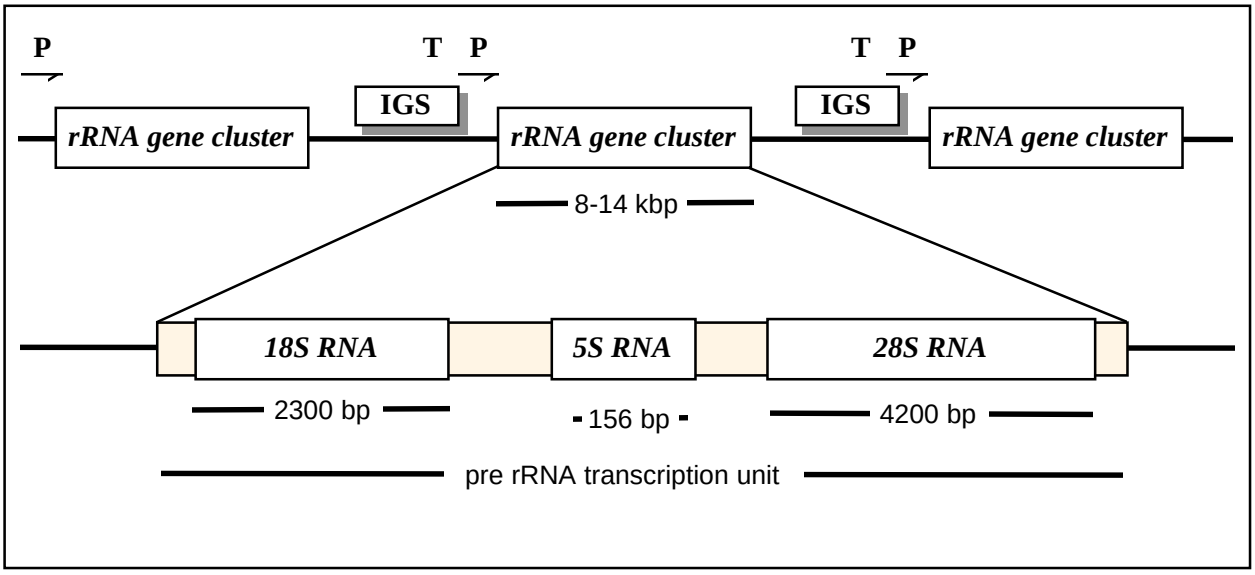
Guide RNAs



Complex interactions among transcipts from different genomic locations may be required to produce a single protein.

The rRNA Loci

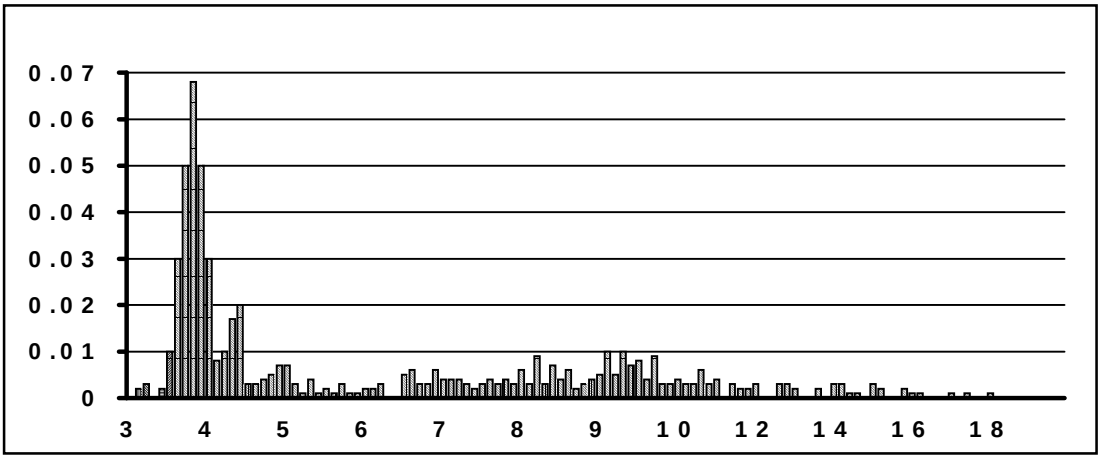
Homo sapiens



Some regions of the genome are characterized by multiple tandem repeats of the "same" gene.

D14S1 is a VNTR Locus

Frequency of PstI fragment sizes (kb)



Balazs, I., Neuweiler, J., Gunn, P., Kidd, J., Kidd, K.K., Kuhl, J., and Mingjun, L., 1992, Human population genetic studies using hypervariable loci, *Genetics*, 131:191-198.

Regions of the genome may differ in length among normal individuals.