Sequence Analysis & Genome Assembly

Learning Goals:

To work with a physical model of DNA and RNA in order to help you to understand:

- rules for both DNA & RNA structure
- transcription including promoters & terminators
- translation including start & stop codons

To work with a computer program in order to help you to understand:

- how to begin to analyze nucleic acid sequences
- the structure of a gene and the effects of mutations

Introduction:

This week's topics in molecular biology will include gene expression and sequence analysis using a computer. The field of bioinformatics is based on sequence analysis and using computer algorithms to answer biological questions. Gene expression is a complex process that allows cells to respond to the environment through the genome. The genome contains the information necessary to produce proteins through the coupled process of transcription and translation. Transcription is the process in which an mRNA molecule is made by RNA polymerase from DNA. Translation is the process in which an mRNA molecule is read by ribosomes (a composite of rRNA and proteins) in order to place amino acids together in a chain to form a polypeptide. Currently, the sequences of genes are analyzed using computer tools, a process called sequence analysis.

	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 START	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

Figure 1: Codon Table

How to decode the genome using a simple eucaryotic gene:

Deciphering the Genetic code requires the following steps: 1) Start transcription, 2) Stop transcription, 3) mRNA processing, 4) Start translation, and 5) Stop translation. A eucaryotic gene contains sequences that specify each of these steps. Currently, research in biology is underway in order to elucidate these sequences. Analyzing the sequence of a protein coding gene sequence will allow you to decipher the resulting protein.

1) Start Transcription: What is the first base of the resulting mRNA?

The actual gene contains more bases than the resulting mRNA. This is because a gene contains non coding regions, such as the promoter. The function of the promoter region is to regulate the recruitment of the RNA polymerase. The TATAA box is sequence motif found in the promoter regions of many genes. For our purposes, the mRNA will begin at the base immediately following the TATAA box.

2) Stop Transcription: What is the last base in the resulting mRNA?

The terminator region of a gene helps to regulate termination of transcription by the RNA polymerase. For our purposes the mRNA will end at the base immediately preceding the GGGGG.

3) mRNA processing

The major processing events are: 1) Introns are spliced out, 2) a poly A tail is added to the 3′ end and, 3) a modified Guanosine triphosphate is added to the 5′ end. For our purposes, introns begin with GTGCG and end with CAAAG

4) Start Translation

Translation does not begin at the very first codon of an mRNA! The AUG codon initiates translation.

5) Amino Acid sequence produced

Each successive codon is read and translated using the codon table until a stop is reached

6) Stop Translation

The amino acid sequence stops after an UAA, UAG, or UGA is reached.

As many genomes have been sequenced and are currently being analyzed by scientists around the world, a standard has been set up for DNA sequences. When a double stranded DNA molecule is written down, it is assumed that the sequence is read from left to right as the 5′ to 3′.

For example: ATGATGCGTAG

Means 5′ ATGATGCGTAG 3′
3′ TACTACGCATC 5′

Remember, the RNA polymerase makes RNA in the 5′ to 3′ direction, thus reading the 3′ to 5′ strand (the BOTTOM strand as written above). This ultimately produces a nucleic acid that is the same as the "top" strand, only with Us replacing Ts.

In Class preparatory questions

The answers to the following questions will be reviewed at the beginning of Lab. It is in your best interest as a student to attempt them prior to Lab, though not required.

1. β-thalassemia is an autosomal recessive genetic disease. This question deals with three alleles of the β-globin gene: the normal allele and two β–thalassemia alleles. The three alleles are described below:

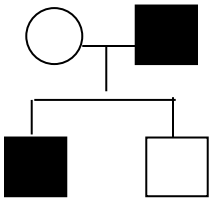
β-globin gene

Allele	**DNA sequence** (◎ means mutated to)	**contribution to phenotype**
H	normal	normal - dominant
h_1	codon 15 TGG ◎ TAG (stop)	β-thalassemia (recessive)
h_2	codon 17 AAG ◎ TAG (stop)	β -thalassemia (recessive)

1) What type of mutation is present in the h_1 and h_2 alleles (missense, nonsense, frameshift)?

2) Based on your knowledge of protein structure, provide a plausible explanation for why the β–globin protein encoded by the h_1 and h_2 alleles is non-functional.

3) Consider the following pedigree. Filled symbols represent individuals with β–thalassemia.



a) For each individual in the pedigree, write his or her genotype next to his or her symbol using the allele symbols defined above. Assume that the h_1 allele is causing β–thalassemia in this family.

b) What is the chance that the couple's next child will have β-thalassemia?

4) What would the phenotype (normal or β–thalassemia) of an individual of genotype h_1h_2 be? Explain your reasoning.

Sequence Analysis

Procedure :

1. Enter the following sequence into the Gene Explorer Program (Brian White, PhD)

TTGTA𝕋AACGT𝔾AT𝔾AAACCℂGAT𝔸TTGC𝔾TGCGTAGCTAGCTCAAAGCTTT𝕋AAGATCGG𝔾GGGCG
ATC

1 2 3 4 5 6 7 8

2. Create a hypothesis for how each of the mutations (a-h) above effects the expression of our gene. Test your hypothesis by mutating your gene and analyzing the output from the Gene Explorer Program.

	HYPOTHESIS	RESULTS
1		
2		
3		
4		
5		
6		
7		
8		

This page has been left blank intentionally.