

A 3DMD Paper BioInformatics Activity©

The β -Globin Gene Exercise

Teacher Notes

Introduction

This β -Globin Gene Exercise© serves as an example of how we can encourage our students to ask thoughtful questions. Our approach is to challenge students with a specific task — in this case finding the β -globin gene — in the absence of the explanations that normally

The problem with the U.S. system of education is that we reward students for knowing the answers to questions they have never asked. — Source Unknown

precede an exercise such as this. It is important to let your students work on this challenge and struggle a little from time to time as they discover the answers to:

- which of the two strands of DNA is the coding strand?
- why are there three different amino acid sequences listed under the DNA?
- what does the * symbol mean?

As your students ask questions and discuss possible answers, we believe they will engage in a memorable experience and gain a deeper understanding of the key concepts conveyed by the β -Globin Gene Exercise©.

The most difficult part may be for you to step back and let your students learn on their own. In our eagerness to share how amazing molecular biology is, we risk depriving our students of the joy of discovery.

Background Information

Where did this sequence data come from? This nucleotide sequence data is contained within the human β -globin region on Chromosome 11 as reported in the GenBank file 455025. This complete file consists of 73,308 nucleotides and indicates that the β -globin gene comprises three separate coding sequences:

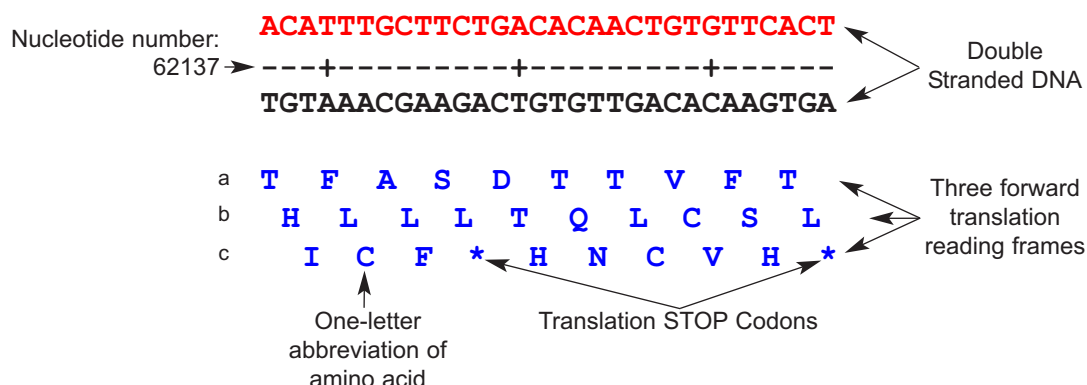
62,187 to 62,278 (Exon 1); 62,409 to 62,631 (Exon 2); 63,482 to 63,610 (Exon 3)

How was the data (Map) obtained? We copied the nucleotide sequence of the β -globin gene from nucleotide 61,981 to 63,773, and then used the *translation tool* that is available online at <http://bio.lundberg.gu.se/edu/translat.html> to generate this output format. We chose to report both strands of DNA and the 3 forward reading frames. We thought it would be confusing to show all six reading frames (three forward and three reverse).



Teacher Notes continued

How do we interpret the Map of the β -Globin Gene® shown below?



- The transcription start site begins at nucleotide 62,137. The map shows both strands of the DNA sequence. Remember that the two strands of DNA are anti-parallel (Watson-Crick double helix). The **top red strand** runs 5' >>> 3', left to right and the **bottom black strand** runs 3' >>> 5', left to right.
- When expressed, RNA polymerase uses the bottom strand as a template to make an RNA copy that corresponds with the sequence shown in the **top red strand** — with Us replacing Ts. The **top strand** represents the *non-template strand* and the **bottom strand** represents the *template strand*.
- The amino acid sequence encoded by DNA is shown in the **three blue lines**. (The one-letter abbreviations of the amino acids are used: T represents threonine; H, histidine; I, isoleucine, and so forth.) The * symbol represents one of three translation STOP codons (UAA, UAG or UGA).

Discussion

Why are there three lines of amino acid sequences? You see three lines since the triplet genetic code can be read in **three reading frames, labelled: a, b or c**. In the diagram above, in the **a reading frame**, the ribosome reads the triplet ACA as T, threonine. In the **b reading frame**, the ribosome shifts one nucleotide and reads the triplet CAT as H, histidine.

Which reading frame is used to translate the β -globin protein? And where does the translation begin? From the GenBank sequence file, we know the coding sequence starts at nucleotide 62,187. Since we know that all proteins begin with M, methionine, we expect to find the translation START codon, AUG, at 62,187. We do, except T replaces U, since we are really seeing the DNA sequence. This establishes the **c reading frame** as the one used to translate the β -globin gene. Therefore, the amino acid sequence of the β -globin protein begins MVHLTPEE and continues for 61 codon triplets before there is a STOP codon in the **c reading frame**.

But, we encounter two problems. First, the β -globin protein contains 146 amino acids. Where is the rest of the gene that encodes the protein? Second, of the 61 amino acids encoded in this **c reading frame**, only the first 31 correspond to the known amino acid sequence of β -globin. Confusing? Yes, this is where your students will struggle to figure out what is going on.

Introns and Exons

What is going on, of course, is that eukaryotic genes are split into segments called **exons** (for **expressed**) and **introns** (for **intervening** sequences). As your students continue to examine the Map of the β -Globin Gene®, they will see that the amino acid sequence of the β -globin protein resumes at nucleotide 62,409

Teacher Notes continued

and continues until nucleotide 62,631, where they will encounter the second intron. The third and last exon stretches from nucleotide 63,482 to 63,610.

Your students will also find in this exercise that the second exon is translated in the ***a reading frame***, and the third is translated in the ***b reading frame***.

Does this mean that the ribosome actually switches from reading frame c to a to b as it translates this mRNA into protein? No. Before the mRNA is translated into protein, the introns are spliced out. As the two introns are removed, the three exons join into one continuous coding sequence, in one reading frame. (The reading frame appears to have switched when looking at the Map of β -Globin Gene®, but simply because the number of nucleotides in each of the two introns is not a multiple of three.)

Can introns split the codon for an amino acid or do they only occur between codons? Look carefully at the first intron in the β -globin gene for the answer.

How are the introns spliced out of the pre-mRNA? **Spliceosomes**, which are complexes of proteins and small nuclear RNAs, are the molecular machines that do the splicing.

How are splice junctions recognized by the spliceosome? Each splice junction starts with GU and ends with AG. For an animation that illustrates the process, as well as internal recognition sequences, go to <http://www.sumanasinc.com/webcontent/animations/content/mRNAsplicing.html>. (Please see *Features of the Teacher's Map* for details.)

Advanced Topics: Promoters and Polyadenylation Sequences

Upstream (5' to the start of the gene coding sequence) are regulatory regions that determine how frequently a gene is transcribed. The β -globin gene has gene-specific promoters in addition to promoters common to all genes. Each promoter has a consensus sequence that lists the most prevalent base (or bases) found at each position within the promoter sequence. Genes with promoters that are identical to the consensus sequence are transcribed frequently. Those with promoters that differ from the consensus are transcribed less often. Students can search for the TATA box, located 20-30 base pairs upstream from the transcription start site. (Note that the gene transcript is LONGER than the coding sequence, since it includes sequences important in the regulation of protein synthesis.) The consensus sequence for the TATA box is TATA(A/T)A(T/A). The notation indicates that in position 5, A is found more often than T, and in position 7, T is found more often than A.

How close is the β -globin gene TATA box to the consensus sequence? With the exception of the first base in the sequence (a C in β -globin), the TATA box is identical to the consensus sequence. This indicates that the β -globin gene is transcribed frequently. Students can also explore additional promoter sequences to determine how close they are to the consensus. (Please see the *Features of the Teacher's Map* for details.)

The stop codon at the 3' end of the gene is a signal for termination of protein synthesis. Transcription continues well beyond this point — often for hundreds of bases. Polyadenylation signals in this region identify the location at which the transcript is cleaved and a poly-A tail is added. The polyadenylation sequence AAUAAA is highly conserved among mammals. Changes in any one base from this consensus result in much lower polyadenylation rates. Genes often have multiple polyadenylation sequences; typically the sequence furthest downstream from the stop codon matches the consensus and is recognized as the cleavage and polyadenylation sequence.





Teacher Notes continued

How many polyadenylation sequences are there in the β -globin gene? There are two polyadenylation sequences, but only one is an exact match to the consensus sequence. The other sequence, AUUAAA, is the most common alternate polyadenylation sequence. Students can explore additional sequences involved in polyadenylation. (Please see *Features of the Teacher's Map* for details.)

Summary

When RNA polymerase transcribes the β -globin gene, it makes an RNA that has the same sequence as the top strand of DNA in the globin map, except that the **Ts** are replaced with **Us**. This initial RNA copy of the gene is not yet messenger RNA (mRNA). It is referred to as precursor-mRNA (or pre-mRNA). In the **splicing process**, the intron sequences are cut out of pre-mRNA, and the three exons are joined to make mature mRNA. The ribosome begins translating mature mRNA in the **c reading frame**, and continues in this reading frame throughout the entire mRNA.

Questions

You may want to ask the following questions, if they haven't occurred to your students:

Why are eukaryotic genes split like this?

Why would higher animals adopt this complicated mechanism to encode the information for making proteins? (Some genes are known to have hundreds of introns.)

Are all introns in all proteins spliced out by the same spliceosome structure or does each gene have its own unique spliceosome?

What happens if the spliceosome makes a small mistake, such as inserting or deleting a single nucleotide from the final spliced mRNA?

Can mutations in the DNA destroy a splice site? What is the consequence for the protein?

Can mutations in the DNA create new splice sites? If so, what is the consequence for the protein?

Are any human diseases caused by aberrant splicing of genes? (Google **hemoglobinopathies**.)

Is a given pre-mRNA always spliced in exactly the same way or can a pre-mRNA be spliced in two different ways to produce two different mature mRNAs that are translated into two different proteins?

References

The book, *Principles of Medical Genetics*, by T. Gelehrter and F. Collins (Williams and Wilkens) provides a good overview of the genetics of hemoglobinopathies.

An Internet search for **hemoglobinopathies** or **thalassemia** will link you to a wealth of information on these topics.

Please send questions or comments to Tim Herman at herman@msoe.edu or Margaret Franzen at franzen@msoe.edu.