

A 3DMD Paper BioInformatics Activity©

The β -Globin Gene Exercise

Features of the Teacher's Map

Normal Features of the β -Globin Gene

Nucleotide 62,137: Transcription Start Site

When this gene is transcribed by RNA polymerase, the mRNA starts at nucleotide 62,137. A multi-protein transcription complex (including RNA polymerase) assembles on the **promoter** region immediately upstream from 62,137, such that mRNA synthesis begins at this nucleotide. Note that the **translation start site (AUG)** is located 50 nucleotides **downstream**, at nucleotide 62,187.

Nucleotides 62,187 to 62,278: Exon I (92 nucleotides, 30 2/3 codons)

The Translation Start Site (AUG) is located at nucleotides 62,187- 62,189.

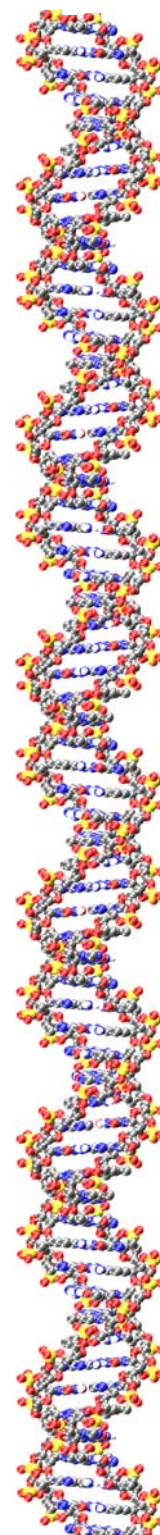
All proteins begin with the amino acid methionine, Met, encoded by nucleotides AUG. This **rule** is a consequence of the mechanism that cells use to begin protein synthesis; a special Met-tRNA initiates assembly of the two subunits of the ribosome at the beginning of each mRNA. (In β -globin protein, and many other proteins, this N-terminal Met is cleaved from the protein soon after it is synthesized on the ribosome. This leaves the second amino acid, valine, as the N-terminal amino acid of the mature β -globin protein.)

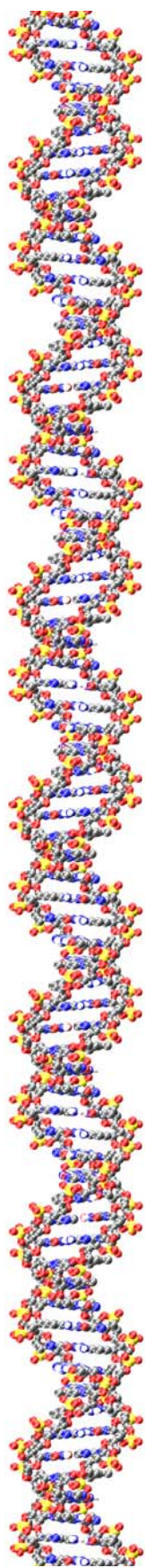
Nucleotides 62,279 to 62,408: Intron I (130 nucleotides)

All introns begin with the nucleotides GU and end with the nucleotides AG. These two invariant dinucleotides are not enough in themselves to define a donor splice site (at the 5' end of the intron) and an acceptor splice site (at the 3' end of the intron). The consensus donor and acceptor splice site sequences are shown on the Map of the β -Globin Gene©. Splice sites that conform closely to these consensus sequences are very efficient, while other sites that do not match the consensus as well are less efficient. As a result, the level of expression of a given gene can be controlled, in part, by the efficiency of splicing of its pre-mRNA. (See *Mutations in the β -Globin Gene* for more information about mutations that affect normal splice sites or create new splice sites in either exons or introns.)

A shift in reading frames? The first intron of the human β -globin gene is 130 nucleotides long. Since this number is not a multiple of 3, the reading frame for the β -globin gene appears to shift from the **c reading frame** in Exon I to the **a reading frame** in Exon II. Since Intron I will be removed from the pre-mRNA before it is exported to the cytoplasm to be translated on the ribosome, this isn't a problem. Upon splicing, the entire β -globin protein will be translated in the **c reading frame** as established at the beginning of Exon I.

Intron I splits a codon. The 30th amino acid of β -globin is arginine (Arg or R), encoded by the codon AGG. This codon is split by Intron I. The first two nucleotides of this codon (AG) are found at the end of Exon I, while the last nucleotide of this codon (G) is the first





Features of the Teacher's Map continued

nucleotide of Exon II. This detail illustrates how precisely splicing reactions must be performed in order to maintain the correct reading frame throughout the entire mature mRNA.

Nucleotides 62,409 to 62,631: Exon II (223 nucleotides, 74 1/3 codons)

This is the longest of the three exons of the β -globin gene, encoding 74 1/3 amino acids. Both His63 and His92 are encoded in this exon. In the β -globin protein, these two amino acids are positioned on either side of the iron atom of the heme group. His92 is known as the *proximal histidine* and directly binds the iron atom of heme.

Nucleotides 62,632 to 63,481: Intron II (850 nucleotides)

Introns are often much longer than exons. At 850 nucleotides, the second intron of the β -globin gene is ~ nine times longer than Exon I (92 nucleotides). The interruption of eukaryotic genes by introns that often are much longer than exons is part of the reason the human genome is *1000-fold* larger than the *E. coli* genome, even though the human genome only codes for about five times as many proteins. (Prokaryotic genes don't split into exons and introns.)

Nucleotides 63,482 to 63,610: Exon III (129 nucleotides, 43 codons)

The last codon of Exon III is UAA, one of three STOP codons. There are no normal transfer RNAs that recognize STOP codons. When STOP codons are encountered, protein release factors bind to the ribosome and cause the protein to be released from the ribosome.

The mRNA continues beyond the STOP codon. The *3' non-translated region* of mRNAs include transcription stop signals and polyadenylation signals.

Mutations of the β -Globin Gene

A large number of mutations in the human β -globin gene have been documented from patients exhibiting a wide range of hemoglobinopathies. It is one of the most thoroughly studied genes.

Nucleotide 62,206: Sickle Cell Anemia

Sickle cell anemia often is referred to as *the first molecular disease*. All cases of sickle cell anemia result from a single base change from an *A* to a *T* in the codon of the 6th amino acid in the mature protein. GAG, which codes for glutamic acid (E) in the wild-type protein is changed to GTG, which codes for valine (V) in the mutant. Negatively charged glutamic acid, which normally contributes to the solubility of the protein in an aqueous environment, is replaced by hydrophobic valine. The hydrophobic patch on the protein's surface has no effect on the protein's oxygen affinity, but hemoglobin tetramers (2 α -chains and 2 β -chains) aggregate in red blood cells exposed to low oxygen pressures. This leads many of the normal red blood cells to assume a sickle shape rather than a normal *donut shape*. Sickle-shaped cells can't pass easily through small capillary beds, which causes clogging and leads to severe pain and cell lysis.

Splicing Mutations

Although interesting, the following four mutations may be too advanced to discuss with high school students.

Nucleotide 62,265: Creation of a Cryptic Donor Splice Site in Exon I

The dinucleotide GU found at nucleotides 62,263 and 62,264 does not normally function as a donor splice site because the flanking nucleotides do not conform to the donor splice site consensus sequence. However, when the G at 62,265 is mutated to an A, this becomes a weak donor splice site that is used approximately 40% of the time when this pre-mRNA is spliced. This aberrant splicing event removes 16 nucleotides from the end of Exon I. When this new

Features of the Teacher's Map continued

cryptic donor site is spliced to the normal acceptor site at nucleotide 62,409, the reading frame for the protein shifts to the ***b reading frame***, since 16 is not a multiple of 3. As a result, 29 amino acids are added to those coded by the truncated Exon I before protein synthesis is terminated by a STOP codon. The 29 amino acids are totally unrelated to β -globin and the truncated protein is completely non-functional. However, 60% of the time, the pre-mRNA from this mutated gene is spliced correctly, resulting in normal β -globin protein. Since there is not enough of β -globin protein to assemble with the chains of α -globin to make hemoglobin, this condition is referred to as β^+ -thalassemia (a deficiency of β -globin).

Nucleotide 62,279: Inactivating the Normal Donor Splice Site of Exon I

Since all donor splice sites start with GU, the mutation of the G at nucleotide 62,279 to A completely inactivates this normal splice site. No functional β -globin protein is made and causes a condition known as β^0 -thalassemia (a complete absence of β -globin).

Nucleotide 62,283: Crippling the Normal Donor Splice Site of Exon I

Mutating the G at 62,283 to an A decreases the efficiency with which the donor splice site is used. The protein is completely normal, but its decreased amount results in β^+ -thalassemia.

Nucleotide 62,388: Creation of a New Acceptor Splice Site in Intron I

The mutation of the G at 62,388 to an A creates a new acceptor splice site within Intron I. When used, an additional 19 nucleotides are added to mature mRNA. This results in a *shifted reading frame*, and a completely non-functional protein. Since some normal β -globin protein is made from the mRNA that is spliced correctly, this mutation results in β^+ -thalassemia.

More Mutations of the β -Globin Gene

Nucleotide 62,540: Hemoglobin Shepherds Bush

Mutation of the G at 62540 to A changes the glycine (G) codon GGC to an aspartic acid (D) codon GAC. In the tetrameric hemoglobin structure, this mutation maps to the interface between the two β -globin subunits where BPG (bisphosphoglycerate) is bound to reduce the oxygen affinity of normal hemoglobin. Since BPG is a negatively charged molecule, the introduction of a negatively charged aspartic acid side chain at this position reduces BPG binding. This leads to the increased oxygen affinity seen in Hemoglobin Shepherds Bush. Patients with Hemoglobin Shepherds Bush are able to load O_2 onto hemoglobin in the lungs, but are not able to efficiently release O_2 in peripheral tissues where the oxygen is needed. (The names given to hemoglobin mutants refer to the geographic location where they were first documented, in this case, the Shepherds Bush area of London.)

Nucleotide 63,561: Hemoglobin Shanghai

Since proline is a cyclic amino acid, it does not comfortably exist in an alpha helix. When A at nucleotide 63,561 is mutated to C, the glutamine (Q) codon CAG is changed to the proline (P) codon CCG. The introduction of this helix-breaking proline in the middle of a long alpha helix disrupts the helix, resulting in an unstable protein that degrades at a much higher rate than wild type β -globin.

Nucleotide 63,605: Hemoglobin Hiroshima

Hemoglobin is a magnificent protein that is designed to respond to its environment. For example, as the pH drops in vigorously exercising muscle, hemoglobin binds the excess H^+ , and this in turn results in the release of more O_2 to these tissues. The imidazole sidechain of the C-terminal histidine 146 is responsible for binding H^+ . Therefore the mutation of the CAC



Features on the Teacher's Map continued

histidine codon to the GAC aspartic acid (D) codon eliminates this critical amino acid and eliminates the ability of hemoglobin to release more oxygen in exercising muscle.

Note to Teachers

Important concepts for students to understand are:

- The relationship between DNA and protein, that is, nucleotide changes (mutations) in the gene result in changes in the amino acid sequence of the protein, and
- Structure and function are tightly linked in biology

Both of these concepts can be reinforced when the mutations documented in this exercise are **mapped** onto the 3D structure of the β -globin protein. We encourage teachers to use this **β -Globin Gene Exercise®** in conjunction with the physical model of the β -globin protein. (See 3D Molecular Designs' Mini-Toober β -Globin Folding Kit®, Prefolded Mini-Toober β -Globin Model or β -Globin Mini Model.) The Mini-Toober β -Globin Folding Kit® also introduces another important concept:

- The amino acid sequence of a protein determines its final 3D folded shape, following basic principles of chemistry and physics.

The Map of the β -Globin Gene® features images of the amino acid sidechains that correspond to those in the β -Globin Folding Kit®.

Additional Features for Advanced Courses

For advanced courses, in which promoter and polyadenylation sequences are discussed, we added new features to the 2.0 version of the Teachers Map of the β -Globin Gene Map®. A description of these sequences follows.

Promoter Sequences

Nucleotides 62027 to 62036 and 62042 to 62051: CACCC Sequences

With the exception of human β -globin, all β -like globin genes of human, mouse, rabbit and goat contain at least one copy of the sequence PuPuCPy(T/A/C)CACCC located 9 or 23-25 base pairs upstream from a CCAAT sequence. Human and rabbit genes contain two copies of this sequence; in rabbit, both copies are necessary, whereas in human β -globin, only the sequence nearest the CCAAT box binds the erythroid-specific nuclear factor (NF-E1). NF-E1 is expressed only in red blood cells, predominantly after birth, and binds much more tightly to β -globin than to γ -globin sequences. Thus, this sequence is important in switching expression to the β -globin gene after birth. A mutation at nucleotide 62,050 leads to β -thalassemia.

Nucleotides 62059 to 62067: CCAAT Promoter Sequence

Many eukaryotic genes contain a CCAAT promoter sequence, located 70-90 bp upstream from the transcription start site. This sequence binds the CAAT-binding transcription factor (CTF) which impacts the transcription efficiency. The CCAAT consensus sequence is GG(C/T)CAATCT. It is highly conserved among mammalian α - and β -globin genes.

Nucleotides 62105 to 62111: TATA Promoter Sequence

The TATA box in eukaryotes, located 20-30 bp upstream from the transcription start site, is involved in recognizing the transcription initiation site and is similar to the -10 sequence in prokaryotes. The TATA-binding protein (TBP) binds to this site and assists in positioning RNA polymerase and other factors of the pre-initiation complex. The consensus sequence for the TATA box is TATA(A/T)A(T/A). Genes whose TATA box is



Features on the Teacher's Map continued

close to the consensus sequence are transcribed more frequently than genes whose TATA sequence diverges from this consensus. The β -globin gene TATA sequence matches the consensus in all but the first base of the sequence, which is a C. The β -globin gene is one of a class of proteins whose TATA box also regulates the efficiency of transcription in addition to the site of transcriptional initiation. A mutation at nucleotide 62,110 from A to C results in β^+ -thalassemia.

Polyadenylation Sequences**Nucleotides 63637 to 63642 and 63718 to 63723: Polyadenylation Sequences**

The mammalian consensus polyadenylation sequence is AAUAAA, though there are numerous one-base variations of this sequence that result in polyadenylation, albeit at lower rates. Many human genes contain multiple polyadenylation signal sequences; typically the signal most 3' to the gene is the consensus AAUAAA, as is the case with β -globin. This sequence is found 10-30 bp upstream of the cleavage/polyadenylation site. The cleavage/polyadenylation specificity factor binds to this sequence and is necessary for both cleavage and polyadenylation.

Nucleotides 63737 to 63741: CATTG Sequence

CATTG is found 3' of the AATAAA before or after the poly(A) site in many vertebrate genes. This sequence is complementary to sequences within the RNA of U4 small ribonucleoprotein, suggesting that U4 may mediate mRNA 3' end formation. In the β -globin gene, the CATTG sequence is immediately upstream from the cleavage and polyadenylation site.

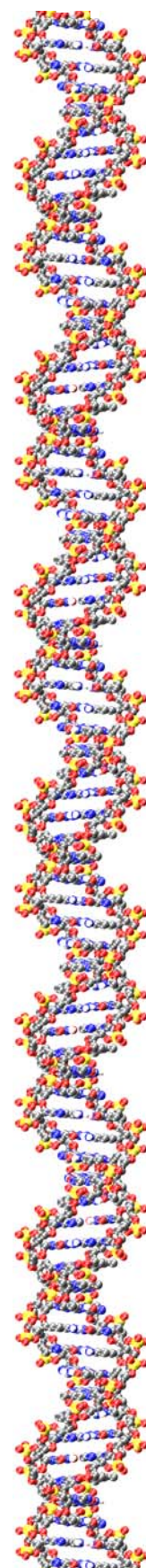
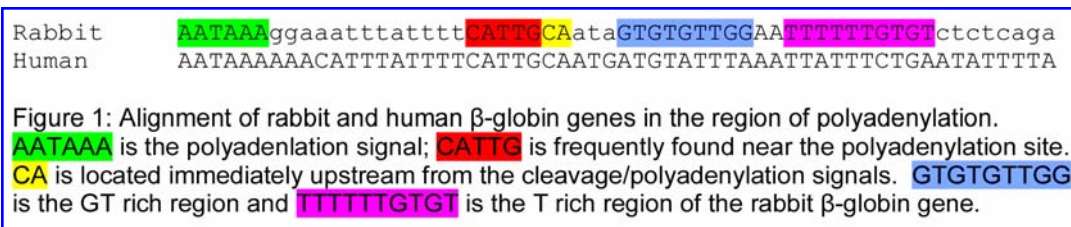
Nucleotides 63742 to 63743: Cleavage and polyadenylation site

Just 5' to the cleavage/polyadenylation site, the dinucleotide CA is the preferred sequence. Though this sequence is not required, in β -globin, the consensus sequence is seen.

Nucleotides 63747 to 63755: GU-rich region involved in polyadenylation**Nucleotides 63758 to 63767: U-rich region involved in polyadenylation**

Efficient mammalian gene polyadenylation requires either a GT or T rich region 23-24 bp downstream from the polyadenylation site. These regions are not highly conserved; the position of these regions, rather than the homology, is important. The β -globin gene contains both sequences, though neither is readily recognized.

Below is the polyadenylation region of rabbit and human β -globin genes. The rabbit sequence (on top) shows the GT and T rich regions; the six differences between the two genes in this region make it more difficult to identify the region in the human β -globin gene without aligning it with the rabbit sequence.



Features on the Teacher's Map continued

Bibliography

Beaudoing, Emmanuel, Freier, Susan, Wyatt, Jacqueline R., Claverie, Jean-Michel and Gautheret, Daniel. 2007. Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Res.* 10:1001-1010.

Colgan, Diana F. and Manley, James L. 1997. Mechanism and regulation of mRNA polyadenylation. (Review) *Genes and Dev.* 11:2755-2766.

deBoer, Ernie, Antoniou, Michael, Mignotte, Vincent, Wall, Lee and Grosveld, Frank. 1988. The human β -globin promoter; nuclear protein factors and erythroid specific induction of transcription. *The EMBO Journal* 7(13):4203-4212.

Dierks, Peter, van Ooyen, Albert, Cochran, Mark D., Dobkin, Carl, Reiser, Jakob and Weissmann, Charles. 1983. Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit β -globin gene in mouse 3T6 cells. *Cell* 32:695-706.

Donze, David, Townes, Tim M. and Bieker, James J. 1995. Role of Erythroid Kruppel-like Factor in Human γ - to β -Globin Gene Switching. *J. Biol. Chem.* 270(4):1955-1959.

Gil, Anna and Proudfoot, Nicholas J. 1987. Position-Dependent Sequence Elements Downstream of AAUAAA Are Required for Efficient Rabbit β -Globin mRNA 3' End Formation. *Cell* 49:399-406.

Grosveld, Frank, van Assendelft, Greet Blom, Greaves, David R. and Kollias, George. 1987. Position-independent, high-level expression of the human β -globin gene in transgenic mice. *Cell* 51:975-985.

Lanoix, Jacqueline and Acheson, Nicholas H. 1988. A rabbit β -globin polyadenylation signal directs efficient termination of transcription of polyomavirus DNA. *The EMBO J.* 7(8):2515-2522.

Mantovani, Roberto, Margaretti, Nicoletta, Nicolis, Sylvia, Giglioni, Barbara, Comi, Paola, Cappellini, Nica, Bertero, Maria Tiziana, Caligaris-Cappio, Federico and Ottolenghi, Sergio. 1988. An erythroid specific nuclear factor binding to the proximal CACCC box of the β -globin gene promoter. *Nucleic Acids Research* 16(10):4299-4313.

Myers, Richard M., Tilly, Kit and Maniatis, Tom. 1986. Fine Structure Genetic Analysis of a β -globin promoter. *Science* 232(4750):613-618.

Orkin, Stuart H., Kazazian, Haig H. Jr., Antonarakis, Stylianos E., Goff, Sabra C., Boehm, Corinne D., Sexton, Julianne P., Waber, Pamela G. and Giardina, Patricia J.V. 1982. Linkage of β -thalassaemia mutations and β -globin gene polymorphisms with DNA polymorphisms in human β -globin gene cluster. *Nature* 296(5858):627-631.

Weaver, Robert Franklin. 2002. *Molecular Biology*. Boston: McGraw Hill.

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

