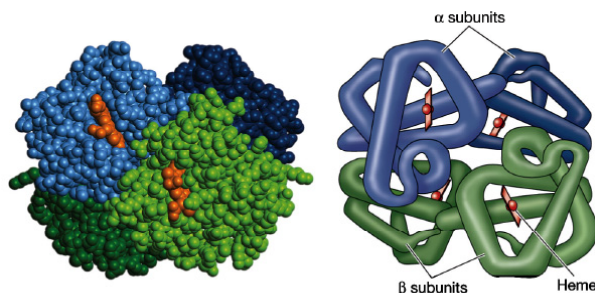
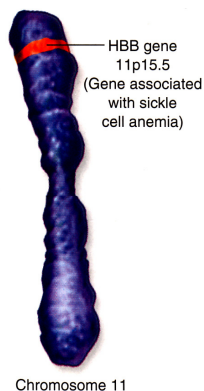


Human β -globin (Hemoglobin) Gene Sequence

Beta-globin (β -globin) is one of the protein subunits that make up the protein, hemoglobin. Hemoglobin is a quaternary protein comprised of two alpha subunits linked with two beta subunits. Each subunit binds a heme group, which in turn binds one iron atom (Fe). Oxygen can bind reversibly to this Fe atom. As a red blood cell passes through the lungs, oxygen diffuses into the cell and binds to the Fe of the heme group on all four subunits of the hemoglobin tetramer.



A common form of sickle-cell anemia is caused by a single point mutation in the nucleotide sequence of β -globin. The mutation is located in the seventh codon (The first codon codes for Met, the leader amino acid in polypeptides.) The seventh triplet should read GAG which codes for glutamic acid, but the middle nucleotide has changed to a thymine, which changes the triplet to GTG, which codes for valine. β -globin's three-dimensional structure places this seventh amino acid on the outside of the protein. Replacement of the normally charged glutamic acid with the hydrophobic valine alters the solubility of hemoglobin, so that at a lower oxygen concentration, the altered protein changes the red blood cell to a sickle shape that is unable to carry oxygen. This causes the symptoms of sickle-cell anemia.



HBB Sequence in Normal Adult Hemoglobin (Hb A):

Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Glu	Glu	Lys	Ser

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Val	Glu	Lys	Ser

In this activity you will be observing a portion of the actual nucleotide sequence for the β -globin polypeptide. This sequence (found at the end of this exercise) was obtained from GenBank, an internet source, under the key word: HUMHBB. There are actually 73,308 nucleotides for hemoglobin, which is located at the tip of chromosome 11. This includes the adult β -globin gene as well as several other related globin genes (fetal, α -globin, etc.). The entire hemoglobin nucleotide sequence takes 40 pages to list. Taped together in the correct order, it would represent just 0.002% of the entire human genome. The sequence that contains the β -globin information is located between nucleotides 62,137 \rightarrow 63,660.

You will be tracing the information flow from DNA \rightarrow mRNA \rightarrow amino acid sequence of the β -globin polypeptide in this exercise.

Exercise I – The β -globin Gene Sequence

Procedure

1. Read "Interpreting the β -globin Gene Sequence Map" below.
2. Once you understand how to interpret the β -globin gene map, turn to the β -globin nucleotide sequence reproduced at the end of this exercise.
3. Can you find the gene? There are 1,473 nucleotides in the gene, enough to code for 491 triplet codons. But the β -globin polypeptide has just 146 amino acids. The excess nucleotides are introns (intervening sequences) that separate the protein coding "exons" (expressed sequences). The β -globin gene consists of three exons and two introns.

Exon #1	62187 – 62279
Intron #1	62280 – 62409
Exon #2	62410 – 62631
Intron #2	62632 – 63481
Exon #3	63482 – 63610

4. The sequence of amino acids that comprise the β -globin's three exons is below. This representation uses an alphabet letter code for each amino acid. See Table 1 at the end of Exercise II for the amino acid and codon equivalents for the letters.

MVHLTPEEKSAVTALWGKVVNVDEVGGEALGR

LLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLAHLNLRKGTFFATLSEHLCDKLHVDPENFR

LLGNVLVCVLAHHFGKEFTTPVQAAAYQKVVAGVANALAHKYH

5. Locate the promoter region of the gene and highlight the TATA box (TCAAA) and CAT box (CATT) sequences in the promoter.
6. Locate the introns and highlight them with a transparent marker.
7. Translate your entire sequence into messenger RNA (mRNA).
 - a. Which side is the coding side that is transcribed into RNA? _____
(If you aren't sure, reread "Interpreting the β -globin Gene Sequence Map".)
8. Highlight the sections that will be removed by the spliceosome during mRNA processing.
 - a. How many nucleotides will be removed per intron section?
9. Recopy the new shorter mRNA message.
 - a. Place a slash mark between each codon.
 - b. Does the amino acid sequence in #4 above match your mRNA? _____
 - c. Record your mRNA message.
 - d. What is the seventh codon of your mRNA? _____
 - e. Confirm that replacing the middle nucleotide of the seventh codon with **U**, the amino acid sequence is changed to valine.

Note: Life Science Tools: <http://www.fr33.net> is a helpful website that provides DNA and RNA tools to convert or translate DNA and RNA sequences.

Exercise II – Restriction Site Analysis of the β -globin Sequence:

Search for the following restriction enzyme recognition sequences in your β -globin sequence.

- Sac I (GAGCTC)
- Bgl II (AGATCT)
- Hpa II (CCGG)
- Xma I (CCCGGG)
- BamHI (GGATCC)
- Ava II (GGACC)
- Hind III (AAGCTT)
- EcoRI (GAATTC).

Which restriction enzyme site(s) would work best for cloning the β -globin sequence of this gene? Explain your answer.

Table 1: Amino Acid Symbols
IUB Symbol 3-letter Meaning Codons Depiction

Letter	Amino Acid Abr	Amino Acid	DNA Triplet (codons)
A	Ala	Alanine	GCT, GCC, GCA, GCG
B	Asp, Asn	Aspartic, Asparagine	GAT, GAC, AAT, AAC
C	Cys	Cysteine	TGT, TGC
D	Asp	Aspartic	GAT, GAC
E	Glu	Glutamic	GAA, GAG
F	Phe	Phenylalanine	TTT, TTC
G	Gly	Glycine	GGT, GGC, GGA, GGG
H	His	Histidine	CAT, CAC
I	ile	Isoleucine	ATT, ATC, ATA
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	TTG, TTA, CTT, CTC, CTA, CTG
M	Met	Methionine	ATG
N	Asd	Asparagine	AAT, AAC
P	Pro	Proline	CCT, CCC, CCA, CCG
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	CGT, CGC, CGA, CGG, AGA, AGG
S	Ser	Serine	TCT, TCC, TCA, ACG, AGT, AGC
T	Thr	Threonine	ACT, ACC, ACA, ACG
V	Val	Valine	GTT, GTC, GTA, GTG
W	Trp	Tryptophan	TGG
X	Xxx	unknown	XXX
Y	Tyr	Tyrosine	TAT, TAC
Z	Glu, Gln	Glutamic, Glutamine	GAA, GAG, CAA, CAG
*		End Terminator	TAA, TAG, TGA

Interpreting the β -globin Gene Sequence Map

The beginning of the globin map file is shown below. Using this file, the following information can be explored:

Nucleotide Number

```

ACATTTGCTTCTGACACAACACTGTGTTCACT
_____ = I
_____ = II
_____ = III
62137 ----- Double-Stranded DNA
TGTA AACGAAGACTGTGTTGACACAAGTGA
  
```

Three Translational Reading Frames:

I	T	F	A	S	D	T	T	V	F	T
II	H	L	L	L	T	Q	L	C	S	L
III	I	C	F	*	H	N	C	V	H	*

Note: The letters in the three reading frames above represent the one-letter abbreviation of amino acids. The * represents the STOP codon. See Table 1 for the amino acid and codon equivalents for the letters.

1. The nucleotide sequence of both strands of DNA is shown, beginning with nucleotide 62,137. Remember that in the Watson-Crick double helix, the two strands of DNA are anti-parallel. The top strand represents the "non-coding" strand and the bottom strand represents the "coding" strand. When this gene is expressed, RNA polymerase uses the bottom strand as a template to catalyze synthesis of an mRNA copy that corresponds to the sequence shown in the top strand - replacing thymine (T) on the DNA with uracil (U) in the mRNA molecule.
2. The amino acid sequence encoded by the DNA is shown in the three lines below the nucleotide sequence. The one-letter abbreviation of the amino acids is used. (See Table 1). For example, the letter "T" represents threonine, "H" represents histidine, "I" represents isoleucine, etc. The * refers to one of the three translation STOP codons.
3. The triplet generic code could be read in three possible "reading frames" on the DNA. The three lines underneath the non-coding DNA strand represent the three "reading options". The translated versions of these three reading frames are shown in representational code below the DNA and labeled "I, II, and III". In the "I" reading frame, the ribosome reads the triplet ACA to mean T (threonine). In reading frame "II", the ribosome shifts one nucleotide and reads the triplet CAT to mean H (histidine). With "III" the frame shifts again and reads ATT to mean I (isoleucine).
4. You need to determine which of the three reading frames above is used to translate the β -globin protein, and then locate where on the DNA the translation begins. Using the sequence documentation found in the HUMHBB file at the end of this exercise, you know that the coding sequence starts at nucleotide 62,187. That establishes the "III" reading frame as the one that is used to translate the β -globin mRNA. Therefore, the amino acid sequence of the β -globin protein begins as MVHLTPEEK, and continues for a total of 146 amino acids in the polypeptide. This coding region constitutes an "open reading frame". Examine the other 2 possible reading frames over this same region of the gene and note all of the translation STOP codons.
5. Finally, examine the triplet that encodes the seventh amino acid – an E (glutamic acid). In sickle-cell anemia, this GAG is changed to GTG as a result of a mutation in the second base of the triplet. The new DNA triplet, GTG encodes the hydrophobic amino acid valine. The replacement of a negatively-charged glutamic acid with a hydrophobic valine at this position of the β -globin protein is responsible for the aggregation of hemoglobin in red blood cells that results in sickle-cell anemia.

Human β -globin sequence from 62137 → 63660

62137 ACAT TTGCTTCTGA CACAACGTGT
62161 TTCACTAGCA ACCTCAAACA GACACCATGG TGCACCTGAC TCCTGAGGAG AAGTCTGCCG
62221 TTACTGCCCT GTGGGGCAAG GTGAACGTGG ATGAAGTTGG TGGTGAGGCC CTGGGCAGGT
62281 TGGTATCAAG GTTACAAGAC AGGTTTAAGG AGACCAATAG AAAC TGGGCA TGTGGAGACA
62341 GAGAAGACTC TTGGGTTTCT GATAGGCACT GACTCTCTCT GCCTATTGGT CTATTTTCCC
62401 ACCCTTAGGC TGCTGGTGGT CTACCCTTGG ACCCAGAGGT TCTTTGAGTC CTTTGGGGAT
62461 CTGTCCACTC CTGATGCTGT TATGGGCAAC CCTAAGGTGA AGGCTCATGG CAAGAAAGTG
62521 CTCGGTGCCT TTAGTGATGG CCTGGCTCAC CTGGACAACC TCAAGGGCAC CTTTGCCACA
62581 CTGAGTGAGC TGCAC TGTGA CAAGCTGCAC GTGGATCCTG AGAACTTCAG GGTGAGTCTA
62641 TGGGACCCTT GATGTTTTCT TCCCCCTTCT TTTCTATGGT TAAGTTCATG TCATAGGAAG
62701 GGGAGAAGTA ACAGGGTACA GTTTAGAATG GGAAACAGAC GAATGATTGC ATCAGTGTGG
62761 AAGTCTCAGG ATCGTTTTAG TTTCTTTTAT TTGCTGTTCA TAACAATTGT TTTCTTTTGT
62821 TTAATTCCTTG CTTTCTTTTT TTTCTTCTC CGCAATTTTT ACTATTATAC TTAATGCCTT
62881 AACATTGTGT ATAACAAAAG GAAATATCTC TGAGATACAT TAAGTAACTT AAAAAAAAAAC
62941 TTTACACAGT CTGCCTAGTA CATTACTATT TGGAATATAT GTGTGCTTAT TTGCATATTC
63001 ATAATCTCCC TACTTTATTT TCTTTTATTT TTAATTGATA CATAATCATT ATACATATTT
63061 ATGGGTTAAA GTGTAATGTT TTAATATGTG TACACATATT GACCAAATCA GGGTAATTTT
63121 GCATTTGTAA TTTTAAAAAA TGCTTTCTTC TTTTAATATA CTTTTTTGTT TATCTTATTT
63181 CTAATACTTT CCCTAATCTC TTTCTTTCAG GGCAATAATG ATACAATGTA TCATGCCTCT
63241 TTGCACCATT CTAAAGAATA ACAGTGATAA TTTCTGGGTT AAGGCAATAG CAATATTTCT
63301 GCATATAAAT ATTTCTGCAT ATAAATTGTA ACTGATGTAA GAGGTTTCAT ATTGCTAATA
63361 GCAGCTACAA TCCAGCTACC ATTCTGCTTT TATTTTATGG TTGGGATAAG GCTGGATTAT
63421 TCTGAGTCCA AGCTAGGCC TTTTGCTAAT CATGTTTATA CCTCTTATCT TCCTCCCACA
63481 GCTCCTGGGC AACGTGCTGG TCTGTGTGCT GGCCCATCAC TTTGGCAAAG AATTCACCCC
63541 ACCAGTGCAG GCTGCCTATC AGAAAGTGGT GGCTGGTGTG GCTAATGCC TGGCCACAA
63601 GTATCACTAA GCTCGCTTTC TTGCTGTCCA ATTTCTATTA AAGGTTTCTT TGTTCCCTAA