Human Genetics--BIOL 102
Summer Lab 2--The Process Whereby Your Genes Make Your Proteins

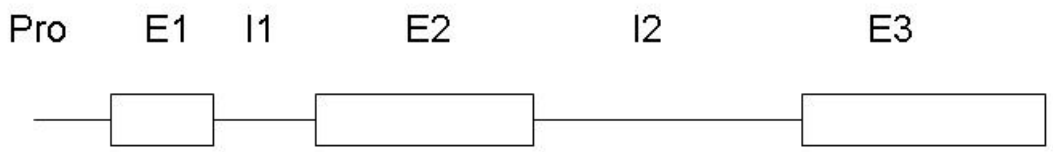
I) Purpose And Procedure

In this exercise, you will dissect the sequence of the human beta-globin gene, to reveal which portions of the gene are used to encode the beta-globin protein and where certain critical regulatory sites are. You will also answer questions regarding whether a particular change in the DNA sequence is expected to change the level of activity in the gene's protein.

The sequence below is the sequence of the human beta-globin gene from chromosome 11. This particular evrsion of the beta-globin gene's sequence is from the National Center for Biotechnology Information's GenBank database. GenBank is the repository for all published DNA sequences, from all the species whose DNA has been or is being sequenced.

II) The Structure And Sequence Of The Human Beta-Globin Gene

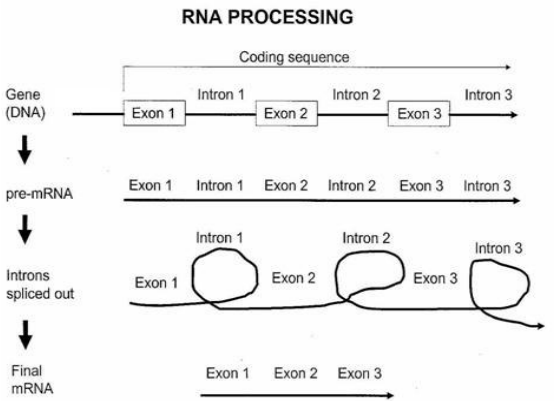
As you can see, this is a sequence of 2052 nucleotides. It includes the promoter region of the beta-globin gene, the coding sequence and some of the sequence that borders the gene on both sides. The arrangement of the promoter region, exons and introns in the human beta-globin gene is depicted in the figure below.



The different elements of the gene's structure correspond to the nucleotides in the sequence as follows:
Pro = promoter region, nucleotides 1-103
E1 = exon 1, nucleotides 104-245
I1 = intron 1, nucleotides 246-375
E2 = exon 2, nucleotides 376-598
I2 = intron 2, nucleotides 599-1448
E3 = exon 3, nucleotides 1449-1709

To help you decipher the information presented, read the notes on the left side of the GenBank entry, which indicate the essential aspects of the sequence. Here's a glossary, so you know what the terms mean:
source = the entire sequence in this entry

primary transcript (aka pre-mRNA) = you will recall that, during transcription, the gene makes a primary transcript, or pre-mRNA, which then gets processed into mRNA (see figure below). The primary transcript begins with the first nucleotide of exon 1, and ends with the last nucleotide in the last exon (in this case, exon 3).



exon and intron = you will recall that portions of the pre-mRNA are spliced out as the pre-mRNA is processed to make the mRNA (see figure above). The portions that remain in the mRNA are called exons, while the portions that get spliced out are called introns. The coding sequence of the beta-globin gene has three exons, and two introns.

CDS = The CDS contains only the nucleotides that are actually read by the ribosome and used to determine the amino acid sequence of the polypeptide being made. This is not the entire mRNA. There is always some of the mRNA, at both ends, that is not actually read by the ribosome and used to string the amino acids together, so the CDS is smaller than the mRNA. The first approximately 50 nucleotides of the mRNA get the ribosome oriented, and the CDS begins at approximately the 51st nucleotide of the mRNA. In addition, the CDS does not go as far as the mRNA, so the CDS will end somewhere in the middle of the last exon (here, exon 3).

Translation = the sequence of amino acids in the protein. The single-letter abbreviations are used for the different amino acids.

GenBank Sequence Repository--Human beta-Globin Gene

```
FEATURES             Location/Qualifiers
     source          1..2052

  primary transcript 104..1709

     exon 1          104..245

     intron 1        246..375

     exon 2          376..598

     intron 2        599..1448

     exon 3          1449..1709

     CDS             join(154..245,376..598,1449..1577)

     Translation     MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFE
                     SFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPE
                     NFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

SEQUENCE

```
       1 ccctgtggag ccacacccta gggttggcca atctactccc aggagcaggg agggcaggag
      61 ccagggctgg gcataaaagt cagggcagag ccatctattg cttacatttg cttctgacac
     121 aactgtgttc actagcaacc tcaaacagac accatggtgc acctgactcc tgaggagaag
     181 tctgccgtta ctgccctgtg gggcaaggtg aacgtggatg aagttggtgg tgaggccctg
     241 ggcaggttgg tatcaaggtt acaagacagg tttaaggaga ccaatagaaa ctgggcatgt
     301 ggagacagag aagactcttg ggtttctgat aggcactgac tctctctgcc tattggtcta
     361 ttttcccacc cttaggctgc tggtggtcta cccttggacc cagaggttct ttgagtcctt
     421 tggggatctg tccactcctg atgctgttat gggcaaccct aaggtgaagg ctcatggcaa
     481 gaaagtgctc ggtgccttta gtgatggcct ggctcacctg gacaacctca agggcacctt
     541 tgccacactg agtgagctgc actgtgacaa gctgcacgtg gatcctgaga acttcagggt
     601 gagtctatgg gacccttgat gttttctttc cccttctttt ctatggttaa gttcatgtca
     661 taggaagggg agaagtaaca gggtacagtt tagaatggga aacagacgaa tgattgcatc
     721 agtgtggaag tctcaggatc gttttagttt cttttatttg ctgttcataa caattgtttt
     781 cttttgttta attcttgctt tctttttttt tcttctccgc aatttttact attatactta
     841 atgccttaac attgtgtata acaaaaggaa atatctctga gatacattaa gtaacttaaa
     901 aaaaaacttt acacagtctg cctagtacat tactatttgg aatatatgtg tgcttatttg
     961 catattcata atctccctac tttatttttct tttatttttа attgatacat aatcattata
    1021 catatttatg ggttaaagtg taatgtttta atatgtgtac acatattgac caaatcaggg
    1081 taattttgca tttgtaattt taaaaaatgc tttcttcttt taatatactt ttttgtttat
    1141 cttatttcta atactttccc taatctcttt ctttcagggc aataatgata caatgtatca
    1201 tgcctctttg caccattcta aagaataaca gtgataattt ctgggttaag gcaatagcaa
    1261 tatttctgca tataaatatt tctgcatata aattgtaact gatgtaagag gtttcatatt
    1321 gctaatagca gctacaatcc agctaccatt ctgcttttat tttatggttg ggataaggct
    1381 ggattattct gagtccaagc taggcccttt tgctaatcat gttcatacct cttatcttcc
    1441 tcccacagct cctgggcaac gtgctggtct gtgtgctggc ccatcacttt ggcaaagaat
    1501 tcaccccacc agtgcaggct gcctatcaga aagtggtggc tggtgtggct aatgccctgg
    1561 cccacaagta tcactaagct cgctttcttg ctgtccaatt tctattaaag gttcctttgt
    1621 tccctaagtc caactactaa actgggggat attatgaagg gccttgagca tctggattct
    1681 gcctaataaa aaacatttat tttcattgca atgatgtatt taaattattt ctgaatattt
    1741 tactaaaaag ggaatgtggg aggtcagtgc atttaaaaca taaagaaatg atgagctgtt
    1801 caaaccttgg gaaaatacac tatatcttaa actccatgaa agaaggtgag gctgcaacca
    1861 gctaatgcac attggcaaca gcccctgatg cctatgcctt attcatccct cagaaaagga
    1921 ttcttgtaga ggcttgattt gcaggttaaa gttttgctat gctgtatttt acattactta
    1981 ttgttttagc tgtcctcatg aatgtctttt cactacccat ttgcttatcc tgcatctctc
    2041 tcagccttga ct
```

Note--there are 2052 total nucleotides in this sequence. They are arranged in blocks of 10, with six blocks of 10 on each line. The number on the left of each line = the number of the first nucleotide in that line.

III) Please answer the following questions. IMPORTANT: When referring to specific nucleotides, give the letters of the bases (A,C,G or T), as well as their numerical positions in the original DNA sequence. For questions 1-6, please just give me the bases and their numbers, or the range of bases--don't include the question or add any editorial material. (Ex. "GAC at 234, 235, 236" or "pre-mRNA = 456-879" if I am asking for a range of nucleotides).

1. a. How many nucleotides in the entire sequence entry?
b. How many exons are in the gene?
c. How many introns are in the gene?
To help you answer later questions, please highlight the three exons.

2. Transcription is initiated by the binding of transcription factors to the promoter region at the front end of the gene. The promoter region usually includes a variable number of nucleotides that lie in front of exon 1, and often includes approximately the first third of exon 1. Transcription factor binding sites sometimes have characteristic sequences. Among these are what people cutely call the "cat box motif," which involves the nucleotides C, A and T in some arrangement that spells out CAT or something like it, examples: CAAT, CCAATT, CATAAAA.
Please identify two sequences that might serve as "cat box" promoter sequences. First remind yourself where to look--think about where the promoter region of the gene is, compared to the gene's coding sequence.

3. Which nucleotides are used to make the pre-mRNA (aka the primary transcript)?

4. Recall what happens as the pre-mRNA is processed into mRNA. Which nucleotides are present in the mRNA?

5. Codon 31 begins with nucleotide 244. Which three nucleotides from the DNA sequence are contained in codon 31 of the mRNA?

6. Recall that the bases at the very beginning and very end of the mRNA are not actually read as instructions to chain the amino acids together; they help stabilize the mRNA and get the mRNA to where it needs to go to be translated. The CDS contains the nucleotides that were actually used to direct the ribosome to chain the appropriate amino acids together during translation. Look at the CDS to determine exactly where translation begins and ends. Which three nucleotides make up the translation initiation codon, aka the START codon? (Hint: its sequence is always ATG)

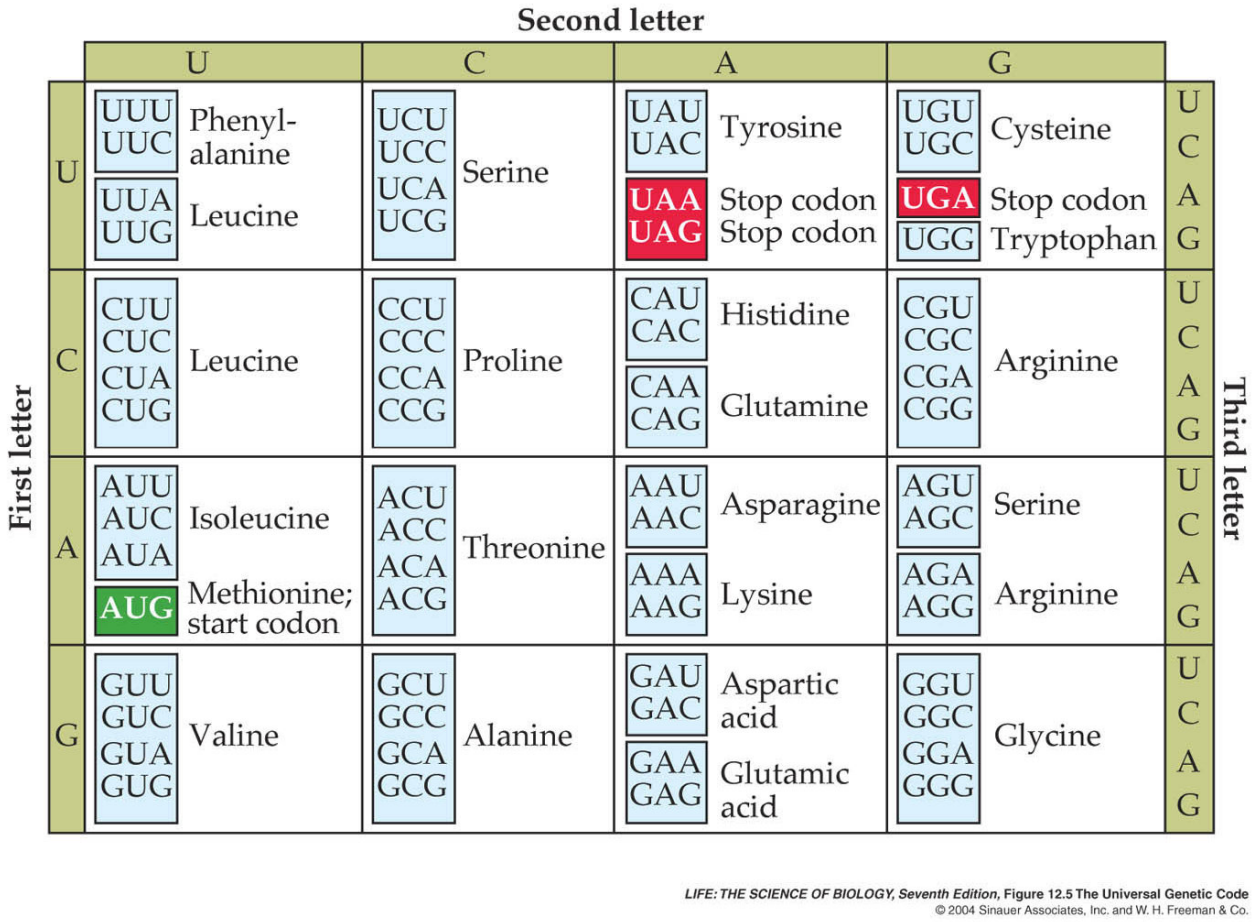
7. Again using the CDS to tell you where translation began and stopped, please tell me which of the three STOP codons the beta-globin gene uses to signal to the ribosome that it can stop adding amino acids to the polypeptide? Give me the three nucleotides and their positions in the original DNA sequence.

8. If there was a deletion that took out nucleotides 1-103, what effect would that deletion have on the beta-globin gene's activity, and why?

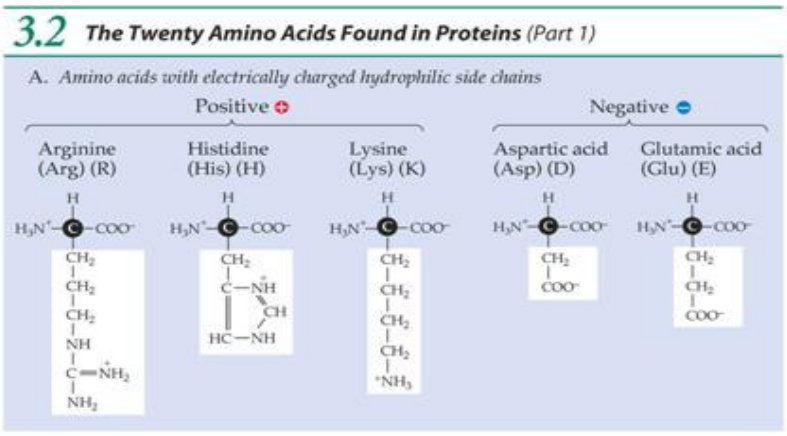
9. Codon 9 is an AAG codon; it causes lysine to be incorporated as the ninth amino acid in the polypeptide. Which would be more deleterious to the function of the protein, an A→G change

involving the first A in the codon (new codon = GAG) or an A→G change involving the second A in the codon (new codon = AGG)? Why do you say this?

Genetic Code Table

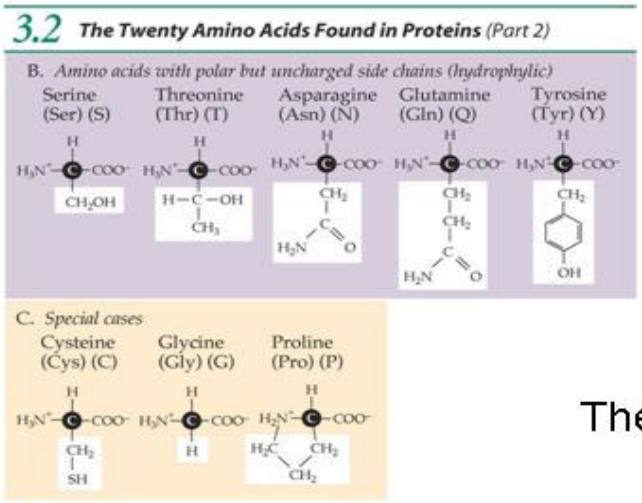


LIFE: THE SCIENCE OF BIOLOGY, Seventh Edition, Figure 12.5 The Universal Genetic Code



Arginine, Histidine and Lysine are positively charged

Glutamic acid and aspartic acid are negatively charged



The others are neutral