**Hemoglobin background**

**Learning goals:**

Students should be able to

* associate the molecular nature of DNA mutations with protein sequence through the use of online bioinformatics resources
* create alignments between nucleic acid or amino acid sequences using online bioinformatics resources
* identify the chromosomal location of hemoglobin monomers in the human chromosomes

**Oxygen binding proteins**

There are different oxygen (O2) binding proteins present in eukaryotic organisms. In animals, these proteins include hemoglobin, myoglobin, and cytochromes. Hemoglobin is involved in transporting oxygen from the lungs to tissues. Myoglobin transports and stores oxygen within muscle tissues. Cytochromes participate in redox reactions, including the electron transport chain. In plants, cytochromes participate in electron transport for both respiration and photosynthesis. These O2-binding proteins all interact with the prosthetic group heme and co-factor Fe2+ to coordinate the binding of one or more O2 molecules.

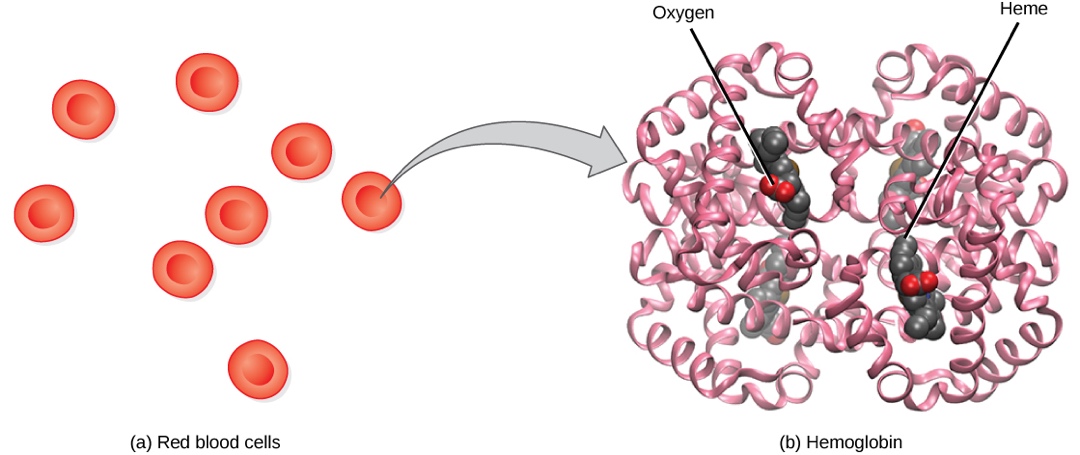


Figure showing hemoglobin. Figure taken from OpenStax Biology 2e with permission (Figure 39.19. The protein inside (a) red blood cells that carries oxygen to cells and carbon dioxide to the lungs is (b) hemoglobin.)

Hemoglobin is the primary oxygen-carrying protein found in red blood cells. In normal adult red blood cells, the protein is composed of four polypeptides (two α-globin and two ß-globin), each coordinately bound to a heme prosthetic group which in turn has the ability to bind to oxygen (O2) through ferrous iron (Fe2+). The molecular mass of the tetrameric hemoglobin protein is approximately 64.5 kDa (64,500 Daltons). The mature α-globin polypeptide is composed of 141 amino acids and the mature ß-globin polypeptide is composed of 146 amino acids. The secondary structures of the α- and ß-globin polypeptides suggest the presence of 7 or 8 alpha helices within the secondary structure, having bends and loops to fold the polypeptide (tertiary structure).

Hemoglobin is commonly found in two states - oxyhemoglobin (O2-bound hemoglobin, also called the R-state) and deoxyhemoglobin (no O2 bound, also called the T-state). These are characterized not only by their binding state to O2, but the resulting protein conformational change. It is possible for carbon monoxide (CO) and nitric oxide (NO) and other small molecules to bind to hemoglobin. Oxygen-binding is driven by a relatively higher concentration of O2 in the lungs than CO or NO. O2-binding is a positively cooperative event, meaning that when one O2 molecule binds, this O2 binding increases the affinity of the other three sites, resulting in a hemoglobin molecule with four bound O2 molecules. Similarly, when one O2 dissociates, the other three also do. Nitric oxide stimulates the blood vessel walls to relax. Recent evidence suggests that NO binding to hemoglobin through interactions with specific cysteine residues may deliver the NO to the blood vessel walls. CO binds with a greater affinity than O2 and binding is essentially irreversible. Carbon monoxide poisoning is due to the high affinity of hemoglobin for CO and the resulting inability of O2 to bind to hemoglobin and be delivered to the tissues.

The regulation of O2-binding to hemoglobin by H+ and CO2 is called the Bohr effect. Both H+ and CO2 are negative effectors of O2-binding. Addition of a proton to the imidazole group of histidine (His coordinately binding to heme prosthetic group) at the C-terminus of ß-globin polypeptide facilitates the formation of a salt bridge between a His and an Asp residue and stabilization of the T-state quaternary structure of deoxyhemoglobin.

The goal of this series of worksheets on hemoglobin will be to connect the DNA sequence with the structure and function of hemoglobin, to demonstrate the effects of mutations on the structure and function of hemoglobin and to demonstrate the connection between the mutations and disease.

The worksheets will also introduce you to techniques and algorithms (mostly how to use them) commonly used in bioinformatics.

**Instructions: Complete the worksheet (it is suggested to type or paste in your answers in the provided spaces). An assignment will be created on Canvas for you to upload this completed document. This is an INDIVIDUAL assignment.**

**Learning objectives:**

At the end of this worksheet, students should be able to:

* Find sequences and data by searching NCBI
* Identify basic information that is found within a GenBank file
* Perform a basic multiple sequence alignment (MSA) between two sequences

**Finding and comparing the nucleotide and amino acid sequences of the adult globin subunit mRNAs and polypeptides**

We will be performing several worksheets investigating the gene and polypeptide structure of several globin genes as well as the structure of the tetrameric hemoglobin protein (normally made up of two α-globin and two ß-globin polypeptides). To do this, we will be using different bioinformatics websites. Bioinformatics is the use of computer algorithms to mine data from collections. We will be investigating the globin gene family using some relatively simple approaches to demonstrate different aspects of the gene family, protein structure and disease relatedness.

In this first worksheet, you will be finding the DNA version of the mRNA sequence for both the α-globin and ß-globin genes from humans (*Homo sapiens*). You will be looking at the features of the mRNA sequences as well as comparing the polypeptide sequences. These are relatively basic approaches that you can use to begin to investigate a sequence and gather some information.

You should also review the basic structure of a eukaryotic gene.

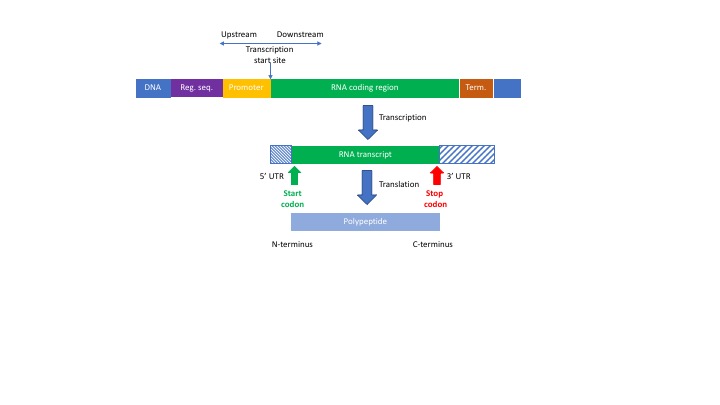


Figure showing an idealized gene structure including the regulatory and promoter sequences (both untranscribed), RNA coding region (transcribed) and terminator of the DNA. The RNA transcript including the coding sequence (encodes the polypeptide that is translated) and the 5’ and 3’ untranslated regions. The start codon (first translated codon) and the stop codon are also indicated.

1. How do the sequences of α- and ß-globin compare at the nucleotide and amino acid level?
   1. Finding the sequences for α- and ß-globin?
      1. α-globin (HbA1) mRNA nucleotide sequence (Genbank number NM\_000558)
         1. Go to [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) and enter the Genbank number.
         2. Click on the link under the ‘Nucleotide Sequence’. *This link shows the DNA version of the transcribed mRNA.*

|  |  |
| --- | --- |
| How long is the mRNA? | 577 bp |
| The coding sequence is the region that will be translated by the cell (in this case, the red blood cell precursors). Find the coding sequence (labeled as *cds*). What range of bases are included in the coding sequence (first to last base)? | 38 to 466 bp |
| What is the region of mRNA before the coding sequence called? (Refer to figure on previous page.) | 5’ UTR |
| What is the region of the mRNA after the coding sequence called? (Refer to figure on previous page.) | 3’ UTR |
| What is the DNA sequence of the stop codon used by α-globin? (Hint: The CDS (identified above) extends to the END of the stop codon.) | TAA |

* + - 1. Go back to the top of the file and click on the link for Fasta. This shows a different file format for the sequence. Copy and paste the sequence in the space below, replacing the “Student add” line with just the sequence. (I*f you wish to make it look neater, you can change the font of the sequence to Courier and the size to 8 or 9 pt.*)

Courier is an equal spaced font, in other words, a lower case l and an upper case M take up the same linear space, this allows sequences to look more uniform. (*M and l are an example of characters but are not nucleotide bases.*)

There is a second α-globin gene (called HbA2) – we will look just at HbA1.

In the Fasta shown, the >HbA1\_mRNA is used as an identifier. Multiple sequences can be entered into a Fasta file by using >seq\_name (no spaces) followed by a return and the sequence. Additional identifiers and sequences can be entered in the same manner. You will use this multiple sequence format later.

|  |
| --- |
| >HbA1\_mRNA  Student add  ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCA |

* 1. ß-globin (HbB) mRNA nucleotide sequence
     1. ß-globin mRNA nucleotide sequence (Genbank number NM\_000518).
        1. Go to [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) and enter the Genbank number.
        2. Click on the link to the mRNA sequence.

|  |  |
| --- | --- |
| How long is the mRNA? | 628 bp |
| The coding sequence is the region that will be translated by the cell (in this case, the red blood cell precursors). Find the coding sequence (labeled as cds). What bases are included in the coding sequence? | 51 to 494 bp |

* + - 1. Go back to the top of the file and click on the link for FASTA. This is a different file format for the sequence. Copy and paste the sequence in the space below, replacing the “Student add” line with just the sequence.

|  |
| --- |
| >HbB\_mRNA  Student add  ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGAT  CCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA  CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA  CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT  GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAA |

* 1. What is the polypeptide sequence of each globin subunit?
     1. α-globin amino acid sequence (GenBank number NP\_000549).
        1. Go to [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) and enter the Genbank number for the polypeptide.

|  |  |
| --- | --- |
| How many amino acids are in the coding sequence? | 142 aa |

* + - 1. Go back to the top of the file and click on the link for FASTA. This is a different file format for the sequence. Copy and paste the sequence in the space below, replacing the “Student add” line with just the sequence.

|  |
| --- |
| >HbA1\_polypeptide  Student add  MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNA  VAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSK  YR |

* + 1. ß-globin amino acid sequence
       1. ß-globin amino acid sequence (Genbank number NP\_000509).

|  |  |
| --- | --- |
| How many amino acids are in the coding sequence? | 147 aa |

* + - 1. Go back to the top of the file and click on the link for FASTA. This is a different file format for the sequence. Copy and paste the sequence in the space below, replacing the “Student add” line with just the sequence.

|  |
| --- |
| >HbB\_polypeptide  Student add  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG  AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN  ALAHKYH |

* 1. How similar are the α and ß-globin polypeptides?.
     1. Copy the FASTA amino acid sequences for α-globin and ß-globin and paste below in a FASTA format (replace “Copy and paste sequences from above”). *Include in your ‘paste’ the sequence identifiers (e.g., >HbA1\_polypeptide) on one line and the sequence on a separate line. Start a new line for the HbB polypeptide sequence.*

|  |
| --- |
| Copy and paste sequences from above  >HbA1\_polypeptide  MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR  >HbB\_polypeptide  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH |

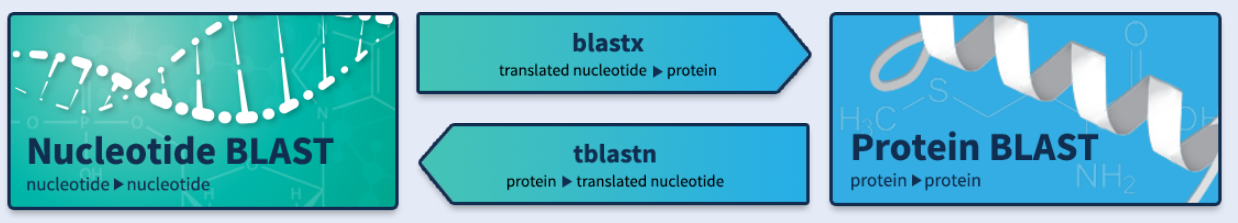
* + 1. To compare the two sequences, you will align the sequences using the program CLUSTAL Omega (*a previous version was called CLUSTALW if you have read about that program*). This program compares two or more sequences and creates an optimal alignment.
       1. Go to <http://www.ebi.ac.uk/Tools/msa/clustalo/> and paste the two FASTA sequences from above into the window.
       2. Use the default settings (change nothing). Scroll down and click on the SUBMIT button. Alignment may take a couple of minutes.
       3. Copy the resulting alignment file and paste in the box below. Copy and convert the alignment font to Courier and 8 or 9 point font to fit and look nice.

|  |
| --- |
| HbA1\_polypeptide -MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFD------LSHGS 53  HbB\_polypeptide MVHLTPEEKSAVTALWGKVNV--DEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58  : \*:\* :\*: \*.\* \*\*\*\*.. .\* \*.\*\*\* \*::: :\* \*: :\* \*. \*.  HbA1\_polypeptide AQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAH 113  HbB\_polypeptide PKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHH 118  :\*\*.\*\*\*\*\* .\*:::.:\*\*:\*:: .::::\*\*:\*\*..\*\*:\*\*\* \*\*:\*\*.: \*: .\*\* \*  HbA1\_polypeptide LPAEFTPAVHASLDKFLASVSTVLTSKYR 142  HbB\_polypeptide FGKEFTPPVQAAYQKVVAGVANALAHKYH 147  : \*\*\*\* \*:\*: :\*.:\*.\*:..\*: \*\*: |

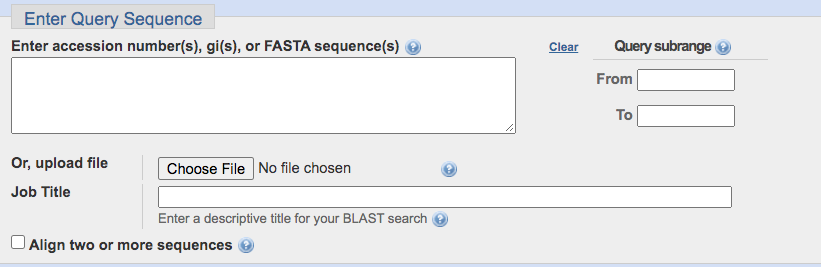
* + - 1. The asterisks (\*) indicate a sequence identity (same amino acid in the same position in both polypeptide sequences. A colon (:) indicates a conserved amino acid in the aligned position. A dash (-) indicates a gap.
    1. A second way to perform this alignment is to use the BLAST program group. This is a series of bioinformatic tools that allow you to search for similar sequences (we will do that later) and in this case, to compare two specific sequences.

BLAST stands for Basic Local Alignment Search Tool. This is a computer program that looks at ‘words’ to compare a short sequence of letters to another sequence and scores the match. The algorithm (program) ‘steps’ through the sequence building scores based on the match.

* 1. Go to [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) and click on the BLAST button (right hand side).
  2. On the next page, there are several different options for the BLAST program. Each option deals with different types of input sequences (nucleotides or amino acids) and compares to different sequence databases. We will be comparing the amino acid sequences of the two globin genes. Click BLASTp (BLAST for proteins).



* 1. Since we will be comparing just these two amino acid sequences, click on the button for “Align two or more sequences”.



* 1. Copy the polypeptide FASTA file (from above) and paste into each window (alternatively, you can paste the HbA1 FASTA sequence into the first window and the HbB FASTA sequence into the second window.
  2. Click on the BLAST button.
  3. The resulting file (after a short computer run) will provide you with different information (more on that later). The first (‘query’) sequence is the first sequence that you entered. It is most likely HbA1 (if you kept them in the order performed above). As you scroll down, you should see an alignment of the ‘query’ sequence with itself. It should be 100% aligned (as subject). This is indicated by a ‘consensus’ sequence in between the ‘query’ and ‘sbjct’ (subject) sequence.
  4. The second alignment is between the two globin polypeptide sequences. The ‘query’ will be the first sequence in the FASTA file and the ‘sbjct’ will be the second sequence. Copy and paste this alignment in the space below - adjusting font and size appropriately to fit.

|  |
| --- |
| Query 3 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQV 56  L+P +K+ V A WGKV + E G EAL R+ + +P T+ +F F DLS G+ +V Sbjct 4 LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV 61  Query 57 KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA 116  K HGKKV A ++ +AH+D++ + LS+LH KL VDP NF+LL + L+ LA H  Sbjct 62 KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK 121  Query 117 EFTPAVHASLDKFLASVSTVLTSKY 141  EFTP V A+ K +A V+ L KY Sbjct 122 EFTPPVQAAYQKVVAGVANALAHKY 146 |

* 1. In this alignment, the middle sequence represents the *consensus* sequence for the alignment. The ‘+’ represents conserved amino acids.

|  |
| --- |
| What does ‘consensus’ mean in relation to the alignment generated?  Consensus refers to the most common nucleotide or amino acid found at a position in an alignment. |
| What makes an amino acid change ‘conservative’?  A ‘conservative’ amino acid change involves an amino acid with similar biochemical properties. |

* 1. From the second alignment, collect the following information:

|  |  |
| --- | --- |
| What is the ‘expect’ value for this alignment (‘expect’ is also known as the e-value or e-score)? *The e-score will range between 0 and 1. The closer it is to 0, the better the match. This is influenced by the length of the alignment and the identities within that region. The e-value may be as small as 1e-200 (1x10-200) before it becomes 0.* | 3e-38 |
| What are the number of identities? *This represents the number of amino acids that are identical between the two sequences in the alignment.* | 63/145(43%) |
| What are the number of positives? *This represents the number of amino acids that are identical and conserved between the two sequences in the alignment.* | 88/145(60%) |

The **Expect value** (**e-value**) of the hit is defined by NCBI as follows:

“The Expect value (E) is a parameter that describes the number of hits one can ‘expect’ to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background noise. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance. The lower the E-value, or the closer it is to zero, the more ‘significant’ the match is.”

* + 1. Based on the sequence alignment, do you think that the α and ß-globin polypeptides are ancestrally related? Explain your answer.

|  |
| --- |
| It should be apparent to the students that the sequences do have an ancestral relationship based on sequence similarity (nucleic acid and amino acid). |

* + 1. Keeping in mind that the hemoglobin protein is naturally a tetramer comprised of two of each subunit (in adults, two α and two ß), do you think that the sequence homology of the α and ß-globin polypeptides might allow for a homotetramer to form? Explain your answer.

|  |
| --- |
| It may be difficult for students to get the answer, so generally this is viewed as more of a thought question and looking at the explanation. While the sequences are different, structurally the proteins are similar enough to allow for homotetramers to form and partially function. See the beginning of the next worksheet for more information. |