BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources

https://bioboot.github.io/bimm143_W18/lectures/#1

Dr. Barry Grant

Jan 2018

Overview: The purpose of this lab session is to introduce a range of bioinformatics databases and associated services available on the Web whilst investigating the molecular basis of a common human disease.

Sections 1 and 2 deal with querying and searching GenBank, GENE and OMIM databases at NCBI. Sections 3 and 4 provide exposure to EBI resources for comparing proteins and visualizing protein structures. Finally, section 5 provides an opportunity to explore these and other databases further with additional examples.

Side-note: The Web is a dynamic environment, where information is constantly added and removed. Servers "go down", links change without warning, etc. This can lead to "broken" links and results not being returned from services. Don't give up - give it a second go and try a search engine using terms related to the page you are trying to access.

Section 1

The following transcript was found to be abundant in a human patient's blood sample.

```
>example1
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTT
```

The only information you are given is the above sequence so you must begin your investigation with a sequence search - for this example we will use NCBI's **BLAST** service at: http://blast.ncbi.nlm.nih.gov/

Note that there are several different "basic BLAST" programs available at NCBI (including nucleotide BLAST, protein BLAST, and BLASTx).

Q1: Which BLAST program should we use in this case?
[HINT, what type of sequence are you provided with]

Searching against the "**Nucleotide collection**" (NR database) that includes GenBank is a good place to start your investigation of this sequence.

Q2: What are the names and accession numbers of the top four hits from your BLAST search?

Q3: What are the percent identities for the top few hits?
[HINT: scroll down to the alignment section of your BLAST result page for details of matched nucleotides]

Q4: How many identical and non identical nucleotides are there in your top hit compared to your last reported hit?

*From the results of your BLAST search you can link to the **GENE** entry for one of your top hits. This link is located under the "Related Information" heading at the right hand side of each displayed alignment (i.e. scroll down to the "Alignments" section).*

Q5: What is the "Official Symbol" and "Official Full Name" for this gene?

Q6: What chromosome is this gene located on?

Q7: What are the names of neighboring genes on this chromosome?

Q8: How many exons and introns are annotated for this gene?

Q9: What is the function of the encoded protein?

Q10: Does the protein have a role in human disease(s)? If so what diseases?
[HINT: Scroll down to the "Phenotypes" section of the GENE entry page and also explore the link to the OMIM database]

Section 2.

By now you should be aware that the example sequence corresponds to human sickle cell beta-globin mRNA and that this disease results from a point mutation in the β globin gene. In the following section, you will compare sickle cell and normal β globin sequences to reveal the nature of the sickle cell mutation at the protein level.

To do this you need to find at least one sequence representing the normal beta globin gene. Open a new window and visit the NCBI home page (http://www.ncbi.nlm.nih.gov)

and select “Nucleotide” from the drop menu associated with the top search box. Then enter the search term: **HBB**

Note that lots of irrelevant results are returned so lets apply some “Filters” (available by clicking in the left-hand sidebar) to focus on RefSeq entries for Homo sapiens.

Remember that we are after mRNA so we can compare to the mRNA sequence from section 1 above.

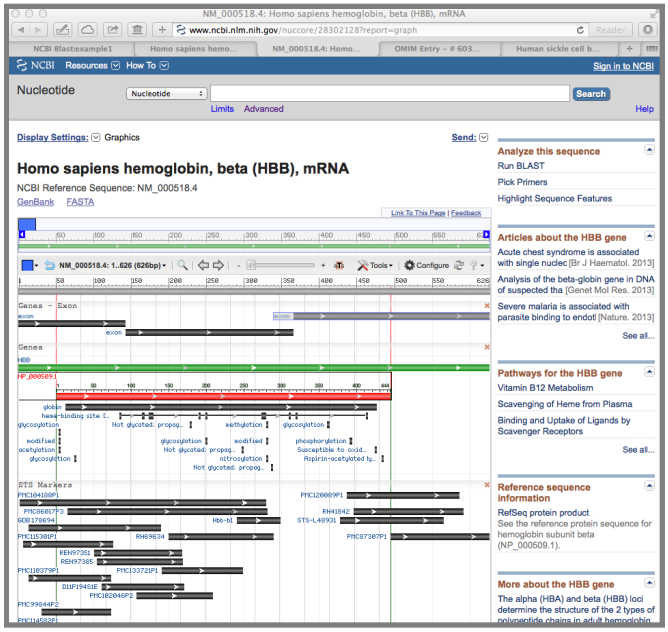
Q11. What is the ACCESSION number of the “Homo sapiens hemoglobin, beta (HBB), mRNA” entry?

NOTE: Boolean operators (NOT, AND, OR) as well as fielded queries (i.e. “`HBB[Gene Name] AND Human[Organism]`”) can be used in ENTREZ searches to filter results for more efficient searching.

Select “Homo sapiens hemoglobin, beta (HBB), mRNA” from the results and scroll down to the FEATURES section to answer the following.

Q12: What are the numbers of the first and last base positions of exon 1 of this entry?
[HINT: You can also find this from selecting the “GRAPHICS” display and placing your mouse over the first exon (see Figure).]

Q13: What are the numbers of the first and last base positions of the CDS?
[HINT: CDS or “coding sequence” refers to the portion of a genomic DNA sequence that is translated, from the start codon to the stop codon. Successful translation of a CDS results in the synthesis of a protein.]

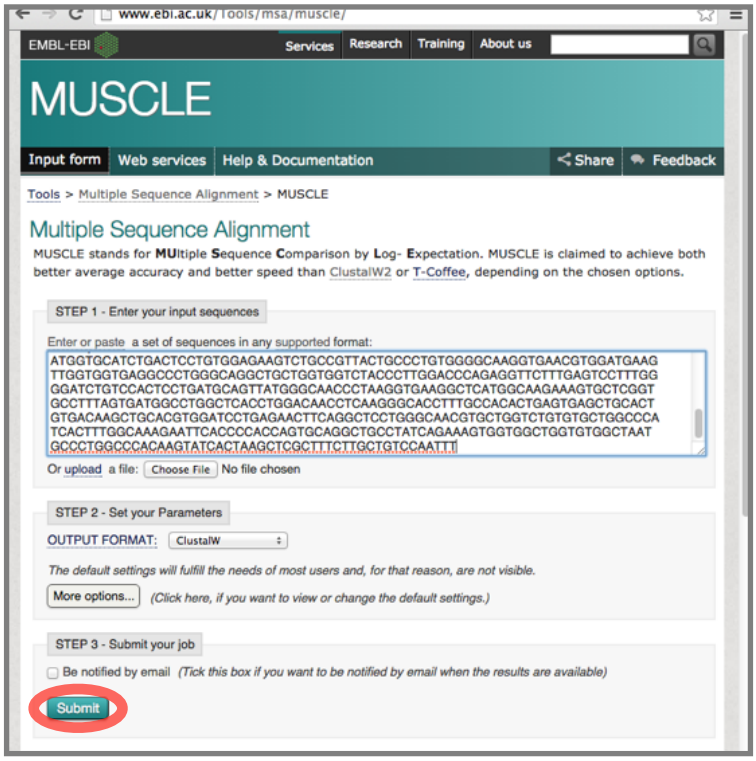


Section 3.

Here we will compare the retrieved sequences by creating a sequence alignment. This will make the difference between the two sequences easy to spot.

To generate the alignment, we will use **MUSCLE** available on the EBI website at: http://www.ebi.ac.uk/Tools/msa/muscle/

Select the FASTA display for the “Homo sapiens hemoglobin, beta (HBB), mRNA” (NM_000518) entry from section 2.



Now copy-and-paste this FASTA format sequence and also the example1 sequence from section 1 into the input box of the **MUSCLE** page. Then click the submit button (see red circle in Figure opposite).

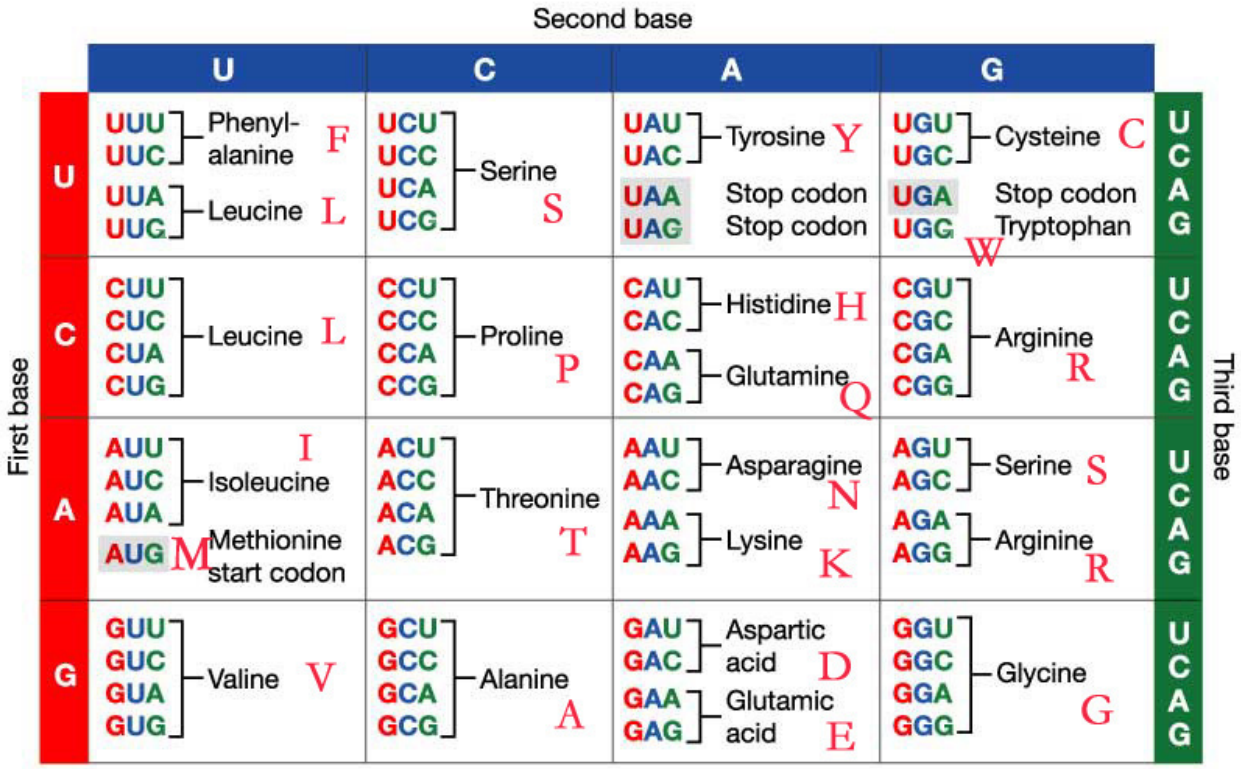
If your alignment is incomplete, please wait until the page refreshes. If the job appears to be in an undefined state try clicking refresh until a result is returned.

The two sequences should now be aligned. Where the aligned sequences are identical, an * is placed under the alignment. Examine the results and note that your sequences are nearly identical. However, being much shorter, the sickle cell sequence has many padding gap characters (-----) to bring equivalent regions into the correct register.

You can also click on the “Results Summary” tab and launch the **JalView** plugin to display a colored version alignment.

Q14: How many gap characters (-) are added to the beginning of the sickle cell beta-globin sequence in order to align it with the beta globin sequence? How might you have guessed this number from information you read in the GenBank annotation? [HINT: See section 2, Q13]

Q15: Ignoring ambiguity codes (Y and N), what are the differences between the two sequences? [HINT: There may be more than one]



Q16: Which codon position from the start of the sickle cell sequence would this difference affect? What amino acid would the different codons encode in the two sequences?
[HINT: use the codon table above to help.]

Section 4

In this section we will retrieve and visualize the 3D protein structure of sickle cell haemoglobin. The aim here is to ascertain how the Glu6 -> Val6 mutation might cause the mutant molecules to oligomerise into fibers, hence deforming erythrocytes. This will require you to examine the structural context of the mutation in the β globin chains.

We could find sickle cell haemoglobin structures via a text search of main PDB website @ http://www.rcsb.org/. However, as we know the nucleotide sequence from our previous work, lets use BLASTX to search the PDB database from the NCBI site.

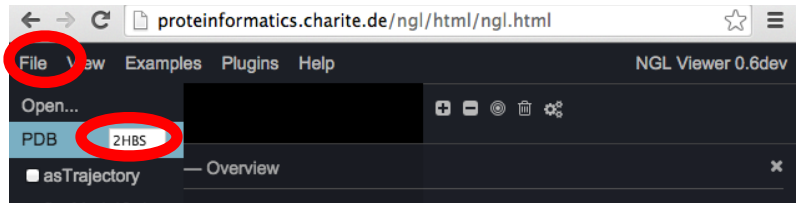
To do this visit http://blast.ncbi.nlm.nih.gov/ select the appropriate BLAST program and make sure the database you are searching against is set to "**Protein Data Bank (pdb)**".

Note the accession numbers and alignment statistics for the top few hits.

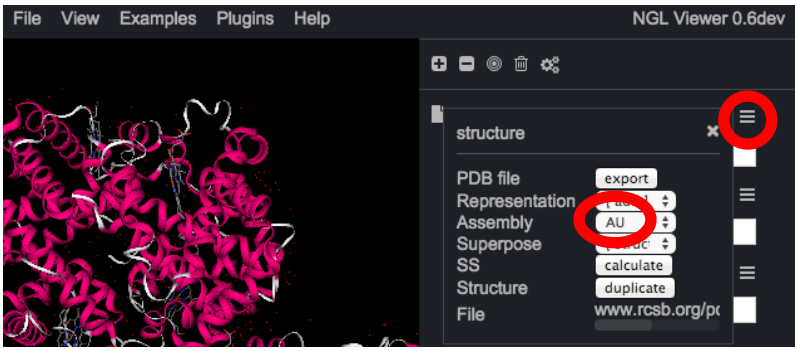
***Q17**: Is there a PDB structure with 100% identity to your `example1` query sequence?*

For this section we will use the online **NGL Viewer**, which has more advanced display options than the viewers currently available at NCBI or the PDB itself.

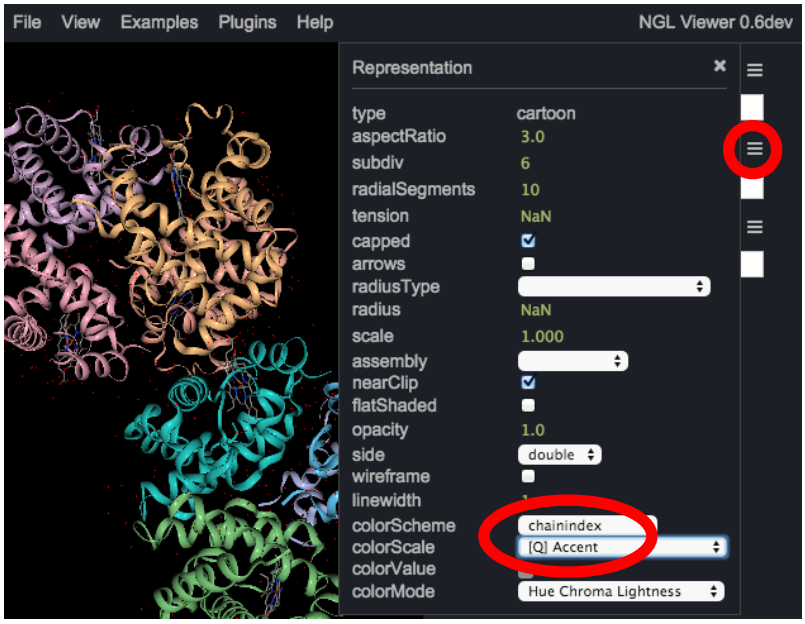
Open a new window, visit the webpage: http://nglviewer.org/ngl/



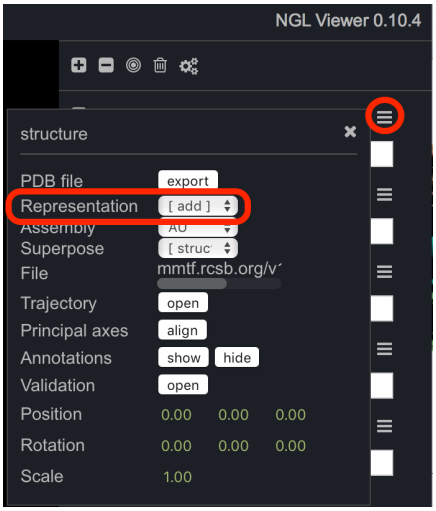
Once loaded, dismiss any splash screen instructions and click on the **File** button, and enter the **PDB code** 2HBS in the appropriate box and press return.



Left click the menu icon to the right of the listed entry (in our case "*2HBS*" as this was our PDB code) and set "**Assembly**" to "**AU**".

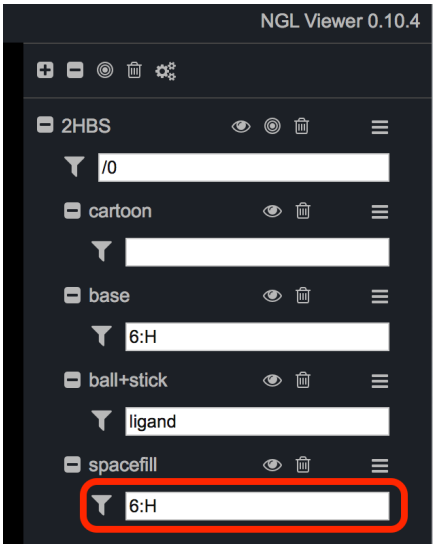


To color the structure by the chain, left click the second menu button on the right of the "*cartoon*" item to display the "*Representation*" menu options. Scroll down and set "**colorScheme**" to "**chainindex**", "**colorScale**" to "**[Q]Accent**".



Now lets add a new "*Representation*" to more clearly display the mutated residue. First click on the menu icon beside our loaded entry "2HBS". From the menu that appears click *Representation* **[add]** and select **spacefill** from the dropdown list of options.

This will result in all atoms of our entry being displayed as so called "sapce-fill spheres" with different atom types in different colors (e.g. oxygens in red, carbons in gray etc.)



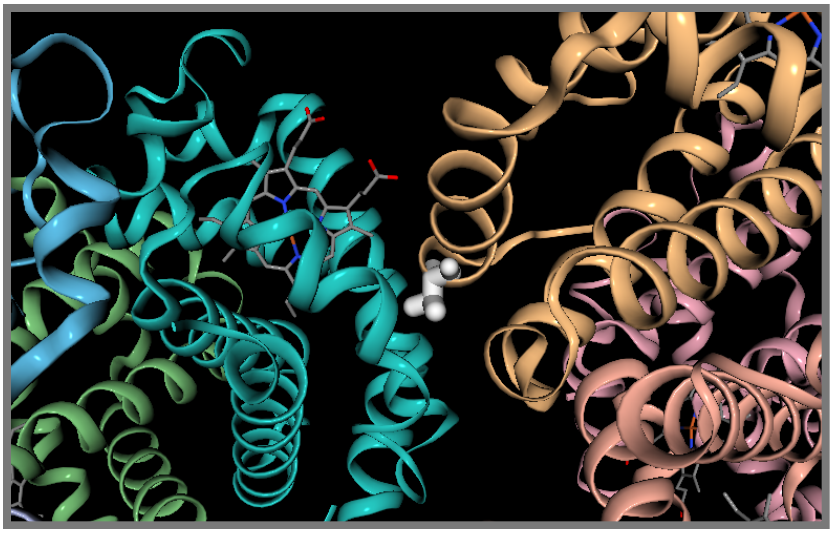
We now want to limit the atoms shown in "spacefill" representation to be be only those of our mutated amino acid residue (namely **Val 6** in **Chain H**).

In the white box beside the new "spacefill" item enter the selection text **6:H**

This will lead to only residue number 6 of chain H being rendered as spacefill. Play around with the settings from the spacefill menu and selection text until you have a reasonable feel for how the program works. Can you see mutated residue position?

Try zooming (via scrolling up and down) and rotating (via clicking and moving your mouse). You can always "reset" the view by clicking the target like circular icon. Also experiment with different settings and views.

***Q18**: What do you notice about the location of the Val6 residue in chain H of the 2HBS structure in relation to porphyrin?*
[HINT: see Figure below where I have used a white "licorice" representation for Val6.]



In this representation, one of the central mutant β chains is highlighted in orange ribbon. Also highlighted is the side chain of the E6V (i.e. Val6) mutation (white) and porphyrin prosthetic group (ball and stick representation).

NOTE: Some folks have reported issues using the NGL with older versions of the Chrome browser. The workaround is to use a different web browser. If, the structure is still not displayed correctly for you, download its coordinates from the **PDB** database at: http://www.rcsb.org/ and **ask for assistance**.

If deemed appropriate, and you are working on your own computer, you may consider updating your version of JAVA by downloading from:
https://www.java.com/en/download/manual.jsp

Discussion:

The original paper discussing the 2HBS crystal structure is available online:

http://www.sciencedirect.com/science/article/pii/S0022283697912535

In this article, Figure 3 demonstrates how the Glu6->Val6 mutation could result in the characteristic "sickle" phenotype. The charged Glu6 mutating to Val6 creates a superficial hydrophobic patch on one HbS molecule that interacts with hydrophobic surface residues of another. The molecules thus polymerize, creating extended fibers that distort the shape of the red blood cell.

Assessment of the disparate biochemical properties of normal and sickle haemgolobin, together with microscopy studies showing long crystal fibres inside sickle cells, led Linus Pauling (1949) to (correctly) predict the morphological effects of these changes.

The abnormal sickle form causes the cells to clump together, hampering their passage through blood vessels, depriving tissues of oxygen. See this YouTube video for an illustration: http://www.youtube.com/watch?v=Qd0HrY2NlwY

The sickled blood cells have a short lifetime and cannot be replaced fast enough, leading to chronic anaemia. Sickle cell anemia was one of the first diseases to be linked to a defect at the molecular level, providing a clear demonstration that a single base mutation can change a single amino acid, which in turn can result in a defective protein.

***Q19**: What one part of this exercise or associated lecture material is still confusing? If appropriate please also indicate the question number from this document and answer the question in the following anonymous form:*
https://tinyurl.com/bimm143-01

[Your comments will let us know which material needs to be further clarified and will help us gain stronger control of the material in this course. Thank you!]

Section 5 (Optional)
Pick one of the following three genes to investigate. Again the only identifying information you are given is a nucleotide or peptide sequence. Use the various NCBI and EBI resources to answer questions 5 to 10 from section 1. However, do not limit yourself to these five questions as you may find other directions of exploration more interesting. As always, please ask for assistance if you get stuck.

Gene 1 – the following cDNA sequence was found to be expressed abundantly in human adipocytes (fat cells).

```
>Transcript 1
gccactgccaacatttcccttcttccagttgcactattctgagggaaaatctgacaccta
agaaatttactgtgaaaaagcattttaaaaagaaaaggttttagaatatgatctatttta
tgcatattgtttataaagacacatttacaatttacttttaatattaaaaattaccatatt
atgaaattgctgatagta
```

Gene 2 – the following cDNA sequence maps to a human genomic location identified by mapping as being of interest.

```
>Transcript 2
```

ctgcgagaagagcagcgacacttgcaaccccctgtcaggcgccttctcaggagtgtccaa
cattttcagcttctgggggacagtcggggccgccagtaccaggagctccctcgatgccc
cgccccaccccagcctcctcaacatcccctctccagcccgggtcggcggccccgggg
cgacgtggagagcaggctggatgccctccagcgccagctcaacaggctggagacccggct
gagtgcagacatggccactgtcctgcagctgctacagaggcagatgacgctggtcccgcc

Appendix

>gi|179408|gb|M25079.1|HUMBETGLA Human sickle cell beta-globin mRNA
ATGGTNCAYYTNACNCCNGTGGAGAAGTCYGCYGTNACNGCNCTNTGGGGYAAGGTNAAYGTGGATGAAG
YYGGYGGYGAGGCCCTGGGCAGNCTGCTNGTGGTCTACCCTTGGACCCAGAGGTTCTTNGANTCNTTYGG
GGATCTGNNNACNCCNGANGCAGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTNCAYGTGGATCCTGAGAACTTCAGGCTNCTNGGCAACGTGYTNGTCTGYGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACCAGTGCANGCNGCCTATCAGAAAGTGGTNGCTGGTGTNGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTNGCYTTYTTGYTGTCCAATTT

>gi|28302128|ref|NM_000518.4| Homo sapiens hemoglobin, beta (HBB), mRNA
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAG
TCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGT
CTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGG
TGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTT
GCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGT
CTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTG
TGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTC
CCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTA
TTTTCATTGC

http://www.rcsb.org/pdb/files/2hbs.pdb

The mutation causing sickle cell anemia is a single nucleotide substitution (A to T) in the codon for amino acid 6. The change converts a glutamic acid codon (GAG) to a valine codon (GTG). Changing a hydrophilic amino acid to a hydrophobic one, see http://themedicalbiochemistrypage.org/sicklecellanemia.php

Note there is also a T -> A difference at position 162 (162/3 => codon 54 GCT -> GCA). This is in the third position of the codon and hence does not change the corresponding amino-acid.