A practical introduction to Bioinformatics.

Step by step instructions.

1. Background and Objectives:

The prime objective of this practical is to use the computer to investigate the molecular biological causes of sickle cell anaemia.

As you should already know, sickle cell anaemia is caused by a point mutation in the beta globin gene. A single base pair change near the beginning of this gene gives rise to a single amino acid change in the beta globin protein which in turn causes the disease.

In this practical you will use the computer to locate both sickle cell and normal beta globin sequences. By comparing the sequences you will see the precise nature and location of the sickle cell mutation. By computing restriction maps of both the sickle cell and normal DNA sequences, you will design experiments to detect the mutated sequence in a patient. By searching reference databases, you will be able to confirm and extend the information you discovered by analysis of the sequence data. Finally, you will locate and visualize the 3D structure of human haemoglobin. You will be led through a series of manipulations of this image that will make clear exactly how such a small mutation in the beta globin protein causes such catastrophic effects.

2. An Overview of the practical:

The stages of the practical are:

- Find DNA sequences of human beta globin both with and without the sickle cell point mutation.
- Compare the beta globin DNA sequences in a way that enable you to see clearly the position and nature of the sickle cell point mutation.
- Seek a method of detecting the sickle cell mutation using restriction enzymes.
- Search the reference databases to confirm and extend the information computed from the sequence data.
- Search the 3D protein structure databases for the structure of human haemoglobin. Visualize this structure and investigate the position and effect of the sickle cell mutation.

As you go through the practical, you will be invited to answer a series of questions. Individually, the questions will generally not require deep thought, but they will draw you attention to fragments of information that will together lead to a balanced and complete picture. The questions appear in shaded italic text boxes with a clear section provided for your answers. At the end of the practical session, a yellow sheet with the "official" answers will be available. Text indicating that you should take some action will be shaded thus.

In addition to a **Web Browser**, the software tools you will use in the course of the practical are:

- **SRS** (**S**equence **R**etrieval **S**ystem). This you will use to search some of the many molecular biological databases available to researchers today, to access data items stored in these databases and to see how these data items are interlinked.
- The program **clustal** to compare families of DNA sequences.

- A program, integrated with **SRS**, for computing restriction maps.
- The program **rasmol** to view the three-dimensional structure of the proteins relevant to the exercise.

3. Getting started and finding the sickle cell beta globin DNA sequence

I am going to assume that you are already logged into your workstation and that a suitable[1] **Web Browser** is running. If this is not the case, and you do not know how to proceed, let me know.

First, go to "**SRS** at the **EBI**" (if this web site is not already in view, try your **Bookmarks**, as a last resort, type in the URL "**srs.ebi.ac.uk**").

Once there, click on the **Library Page** tab. You will arrive at a page listing the databases available for you to search.

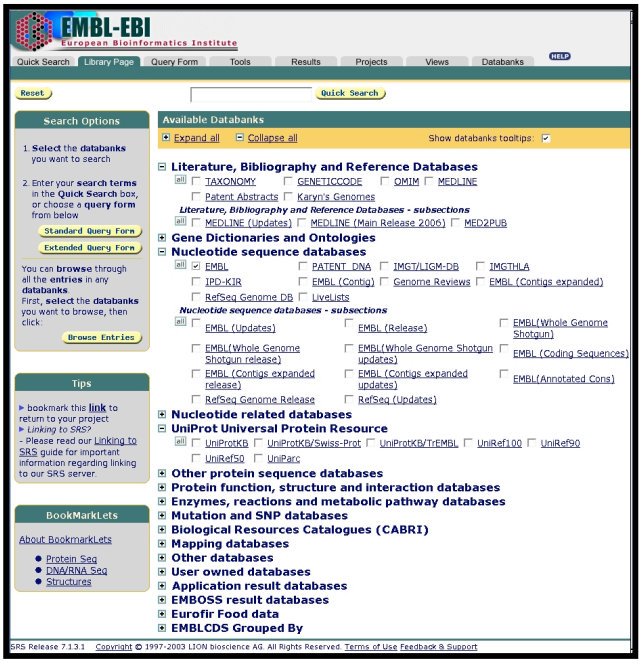
1. To select the database you want to search, you need to click on the box by the name of the database. To find out what the database might be you would click or hover over the name of the database.

a) *What sort of data is stored in the **EMBL** database and where is it maintained?*

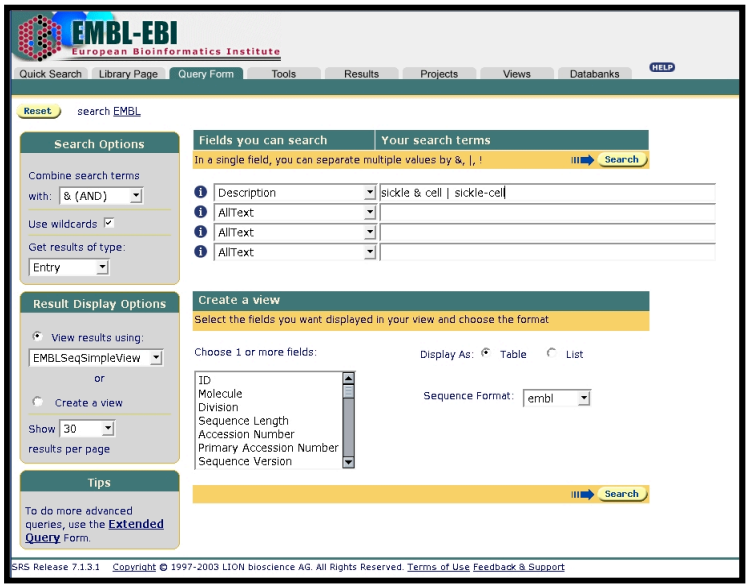
b) *What sort of data is stored in the **UniProtKB/Swiss-Prot**?*

*Use the **Web Browser Back** button to return from the database information pages to your database selection page.*

[1] I will assume that your Web Browser is not too dissimilar from those installed on my machines. If this is not this case, you will need to be flexible.



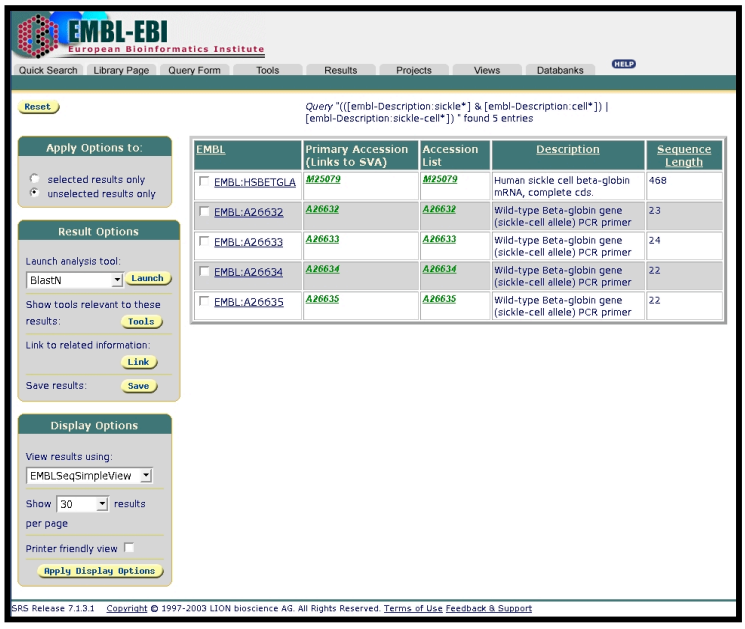
First, you want to look for the DNA sequence(s) that relate to sickle cell anaemia, so select the **EMBL** database by ticking the box next to the database name and click on the **Standard Query Form** button.



Then select **Description** from the pull down menu associated with the first search field and type in the keywords[2]:

`sickle & cell | sickle-cell`

You are asking **SRS** to search for all **EMBL** database entries that contain the words **sickle** AND **cell** in the section of their annotation intended to be a succinct description of the entry. Click on the **Search** button to start the search going.



After a short pause, **SRS** should deliver a short list of hits. For each hit, **SRS** will show you the database name (consistently **EMBL** as this was the database searched), the entry name, the accession code (a more formal label for the entry, sometimes the same as the entry name), the description line (the portion of the entry annotation that you searched) and the sequence length.

[2] The | symbol, which stands for the boolean operator "**OR**" in this context, is immediately to the left of the **Z** on your keyboards. This is easy to miss as it is represented by a broken vertical line even though it displays and prints as an unbroken vertical line?

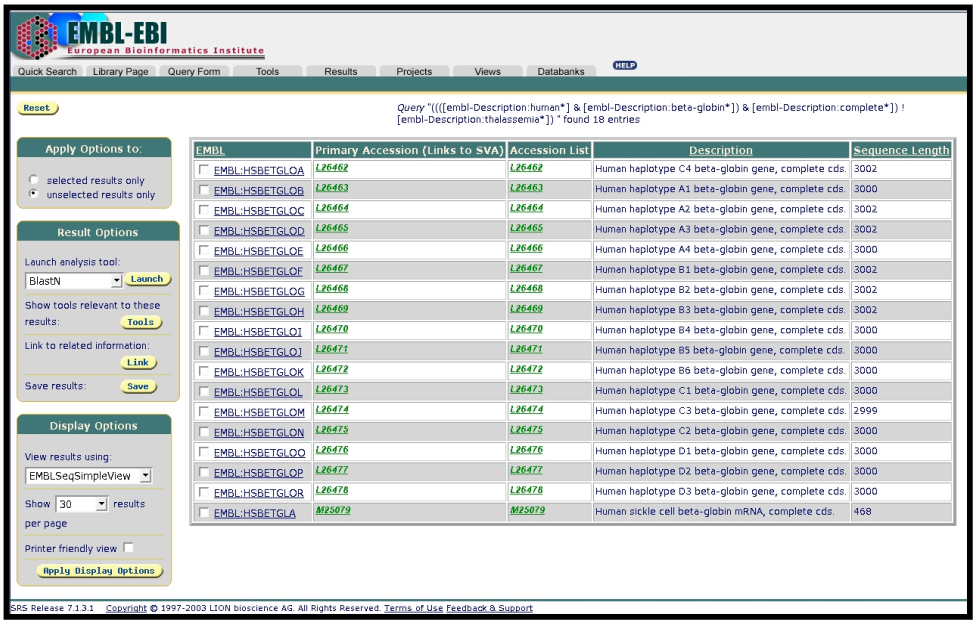
The longest entry returned is clearly the most promising. It claims to be messenger RNA generated from the complete Human sickle cell beta globin gene[3]. Click on the link of the hit and **SRS** will show you the **EMBL** entry itself. Look for the Description header near the top and confirm that this entry matched your keywords.

At the bottom of the page you will see the sequence itself. Note that the sequence includes many of the characters **n** and **y**. These are ambiguity codes. An **n** means a**n**y base and a **y** means either of the p**y**rimidines (i.e. **t** or **c**). If you look at the reference for this (under the reference header just above the sequence), you will see that it was published way back in 1976, which probably explains the poor quality of the data. Unfortunately, as there is no entirely unambiguous sequence in the EMBL database that has the properties we require, this will have to do.

4. Finding a family of beta human beta globin DNA sequences, including the sickle cell mutation.

Having found an mRNA sequence with the sickle cell mutation, you now need to find at least one sequence representing the normal beta globin gene. Go back to the page from which you started your search. From the **EMBL** entry, this can be done by clicking on the **Web Browser Back** button twice. After ensuring you are still searching only the **Descripton** annotation field, change your search to[4]:

`human & beta-globin & complete ! thalassemia`



This time, you are asking for **EMBL** entries that have the words **human** and **beta-globin** and **complete** in their **Description** line(s) ***but not*** the word **thalassemia**. Set the search going by clicking on the **Search** button once more. You should find a few more hits this time.

In the list will be the mRNA of the sickle cell sequence once again. Also you should have 17 haplotypes of the genomic sequence for beta globin in human[5].

[3] The entry "Human sickle cell beta-globin mRNA complete cds". **cds** is the accepted shorthand for **co**d**i**ng **s**equence.

[4] This combination of search terms may seem a bit arbitrary. It is not a first guess, I took several attempts to design this search, which generates exactly the sequences you need for this exercise. In "real life" you would expect to have to try several searches before finding exactly what you require.

[5] The 17 haplotype sequences resulted from a study of 36 Melanesian people, **none of whom had sickle cell anaemia**. A region of genomic DNA including the entire beta globin gene was taken from each and sequenced. These are the 17 slightly varying DNA sequences (or **haplotypes**) that were discovered amongst the group.

2. *Look at one of the haplotype entries by clicking on its **EMBL** link. Scroll down to the feature table. You will see some sections of the table give information about the **variation**s (alleles) that define the haplotype and others that show the positions of the **intron**s and **exon**s within the sequence (as the haplotype sequences are all genomic). If you look at more than one entry, you will notice that the **variation**s are different for each entry whereas the reference to the **UniProtKB/Swiss-Prot** entry (In the **cds** section under **db_xref**) is the same in each case. The numbers of **exons** and **introns** are also the same for all haplotypes.*

   a) *What is the entry name for the **UniProtKB/Swiss-Prot** entry referenced by all haplotypes?*

   b) *How would you explain the differences between the 17 haplotypes, given that the amino acid sequence for each appears to be the same (implied by the consistent **UniProtKB/Swiss-Prot** reference)?*

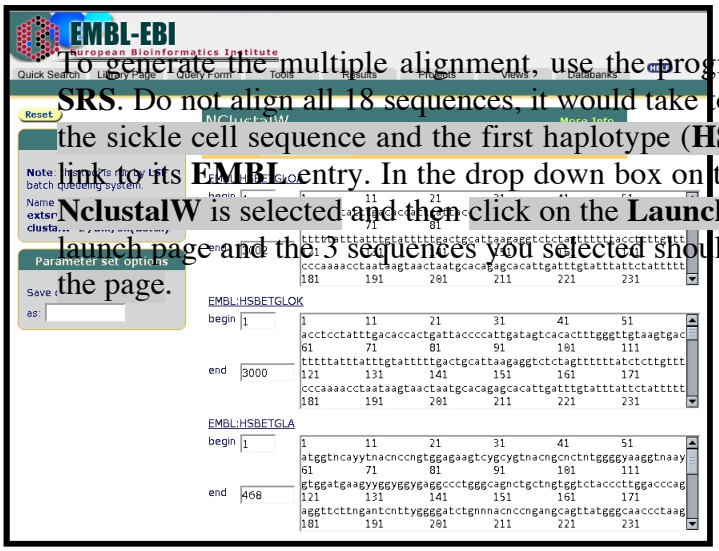
   c) *How many exons do all of the haplotype entries have?*

   d) *Look at the entry **HSBETGLOA** (should be first in the list). What are the numbers of the start and finish base positions of **exon 1**?*

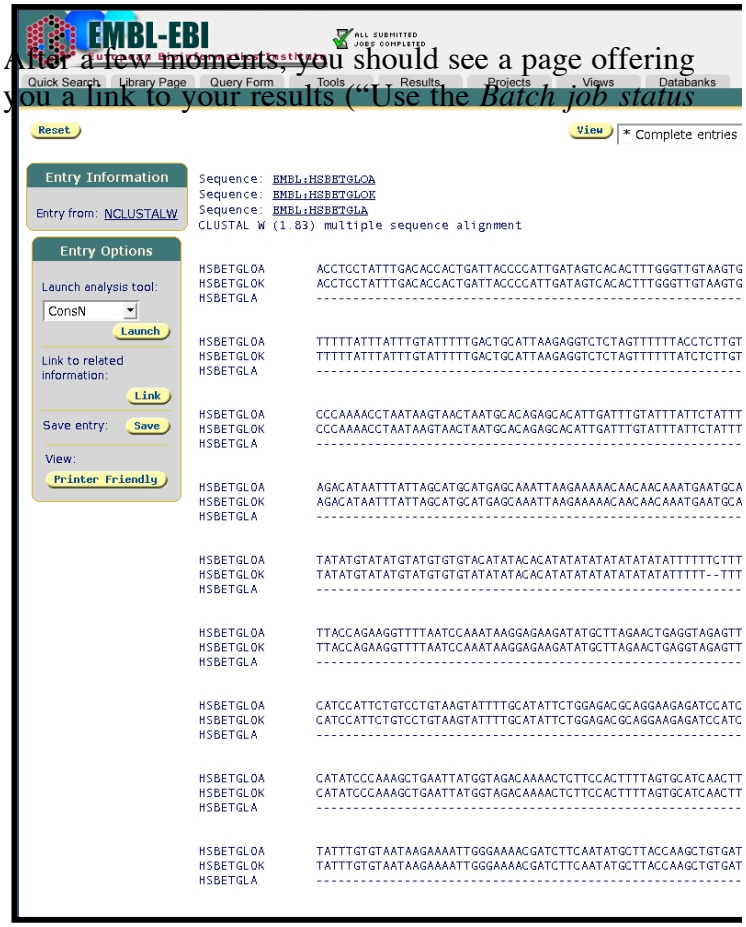
   e) *The Feature Section includes links to the region of sequence to which they refer. Click on the links to individual **intron** and **exon** Features (use the **Web Browser Back** button to return to the main entry each time). What do you notice about the bases at the beginning and end of the introns that you might have expected?*

5. Looking for the sickle cell point mutation by aligning the beta globin DNA sequences.

Now to compare some of the sequences you have retrieved from the database with each other. The prime objective here is to align the sequences in a way that will make the difference between the sickle cell sequence and the others easy to spot. Such an alignment of more than two sequences is called a "multiple alignment".



To generate the multiple alignment, use the program **NclustalW** (see below), which is available from **SRS**. Do not align all 18 sequences, it would take too long, instead select any set of 3 sequences including the sickle cell sequence and the first haplotype (**HSBETGLOA**) by clicking on the square button by the link to its **EMBL** entry. In the drop down box on the left of the **Launch** button, Ensure that the program **NclustalW** is selected and then click on the **Launch** button. You should now be looking at the **NclustalW** launch page and the 3 sequences you selected should be in view[6]. Click on the **Launch** button at the top of the page.



After a few moments, you should see a page offering you a link to your results ("Use the *Batch job status page* to view the **results**"). For the best chance of success, count to ten slowly and then click on this link. You should then see a reference to your alignment including a **a** link to your results[7]. Click on the **a**.

Now you should see your 3 sequences aligned. Where all three of your aligned sequences are identical, an * is placed under the alignment. Move up and down the alignment display. Note that your 2 haplotypes are pretty much identical. Note also that where your sickle cell sequence matches the others it matches well (as indicated by the *s). However, being a much shorter sequence, the sickle cell sequence needed the introduction of many padding characters (minus signs) in order for the matching regions to be displayed.

[7] If you see any symbol other than a **a**, count to 10 again and then click on your browser's **Reload** button. If that does not work, ask for help.

3. *Looking at the multiple alignment of your DNA sequences you should see that sections of the mRNA sickle cell sequence have been aligned convincingly with sections of the haplotype alignment.*

   *Into how many sections is the mRNA sickle cell sequence split in order to align with the haplotype sequences? How might you have guessed this number from information you read in the annotation of the haplotype **EMBL** entries {**Hint:** refer back to your answer to question **2c**}?*

4. *Look particularly at the first few bases (no more than 30) of the first region where the sickle cell sequence is aligned with the rest.*

   a) *Ignoring ambiguity codes (**Y** and **N**), what single difference can you see between the sickle cell sequence and the others?*

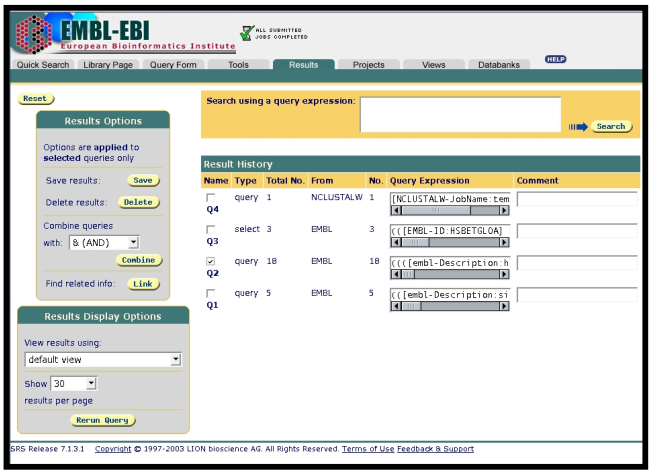
   b) *Which codon (in terms of "How many from the start of the sequence?") of the sickle cell sequence would this difference affect?*

   c) *What amino acid would the codon code for in the haplotype (see Appendix for Amino acid codes)?*

   d) *What amino acid would the codon code for in the sickle cell sequence?*

6. Looking for a diagnostic method for detecting the sickle mutation using restriction mapping.

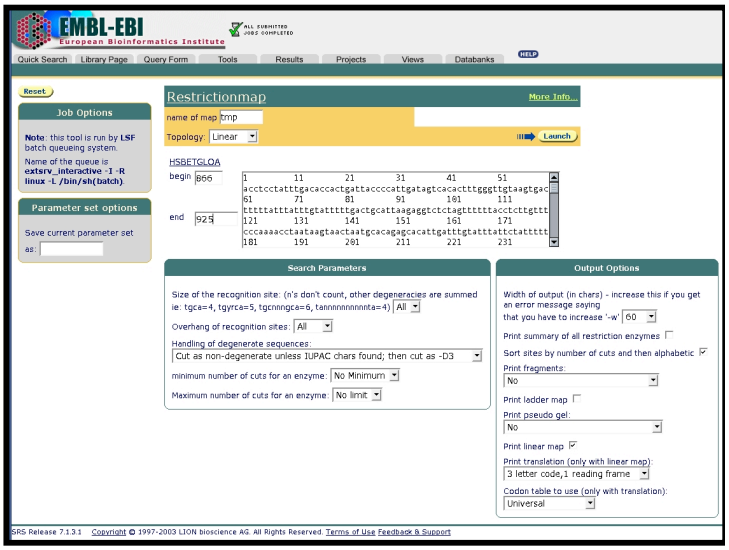
Once you have admired your **NclustalW** results sufficiently and answered all the relevant questions, you should have a clear idea of what is different, at the DNA level, about the sickle cell sequence. The next step is to use the computer to predict where various restriction enzymes would cut the sequences you have compared in the region of their vital difference. If you can discover restriction enzymes that cut the normal beta globin gene differently to the way they cut the sickle cell sequence, you may be in a position to suggest a way to detect the sickle cell mutation.



First, return to your list of **EMBL** beta globin entries. To do this, click on the **Results** tab at the top of your current **SRS** page. You will be offered a list of all the results of analyses and searches you have executed to date. Select the query of **EMBL** that generated the list of 18 sequences (**Q2,** if you have followed the instructions exactly). Then click on the **Rerun Query** button and you will be returned to your list of 17 haplotypes plus the sickle cell mRNA.

Now the plan is to discover which restriction enzymes would cut the most interesting part[8] of your selected sequences and to compute exactly where those cuts would occur. So, from the pull down menu just below the **Launch** button, select the **Restrictionmap** option (from the drop down box on the left of the **Launch** button). Next select the **HSBETGLOA** sequence by clicking on the button next to the link to its **EMBL** entry. Then invoke the restriction mapping program by clicking on the **Launch** button.

SRS offers to generate a restriction map of the whole of **HSBETGLOA**. Instead, just ask for the first 20 codons of the first exon of this sequence. Set the **begin** and **end** values according to what you previously discovered about this sequence (refer to your answer to **Question 2d**[9]).



Set the **Restrictionmap** program rolling by clicking on the **Launch** button. After a few moments, you should see a page offering you a link to your results ("Use the *Batch job status page* to view the **results**"). As before, wait a few seconds and then click on this link. You should then see a reference to your restriction map including a **a** link to your results[10]. **With your right hand mouse button**, click on the **a**. From the proffered list, select the option **Open Link in new window**. Your map of a normal beta globin DNA should appear in a new browser window.

The aim is now to generate the restriction map for the 20 codons of the sickle cell sequence (i.e. the 20 codons corresponding to those mapped in **HSBETGLOA**) in a separate window for easy comparison. The obvious way to do this would be by selecting the sickle cell **EMBL** entry and giving the relevant portion to **Restrictionmap**. However, this would not produce satisfactory results because of the considerable ambiguity of its sequence (i.e. the large number of "**y**"s and "**n**"s). A far better plan is to edit the relevant part of **HSBETGLOA** so that it looked like a non-ambiguous version of the mutated sickle cell sequence and then to map that. As you will have discovered previously, this entails editing just 1 base pair (to remind yourself which base pair to edit, refer to your answer to **Question 4a**[11]).

---

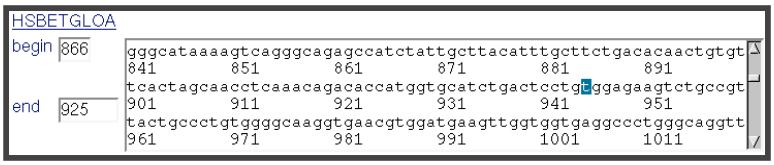
[8] I.e. the region of sequence including the sickle cell mutation.

[9] If you have with good eyesight, read the required values from the little picture by the side of this paragraph. It says begin at **866**, end at **925**.

[10] Again, if you see any symbol other than a **a**, wait a while and then click on your browser's **Reload** button. If that does not work, ask for help.

[11] Once more, if you need it, I offer a pictorial clue. Even more explicitly, you need to change base **885** from an **A** to a **T**. Note that the base numbers are positioned **ABOVE** the bases and the most significant digit is directly above the numbered base.

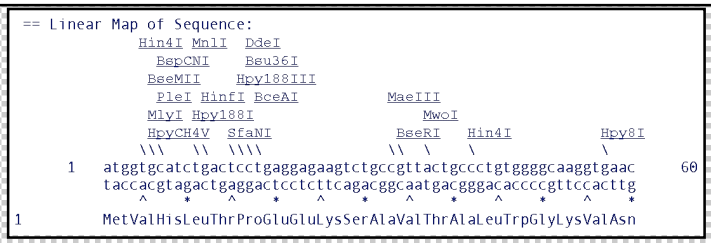
Accordingly, to compute the restriction map of the mutated sequence, move back to your original browser window and click on the **Results** tab at the top of the page. As you did previously from this page, select the query of **EMBL** that generated the list of 18 sequences (probably **Q2**) and then click on the **Rerun Query** button. Select again the **Restrictionmap** option (from the drop down box on the left of the **Launch** button). Select **HSBETGLOA**. Invoke the restriction mapping program by clicking on the **Launch** button. The **Restrictionmap** launch page will now appear displaying the **HSBETGLOA** sequence.



In the display of the **HSBETGLOA** sequence, scroll down to the mutated region and make the edit required to introduce the sickle cell mutation. Then, as you did for your first map, change the **begin** and **end** values to ask that only the first 20 coding **codons** be mapped (see illustration for guidance). Now click on the **Launch** button once again. . When you again see the page offering you a link to your result, wait a few seconds and then click on the link. You should arrive once more at a list of all your analysis thus far. Click on the **a** that corresponds to your second restriction map .Your restriction map for the mutated sequence should then appear.

You should now have two **Web Browser** windows each displaying a restriction map of the first 20 codons of coding sequence of human beta globin. One mapped section of sequence will include a point mutation that causes sickle cell anaemia, the other will not. Position your maps so that you can see as much of both as is possible. I suggest you start by looking at the "Linear maps of the sequence" which are at the bottom of the outputs. The differences between the two maps are the important things to spot and understand.

5. *Compare first the two "Linear maps" you have generated. Each map is displaying the 20 codons (60 base pairs) that you have mapped (both strands). Above the base pairs are displayed the names of the enzymes that cut that portion of sequence.*



'\' Characters are used to indicated the precise location of the cuts. Directly underneath the DNA sequence display, '' and '^' Characters are placed to make it easy to count along the sequence. Right at the bottom of the display, the amino acid translation of the mapped 20 codons is displayed in three letter codes.*

a) *What single difference is there between the amino acid translations of the two maps'?*

b) *Are there enzymes that will cut the sickle cell sequence but will not cut the haplotype sequence?*

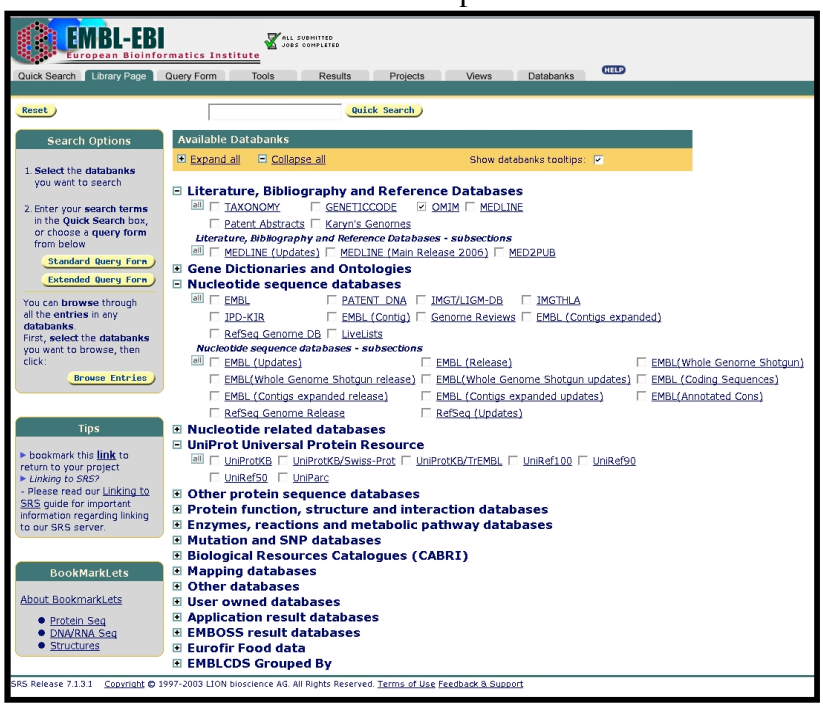
c) *How many enzymes will cut the haplotype sequence but will not cut the sickle cell sequence? Name two of them.*

d) *How might you use the fact that an enzyme cuts one version of the sequence but not the other to test for sickle cell anaemia?*

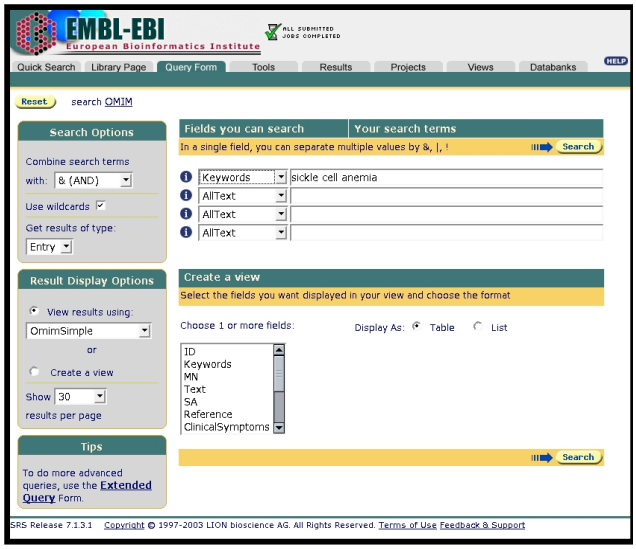
7. Finding answers from the references databases.

Having fully examined your restriction maps, and answered all the relevant questions, you should have at least one suggestion for a diagnostic test for sickle cell anaemia using restriction enzymes. This you achieved by retrieving, examining and analyzing molecular biological data from the highly interlinked databases available over the network.

It should not surprise you that others have looked at the same data as you considered today and have come to similar conclusions. Further, they have investigated their observations properly and published their results. It should therefore be possible to arrive at a condition of enlightenment very similar to that which you have achieved, by simply searching the literature. This is what we will do now.



A clean start is required, so click on the **Library Page** tab of your current **SRS** page. You will be returned right back to the **SRS** database selection page. This time, you want to select the **OMIM**[12] database. This is a database of human diseases caused by genetic defects. **OMIM** is in the **Literature, Bibliography and Reference Databases** section. Click on the button beside the word **OMIM** to select the database for searching. Finally, click on the **Standard Query Form** button to request the **Standard** Query form.



Change the search type of the first query field from **AllText** to **Keywords** and enter the search term[13]:

`sickle cell anemia`

Then click the **Search** button and you should be rewarded by a list of 2 **OMIM** entries. Both are very interesting, but the one labeled "**OMIM603903**" and with the title "**SICKLE CELL ANEMIA**" is what were after. Click on it and you will be able to read the **OMIM** entry.

[12] **O**nline **M**endelian **I**nheritance in **M**an

[13] No, **anemia** is not a typing mistake. **OMIM** is an American product. Note also that the lack of **&** characters is deliberate. **Keywords** are one of the few searchable items that are allowed to be more than one word in **SRS**.

6. *You should now be looking at the **OMIM** entry for **SICKLE CELL ANEMIA. OMIM** entries are comprised of short précis of papers relating to human phenotypes with genetic causes. They provide the user with a quick and relatively painless way of discovering what research has been undertaken in a particular field without having to read all the literature. They often include links to the database entries describing the allele(s) that give rise to the phenotype. Allele descriptions are organized in a separate database called **OMIM_allele**.*

a) *What is the mutation that most commonly causes sickle cell anaemia?*

b) *What is the number of the **OMIM_allele** entry for this mutation?*

Once you have read all you need to of this entry, and have answered all the relevant questions, click on the link to the **OMIM_allele** entry[14] that corresponds to the most common cause of sickle cell anaemia. You will be rewarded by an entry from a database closely related to **OMIM** that contains information about the specific alleles that cause the phenotypes in the **OMIM** database.

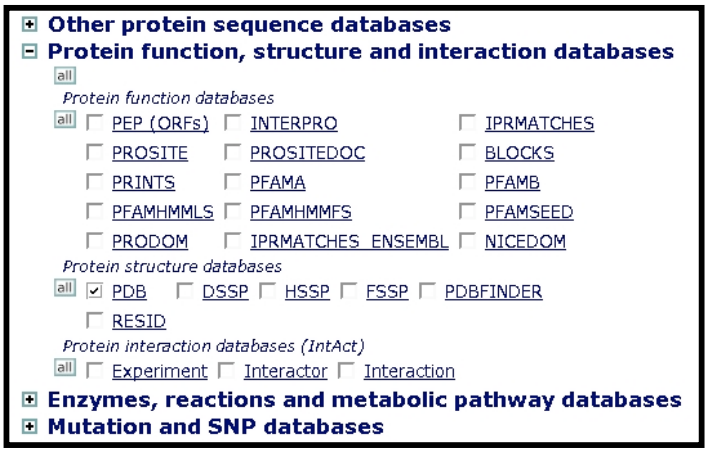
7. *You should now be looking at the **OMIM_allele** entry for the mutation that most commonly causes sickle cell anaemia. Like **OMIM** entries, **OMIM_allele** entries are comprised of short précis of papers.*

The paper mentions a number of restriction enzymes that have been use in the diagnosis of sickle cell anaemia. Name those whose usefulness relies on the sickle cell mutation eliminating one of their cut sites (see paragraph 3).

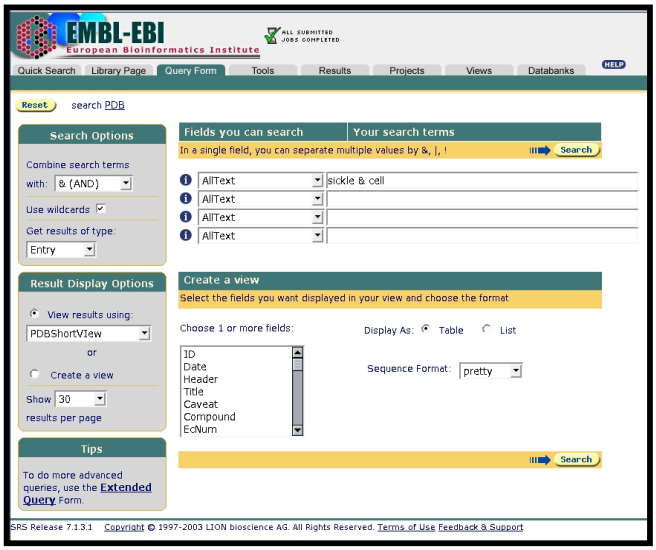
[14] The link to the **OMIM_allelle** entry is the 6 digit number followed by a 4 digit number with a "**.**" In the middle that you (hopefully) wrote down in answer to **Question 6b**.

8. Finding the crystal structure of Sickle cell Haemoglobin.

The tertiary structure of sickle cell haemoglobin is known and is stored in a database of protein 3D structures. With the aid of a structure-viewing program called **rasmol** you should be able to also see how the position and nature of the mutation causes the problems that it does.



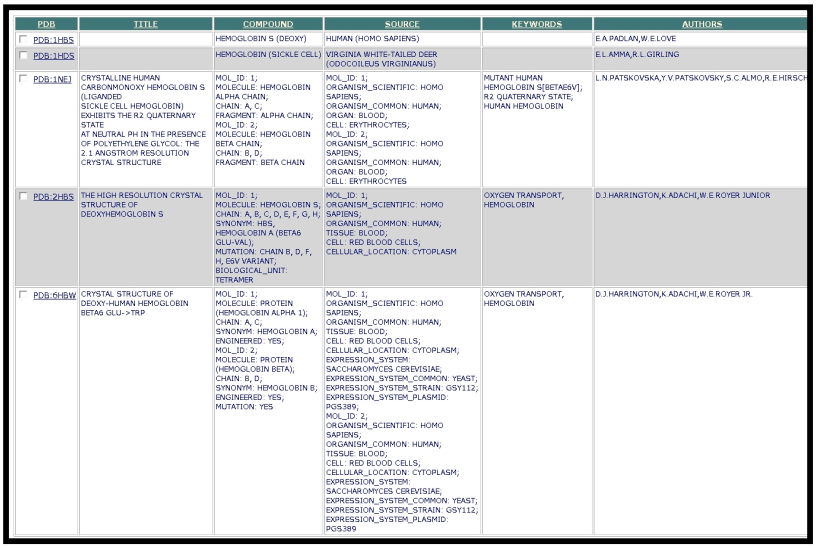
First, you must retrieve the beta globin 3D structure from the appropriate database. Another clean start required, so click on the **Library Page** tab to go back to the database selection page of **SRS**. The 3D structure database you need to search is called **PDB**[15]. It is in the **Protein function, structure and interaction databases** folder, so click on the appropriate **+** sign. Once the **Protein3Dstruct** databases are in view, select **PDB** by clicking in the box to the left of its name and then click on the **Standard Query Form** button to continue.



Leave the search type of the first query field as **AllText** and enter the search terms:

```
sickle & cell
```

and press the **Search** button.



PDB	TITLE	COMPOUND	SOURCE	KEYWORDS	AUTHORS
PDB:1HBS		HEMOGLOBIN S (DEOXY)	HUMAN (HOMO SAPIENS)		E.A.PADLAN,W.E.LOVE
PDB:1HDS		HEMOGLOBIN (SICKLE CELL)	VIRGINIA WHITE-TAILED DEER (ODOCOILEUS VIRGINIANUS)		E.L.AMMA,R.L.GIRLING
PDB:1NEJ	CRYSTALLINE HUMAN CARBONMONOXY HEMOGLOBIN S (LIGANDED SICKLE CELL HEMOGLOBIN) EXHIBITS THE R2 QUATERNARY STATE AT NEUTRAL PH IN THE PRESENCE OF POLYETHYLENE GLYCOL: THE 2.1 ANGSTROM RESOLUTION CRYSTAL STRUCTURE	MOL_ID: 1; MOLECULE: HEMOGLOBIN ALPHA CHAIN; CHAIN: A, C; FRAGMENT: ALPHA CHAIN; MOL_ID: 2; MOLECULE: HEMOGLOBIN BETA CHAIN; CHAIN: B, D; FRAGMENT: BETA CHAIN	MOL_ID: 1; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_COMMON: HUMAN; ORGAN: BLOOD; CELL: ERYTHROCYTES; MOL_ID: 2; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_COMMON: HUMAN; ORGAN: BLOOD; CELL: ERYTHROCYTES	MUTANT HUMAN HEMOGLOBIN S[BETA6V]; R2 QUATERNARY STATE; HUMAN HEMOGLOBIN	L.N.PATSKOVSKA,Y.V.PATSKOVSKY,S.C.ALMO,R.E.HIRSCH
PDB:2HBS	THE HIGH RESOLUTION CRYSTAL STRUCTURE OF DEOXYHEMOGLOBIN S	MOL_ID: 1; MOLECULE: HEMOGLOBIN S; CHAIN: A, B, C, D, E, F, G, H; SYNONYM: HBS, HEMOGLOBIN A (BETA6 GLU-VAL); MUTATION: CHAIN B, D, F, H; ENV VARIANT; BIOLOGICAL_UNIT: TETRAMER	MOL_ID: 1; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_COMMON: HUMAN; TISSUE: BLOOD; CELL: RED BLOOD CELLS; CELLULAR_LOCATION: CYTOPLASM	OXYGEN TRANSPORT, HEMOGLOBIN	D.J.HARRINGTON,K.ADACHI,W.E.ROYER JUNIOR
PDB:6HBW	CRYSTAL STRUCTURE OF DEOXY-HUMAN HEMOGLOBIN BETA6 GLU->TRP	MOL_ID: 1; MOLECULE: PROTEIN (HEMOGLOBIN ALPHA 1); CHAIN: A, C; SYNONYM: HEMOGLOBIN A; ENGINEERED: YES; MOL_ID: 2; MOLECULE: PROTEIN (HEMOGLOBIN BETA); CHAIN: B, D; SYNONYM: HEMOGLOBIN B; ENGINEERED: YES; MUTATION: YES	MOL_ID: 1; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_COMMON: HUMAN; TISSUE: BLOOD; CELL: RED BLOOD CELLS; CELLULAR_LOCATION: CYTOPLASM; EXPRESSION_SYSTEM: SACCHAROMYCES CEREVISIAE; EXPRESSION_SYSTEM_COMMON: YEAST; EXPRESSION_SYSTEM_STRAIN: GSY112; EXPRESSION_SYSTEM_PLASMID: PGS389; MOL_ID: 2; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_COMMON: HUMAN; TISSUE: BLOOD; CELL: RED BLOOD CELLS; CELLULAR_LOCATION: CYTOPLASM; EXPRESSION_SYSTEM: SACCHAROMYCES CEREVISIAE; EXPRESSION_SYSTEM_COMMON: YEAST; EXPRESSION_SYSTEM_STRAIN: GSY112; EXPRESSION_SYSTEM_PLASMID: PGS389	OXYGEN TRANSPORT, HEMOGLOBIN	D.J.HARRINGTON,K.ADACHI,W.E.ROYER JR.

SRS should find you 5 matching entries[16] this time. All are fine 3D structures, but **2HBS** is the highest resolution and the most interesting of the five, so click on its link and wait for it to load (3D structures are big, so this could take a few seconds).

Once it is loaded, click on the **Save** button. You will be linked to a new SRS page offering various save options. In this new page, ensure that the *Output To:* field is set to **File(text)** then ensure the *Save with*

[15] It stands for **P**rotein **D**ata **B**ank.

[16] Once more, there should be fuller information on your screen than is suggested by this document, and an extra hit.

view field is set to **PDBData** and then click on the new **Save** button. **rasmol** will now burst forth and translate the **PDB** entry you saw displayed by **Web Browser** as pages of co-ordinates[17] into a 3D graphical representation of the structure. Note that **rasmol** has both a graphical window and a textual window for issuing commands.

11. Analysis of the 3D structure of sickle cell haemoglobin

You should now be looking at the 3D structure of sickle cell haemoglobin (HbS). In fact you are looking at two interacting HbS complexes and since each HbS is a tetramer (with two alpha chains and two beta chains) you are therefore viewing eight separate protein chains. By using the mouse with the button depressed you can rotate the structure. You can zoom in on the structure by holding the shift button down and moving your mouse towards you. To zoom out, hold the shift button down and move the mouse away from you.

Perhaps one's first impression is "I now have a jumble of completely meaningless lines in front of me". Such an impression is understandable, there are close to 10,000 atoms in front of you! We can use molecular graphics however, to simplify and highlight particular features. We have generated a number of simplified representations of HbS that should facilitate your analysis. The aim is for you to ascertain *how* the Glu6 -> Val6 mutation might cause the HbS complexes to oligomerise into fibres, hence deforming erythrocytes. This will require you to examine the *structural context* of the mutation in the beta globin chains.

rasmol has both a text window and a graphics window. In the text window, type in:

```
script 1hbs
```

You should now be looking at a single HbS. With the exception of the Val6 side chains in the beta chains of HbS and the haem prosphetic groups, the other atoms have been largely removed with just the path of the polypeptide backbone showing. The protein chains are coloured differently. The alpha sub-units are purple and the beta sub-units are green. This simplified representation is somewhat easier on the eye and should allow you to examine the position of the Val6 side chains (the atoms of which are shown as solid spheres) in each of the beta globin chains.

8. *You should now be looking at a single HbS complex.*

   *What do you notice that is unusual about the location of the **Val6** side chains?*

Next, in the **rasmol** text window, type in:

```
script 2hbs
```

In this representation the interaction between two sickle cell haemoglobins is demonstrated. Highlighted are the side chains of the Val6 introduced by the mutation.

[17] Defining the relative positions of all the atoms of the protein molecule in 3D space.

9. *Now, you should be looking at a representation of the interaction between two sickle cell haemoglobins with the side chains of the **Val6** mutation highlighted.*

*What is suggested by the location of the **Val6** side chains?*

In the **rasmol** text window type in:

```
script 3hbs
```

Highlighted here are the amino acid side-chains in the vicinity of the Val6 at the interface of the two haemoglobins. By clicking on the atoms, the text window will tell you the identity of the amino acid.

10. *Comment upon the nature of the interface between the two haemoglobins you should now have in view?*

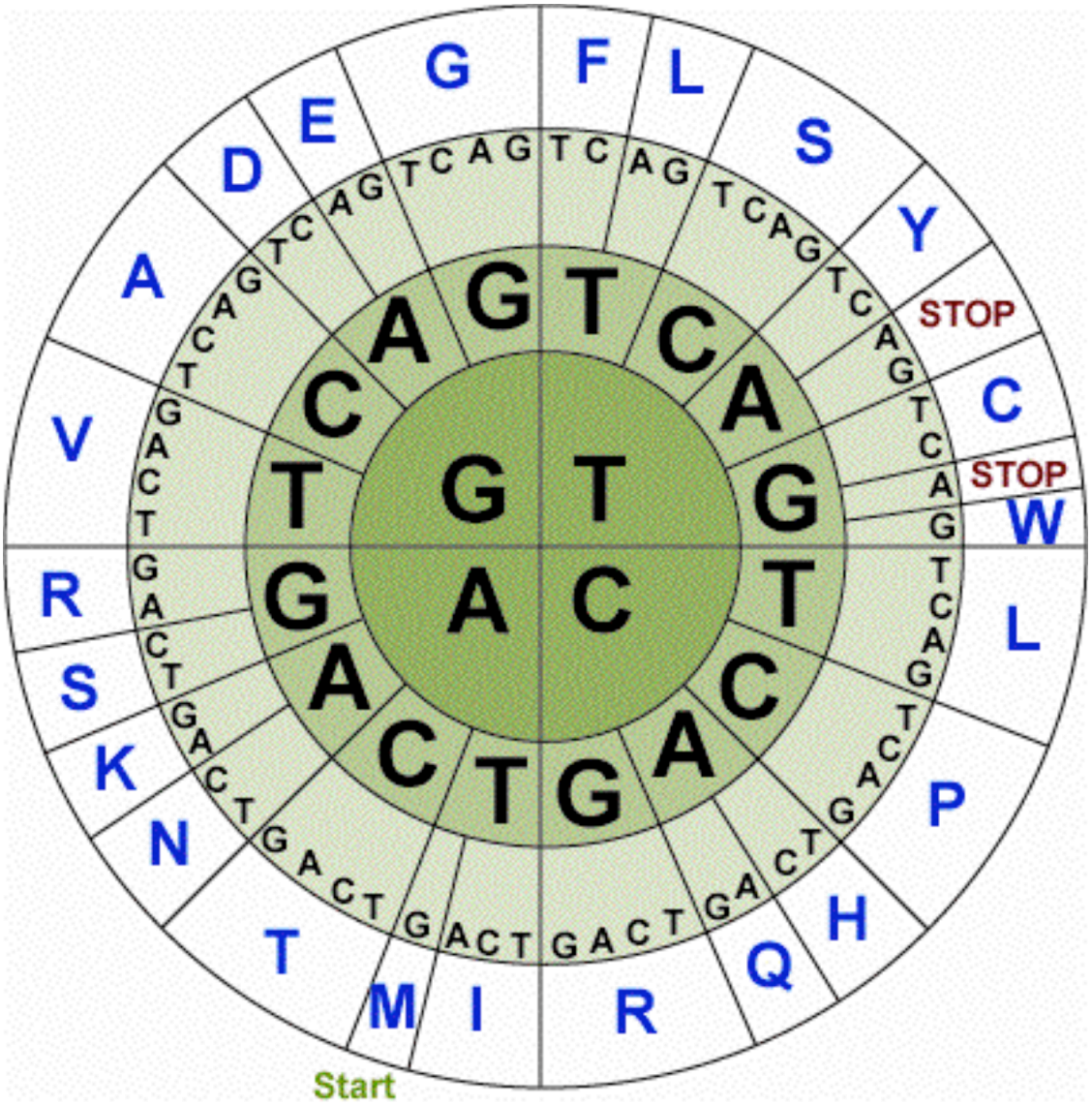
11. *From your observations, what conclusions can you draw about the molecular basis of sickle cell anaemia?*

You can compare your conclusions with those of the scientists who obtained this crystal structure by clicking once on the file called **paper.pdf** which is on your desktop (or, if you're feeling clever, find the paper on the web via the Cambridge University library, it is from the journal of molecular biology, 1997, volume 272 pages 398-407). In particular, look at Figure 3 in the paper which illustrates how the Glu6->Val6 mutation could cause HbS to form a double-helical fibre.

THE END

Appendix – The Standard Genetic Code

First Position	Second Position of Codon: T	C	A	G	Third Position
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G



DPJ & PDFJ 16/05/2006