A practical introduction to Bioinformatics.

Model answers.

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 on 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**?

   a) **EMBL** is a **Nucleotide sequence database** and is maintained by the **EBI** at Hinxton in England.

   b) **UniProtKB/Swiss-Prot** is a **Protein sequence database** and is maintain by the **EBI** (or **EMBL-EBI**)at Hinxton in England and **Amos Bairoch** at the University of Geneva (or **S**wiss **I**nstitute of **B**ioinformatics, **SIB**) in Switzerland.

2. Look at one of the haplotype entries by clicking on its **EMBL** link. Scroll down to the **F**ea**T**ure table (the section of lines beginning **FT**). You will see some **FT** lines 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 (to be found on one of the lines beginning **DR**) 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?

   a) The entry name for the **SWISSPROT** entry referenced by all haplotypes was **HBB_HUMAN**, accession code **P02023**.

   b) The Feature Section lines for the **variations** suggest the DNA sequence is different for each beta globin haplotype and yet the corresponding protein sequence is always the same. This implies that all the differences between the haplotypes are either within the introns (i.e. the non-coding regions) or are within the exons (i.e. the protein coding regions) but are silent differences (i.e. differences that change the DNA sequence but do not alter the sequence of the expressed).

   c) All the haplotypes have **3** exons.

   d) Exon 1 of the **EMBL** entry **HSBETGLOA** starts at position **866** and ends at position **957**.

   e) In your recent lectures it was suggested that introns usually begin with **GT** and end with **AG**. You should have seen that this rule is obeyed by both the introns of **HSBETGLOA**.

3. Looking at the multiple alignment of your DNA sequences you should see that sections of the mRNA sickle cell sequence has 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?

> The haplotype sequences are genomic and so will contain both protein coding regions (**exons**) and intermediate regions (**introns**). The sickle cell sequence is messenger RNA. Generating mRNA from genomic sequence involves splicing out all the introns. Therefore, the sickle cell sequence will only consist of the coding regions of the genomic sequence. **NclustalW** compensates for this by inserting minuses signs where the introns have been spliced out.
>
> The mRNA sickle cell sequence was split into three regions in order to align with the haplotype sequences. Each region corresponds to an exon in the genomic sequences. When looking at the annotation of the **EMBL** entries, you should already have noted that all the haplotype sequences all have three exons, so this should have been no surprise.

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 sequences?

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

> a) Other than those due to the existence of ambiguity codes, there is only one difference between the sickle sequence and the others within the first 30 or so base pairs. The **20th** base pair of the sickle cell sequence is a **T**. The corresponding base pair of all the other sequences is an **A**.
>
> ```
> -----ATGGTNCAYYTNACNCCNGTGGAGAAGTCYGCYGT
> ACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGT
> ACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT
> ACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT
> ACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT
> ACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT
> ACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT
>      ***** **  * ** ** * ********* ** **
> ```
>
> b) There are three base pairs to a codon, so the difference in the **20th** base pair must affect the **7th** codon.
>
> c) The **7th** codon in the haplotype sequences is **GAG** which codes for **Glutamic acid** (**Glu** or **E**).
>
> d) The **7th** codon in the sickle cell sequence is **GTG**, which codes for **Valine** (**Val** or **V**).

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?

a) The **7th** amino acid in the unedited map is a **Glu**. In the map of the edited sequence it is a **Val**.

b) There are **no** enzymes that cut the sickle cell sequence that do not also cut the beta globin haplotype sequence.

c) There are **8** enzymes that cut the haplotype sequence (**Hin4I** cuts twice) but do not cut the sequence with the sickle cell mutation edited in. They are:

BseMII	**Hpyl88III**
BseRI	**Hin4I**
BspCNI	**MnlI**
Bsu36I	**DdeI**

Map of Unedited Haplotype sequence:

```
== Linear Map of Sequence:
            Hin4I MnlI  DdeI
              BspCNI    Bsu36I
             BseMII    Hpy188III
              PleI HinfI BceAI          MaeIII
             MlyI Hpy188I                MwoI
             HpyCH4V  SfaNI             BseRI   Hin4I          Hpy8I
            \\\   \\  \\\\              \\  \    \              \
     1   atggtgcatctgactcctgaggagaagtctgccgttactgccctgtggggcaaggtgaac     60
         taccacgtagactgaggactcctcttcagacggcaatgacgggacaccccgttccacttg
            ^    *    ^    *    ^    *    ^    *    ^    *    ^    *
1        MetValHisLeuThrProGluGluLysSerAlaValThrAlaLeuTrpGlyLysValAsn
```

Map of Edited Haplotype sequence (i.e. with sickle cell mutation):

```
== Linear Map of Sequence:
              PleI HinfI BceAI
             MlyI Hpy188I              MaeIII
             HpyCH4V  SfaNI               MwoI              Hpy8I
             \\   \\  \  \              \   \                \
     1   atggtgcatctgactcctgtggagaagtctgccgttactgccctgtggggcaaggtgaac     60
         taccacgtagactgaggacacctcttcagacggcaatgacgggacaccccgttccacttg
            ^    *    ^    *    ^    *    ^    *    ^    *    ^    *
1        MetValHisLeuThrProValGluLysSerAlaValThrAlaLeuTrpGlyLysValAsn
```

d) One might test a beta globin sequence for the sickle cell mutation by cutting it with one of the restriction enzymes whose cut site is destroyed by the mutation. The missing cut site will result in a recognizably different set of restriction fragments to that expected for the normal sequence.

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?

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

   a) According to the **OMIM** entry, the most common cause of sickle cell anaemia is the mutation **Hb S**.

   b) The OMIM_allele entry for this mutation is **141900.0243**.

TEXT

A number sign (#) is used with this entry because sickle cell anemia is the result of mutant beta globin (HBB; 141900) in which the mutation causes sickling of hemoglobin rather than reduced amount of beta globin which causes beta-thalassemia. The most common cause of sickle cell anemia is Hb S (141900.0243), with SS disease being most prevalent in Africans.

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.

The **OMIM_allele** article mentions a number of restriction enzymes used in the diagnosis of sickle cell anaemia. The first mentioned is **HpaI**, but the use of this enzyme did not depend upon the sickle cell mutation eliminating one of its cut sites. It depended upon a separate base pair difference some 5Kb away from the beta globin gene. This second change eliminated an **HpaI** cut site and so gave rise to a **R**estriction **F**ragment **L**ength **P**olymorphism (**RFLP**). In certain African populations, this **RFLP** is significantly linked to the sickle cell mutation and so of use in diagnosis.

Other enzymes mentioned, whose usefulness in diagnosis depends upon the elimination of their cut site by the sickle cell mutation are:

MnlI ..... {GAGG}
MstII..... {CCTNAGG}
DdeI ..... {CTNAG}

The restriction enzyme MnlI recognizes the sequence G-A-G-G, which also is eliminated by the sickle mutation. The MstII enzyme recognizes the sequence C-C-T-N-A-G-G. Predictably, the resulting fragments are larger than those produced by some other enzymes, and MstII is, therefore, particularly useful in prenatal diagnosis (Wilson et al., 1982). The sickle cell mutation can be identified directly in DNA by use of either of 2 restriction endonucleases--DdeI or MstII (Geever et al., 1981; Kazazian, 1982). The nucleotide substitution alters a specific cleavage site recognized by each of these 2 enzymes. The fifth, sixth, and seventh codons of Hb A are CCT-GAG-GAG; in Hb S, they are CCT-GTG-GAG. The recognition site for DdeI is C-T-N-A-G, in which N = any nucleoside. Chang and Kan

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

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

The **Val6** side chains on the two beta globin chains are on the surface of the proteins. This is noteworthy as hydrophobic residues such as valine are more commonly found buried in the hydrophobic core of globular proteins.

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 do you notice about the location of the **Val6** side chains?

One of the **Val6** side chains forms part of an interface between the two **HbS** complexes. The other 3 point out away from the proteins (though we could expect these to form similar interfaces with other HbS complexes).

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

The interface consists largely, though not exclusively, of hydrophobic residues. These hydrophobic residues become buried from water when the two **HbS** complexes associate.

11. From your observations, what conclusions can you draw about the molecular basis of HbS oligomerisation?

The **Glu -> Val** mutation introduces a hydrophobic side chain on the surface of the protein. This forms a hydrophobic patch that can encourage **HbS** complexes to associate in order to bury from water the hydrophobic side chains. As a result, since each **HbS** has two beta chains, arrays of associated **HbS** complexes can form into fibres.

DPJ 16/05/2006