

Advanced DNA coding

Previously you have learnt the basics of DNA coding, and you have translated some DNA using the codon usage table. You are now going to move on and do some advanced DNA decoding and learn more about how the protein sequence code lies within the DNA.

Firstly, let's recap. You decoded a piece of DNA by grouping the nucleotides of the DNA together in groups of three, these groups are called codons. You then used the codon usage table to find the amino acid which corresponds to that codon.

ATGCGCGTAATT

ATG CGC GTA ATT

However, DNA is double stranded, and the above DNA sequence should be

ATGCGCGTAATT
TACGCGCATTAA

Remember that A pairs with a T and G pairs with a C.

The DNA at the top is called the Top Strand (red), the DNA at the bottom is called the Bottom Strand (blue). See picture above.

The DNA strand also has an end; there are four ends on a double stranded piece of DNA. The ends are called 5' and 3', so there are two 5' ends and two 3' ends (see picture below). Genes code for proteins, genes always run in the 5' to 3' direction, this can be on the top strand, or bottom strand. See the direction of the arrows on the diagram below.

→
5'-ATGCGCGTAATT-3'
3'-TACGCGCATTAA-5'
←

When you did your first DNA decoding, you began grouping into three with the 'A' on the top DNA strand. This is called the **+1 reading frame**.

ATGCGCGTAATT
↓
ATG CGC GTA ATT

If you begin on the second nucleotide (the T), and then group into three, and decode from here, it is called the **+2** reading frame.

A TGC GCG TAA TT
↓
A TGC GCG TAA TT

If you begin on the third nucleotide (the G), and then group into three, and decode from here, it is called the **+3** reading frame.

AT GCG CGT AAT T
↓
AT GCG CGT AAT T

This has now completely decoded the top strand of DNA, but there is still the bottom strand to deal with.

3'-TACGCGCATTAA-5'

If you begin at the A on the far right (this is the 5' end), this is called the -1 reading frame.

3'-TAC GCG CAT TAA-5'
↓

If you then start to decode from the next A along, this is called the -2 reading frame.

3'-TA CGC GCA TTA A-5'
↓

If you then start to decode from the third nucleotide along, the T, this is called the -3 reading frame.

3'-T ACG CGC ATT AA-5'
↓

Exercise 1

Have a go at translating this short DNA sequence into all six reading frames.

5'-ATGGGATTCAGGCCACAT-3'
3'-TACCCTAACTCCGGTGTA-5'

What have you learnt?

- DNA has a code.
- Nucleotides in the DNA sequence pair together
- DNA is double stranded

- Nucleotides are grouped into three's called codons.
- The codon usage table can be used to find the amino acid sequence of the protein coded for by the DNA
- Genes are coded for in the 5' to 3' direction
- DNA has 6 reading frames (+1,+2,+3, -1, -2,-3)

It probably won't surprise you to learn that scientists don't sit and decode the entire DNA sequence by themselves, they use a computer to do it for them. This along with other DNA and protein analysis is a new area of science called Bioinformatics. Bioinformatics can be really exciting, especially now all the human DNA has been sequenced, we are finding out new things every day.

You are now going to do some Bioinformatics for yourself in the next exercise. You are going to decode a section of DNA using software available on the web.

Exercise 2

- 1) In order to do this you are going to need a DNA sequence. Copy the sequence given at the end of this document. Highlight the sequence then press 'Ctrl' and 'C'.
- 2) Now you access the site on the web which will translate the DNA sequence for you.
- 3) Enter the following URL
<http://bioweb.pasteur.fr/seqanal/interfaces/sixpack.html>
- 4) In order to get this programme to run, you must provide it with an e-mail address, type it into the space provided.
- 5) Then paste the sequence into the 'Input' section. Do this by hold down 'Ctrl' and pressing 'V'
- 6) You don't need to alter anything else, simply press the 'Run SixPack' button and wait for your results!
- 7) To see the results click onto the outfile.out link
- 8) All six reading frames are displayed. Three above (+1,+2 and +3), and three below (-1, -2 and -3).

What have you learnt?

- That DNA can be decoded using programmes on the web
- How to use some of these web based analysis sites.

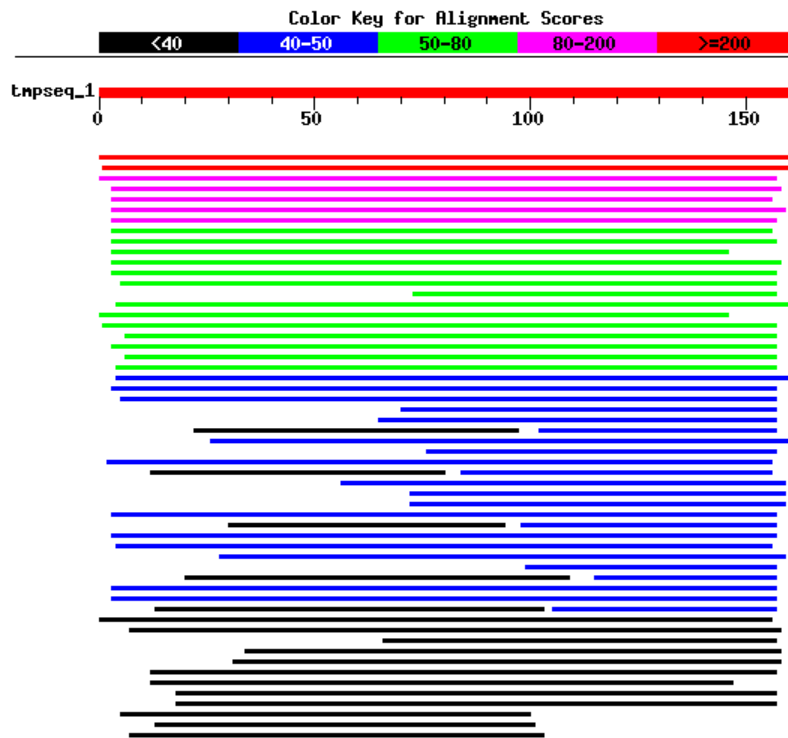
Exercise 3

You may be curious to know what this DNA sequence actually codes for, well you are now going to find out.

- 1) Copy the DNA sequence as before.
- 2) Type in the following URL, <http://www.ncbi.nlm.nih.gov/BLAST/>
- 3) This is a database containing all of the DNA sequences which are currently known. It is a wonderful resource and freely available to all.
- 4) Go to the link which says [Nucleotide-nucleotide BLAST \(blastn\)](#)
- 5) Paste the sequence (Ctrl 'V') into the box called 'Search' Don't change anything else.
- 6) Click onto the button called 'BLAST'

- A new window will open, then click on the button called 'Format' to look at your results.

Your results will be displayed to you in three different ways, the first is a graphical output, this simply shows you if you have any 'strong' hits, ie if your sequence matches closely to something in the database. These are colour coded, red being a very good match, followed by pink, then green, blue and finally black, which represents poor matches.



The next way your results are represented gives the actual hit from the database, it will tell you what the sequence codes for. The information on the left hand side gives the Accession Number (database entry) of the match. The information on the far right tells you how significant the match is, the lower the value the more significant the match. In this case 0 is a perfect match. If you click onto the Accession number, it will take you to the sequence of the database entry.

Sequences producing significant alignments:	(Bits)	Value	
gi 38016908 ref NM_000545.3 Homo sapiens transcription facto...	1929	0.0	U
gi 416528 emb X71346.1 HSHNF1B Homo sapiens HNF1-B mRNA	1929	0.0	U
gi 184264 gb M57732.1 HUMHNF1 Human hepatic nuclear factor 1 ...	1929	0.0	U
gi 416527 emb X71347.1 HSHMF1C H.sapiens HNF1-C mRNA	1913	0.0	U
gi 85397793 gb BC104910.1 Homo sapiens transcription factor ...	1907	0.0	G
gi 85396995 gb BC104908.1 Homo sapiens transcription factor ...	1899	0.0	G

The final way your results are presented are as a pairwise alignment. Your sequence is the 'Query' sequence, and the hit from the database is the 'Subject' sequence. The lines in the middle show those nucleotides which are identical between the two sequences.

> [gi|38016908|ref|NM_000545.3|](#) **UEG** Homo sapiens transcription factor 1, hepatic; LF-B1, hepatic nuclear factor (HNF1), albumin proximal factor (TCF1), mRNA
Length=3249

Score = 1929 bits (973), Expect = 0.0
Identities = 999/1000 (99%), Gaps = 0/1000 (0%)
Strand=Plus/Plus

```
Query 1 CGTGGCCCTGIGGCAGCCGAGCCATGGTTTCTAAACTGAGCCAGCTGCAGACGGAGCTCC 60
      |||
Sbjct 1 CGTGGCCCTGIGGCAGCCGAGCCATGGTTTCTAAACTGAGCCAGCTGCAGACGGAGCTCC 60

Query 61 TGGCGGCCCTGCTCGAGTCAGGGCTGAGCAAAGAGGCACTGATCCAGGCAC TGGGTGAGC 120
      |||
Sbjct 61 TGGCGGCCCTGCTCGAGTCAGGGCTGAGCAAAGAGGCACTGATCCAGGCAC TGGGTGAGC 120

Query 121 CGGGGCCCTACCTCCTGGCTGGAGAAGGCCCTGACAAAGGGGGAGTCTGCGGCGGG 180
      |||
Sbjct 121 CGGGGCCCTACCTCCTGGCTGGAGAAGGCCCTGACAAAGGGGGAGTCTGCGGCGGG 180

Query 181 GTCGAGGGGAGCTGGCTGAGCTGCCAATGGGCTGGGGGAGACTCGGGGCTCCGAGGACG 240
      |||
Sbjct 181 GTCGAGGGGAGCTGGCTGAGCTGCCAATGGGCTGGGGGAGACTCGGGGCTCCGAGGACG 240

Query 241 AGACGGACGACGATGGGGAAGACTTCACGCCACCCATCCTCAAAGAGCTGGAGAACCTCA 300
      |||
Sbjct 241 AGACGGACGACGATGGGGAAGACTTCACGCCACCCATCCTCAAAGAGCTGGAGAACCTCA 300

Query 301 GCCCTGAGGAGGCGGCCACCAGAAAGCCGTGGTGGAGACCCTTCTGCAGGAGGACCCGT 360
      |||
Sbjct 301 GCCCTGAGGAGGCGGCCACCAGAAAGCCGTGGTGGAGACCCTTCTGCAGGAGGACCCGT 360

Query 361 GCGCTGTGGCGAAGATGGTCAAGTCTACCTGCAGCAGCACAACATCCACAGCGGGAGG 420
      |||
Sbjct 361 GCGCTGTGGCGAAGATGGTCAAGTCTACCTGCAGCAGCACAACATCCACAGCGGGAGG 420

Query 421 TGGTCGATACCACTGGCTCAACCAGTCCCACCTGTCCCAACACCTCAACAAGGGCACTC 480
      |||
Sbjct 421 TGGTCGATACCACTGGCTCAACCAGTCCCACCTGTCCCAACACCTCAACAAGGGCACTC 480

Query 481 CCATGAAGACGCGAGAAGCGGGCCGCCCTGTACACCTGGTACGTCGCAAGCAGCGAGAGG 540
      |||
Sbjct 481 CCATGAAGACGCGAGAAGCGGGCCGCCCTGTACACCTGGTACGTCGCAAGCAGCGAGAGG 540

Query 541 TGGCGCAGCAGTTCACCCATGCAGGGCAGGGAGGGCTGATTGAAGAGCCCACAGGTGATG 600
      |||
Sbjct 541 TGGCGCAGCAGTTCACCCATGCAGGGCAGGGAGGGCTGATTGAAGAGCCCACAGGTGATG 600

Query 601 AGCTACCAACCAAGAAGGGGCGGAGGAACCGTTTCAAGTGGGGCCAGCATCCCAGCAGA 660
      |||
Sbjct 601 AGCTACCAACCAAGAAGGGGCGGAGGAACCGTTTCAAGTGGGGCCAGCATCCCAGCAGA 660

Query 661 TCCTGTTCCAGGCCTATGAGAGGCAGAAGAACCCTAGCAAGGAGGAGCGAGAGACGCTAG 720
      |||
Sbjct 661 TCCTGTTCCAGGCCTATGAGAGGCAGAAGAACCCTAGCAAGGAGGAGCGAGAGACGCTAG 720

Query 721 TGGAGGAGTGCAATAGGGCGGAATGCATCCAGAGAGGGGTGTCCCATCACAGGCACAGG 780
      |||
Sbjct 721 TGGAGGAGTGCAATAGGGCGGAATGCATCCAGAGAGGGGTGTCCCATCACAGGCACAGG 780

Query 781 GGCTGGGCTCCAACCTCGTCACGGAGGTGCGTGTCTACAACCTGGTTTGCCAACCGGCGCA 840
      |||
Sbjct 781 GGCTGGGCTCCAACCTCGTCACGGAGGTGCGTGTCTACAACCTGGTTTGCCAACCGGCGCA 840

Query 841 AAGAAGAAGCCTTCCGGCACAAGCTGGCCATGGACACGTACAGCGGGGnnnnnnnnAGGGC 900
      |||
Sbjct 841 AAGAAGAAGCCTTCCGGCACAAGCTGGCCATGGACACGTACAGCGGGCCCCCCCCAGGGC 900

Query 901 CAGGCCCGGGACCTGCGCTGCCCGCTCACAGCTCCCCTGGCCTGCCTCCACCTGCCCTCT 960
      |||
Sbjct 901 CAGGCCCGGGACCTGCGCTGCCCGCTCACAGCTCCCCTGGCCTGCCTCCACCTGCCCTCT 960
```

```
Query 961 CCCCAGTAAGGTCCACGGTGTGCGCTNTGGACAGCCTGC 1000
        |||
Sbjct 961 CCCCAGTAAGGTCCACGGTGTGCGCTATGGACAGCCTGC 1000
```

So, your piece of DNA was a protein found in human blood

What have you learnt?

- You can use freely available databases to find out what DNA codes for
- How to interpret results from a DNA database