The FASTA Format [1]

The FASTA format is a very simply format to describe nucleic acid or amino acid sequences. A FASTA file begins with a single-line containing a description of the sequence, followed by lines of sequence data.
The description line is identified by a greater-than (">") symbol in the first column. It is recommended that all lines of text should be shorter than 80 characters in length. For example, from the description line it is clear that the FASTA file AE005174v2-1.fas file contains the first contig of the E.coli sequence:

```
>AE005174-1 Genome sequence of enterohemorrhagic Escherichia coli O157:H7, segment 1 of 2.
agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtctctgacagcagcttctgaactg
gttacctgccgtgagtaaattaaaattttattgacttaggtcactaaatactttaaccaatataggcatagcgcacagac
agataaaaattacagagtacacaacatccatgaaacgcattagcaccaccattaccaccaccatcaccaccaccatcacc
attaccattaccacaggtaacggtgcgggctgacgcgtacaggaaacacagaaaaaagcccgcacctgacagtgcgggct
```

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable query letters (see below).
Before submitting a request/query, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue). (In the given E.coli FASTA sequence there are no numbers.)

The nucleic acid codes supported are:

```
A --> adenosine           M --> A C (amino)
C --> cytidine            S --> G C (strong)
G --> guanine             W --> A T (weak)
T --> thymidine           B --> G T C
U --> uridine             D --> G A T
R --> G A (purine)        H --> A C T
Y --> T C (pyrimidine)    V --> G C A
K --> G T (keto)          N --> A G C T (any)
                          -  gap of indeterminate length
```

In our two FASTA files all these codes appear except for the code '-'.

Note that for amino acid query sequences (BLASTP and TBLASTN) other amino acid codes are used (see [1]).

The Gene Positions

The gene positions given in the zip-file (and MS Excel file) are given by the starting position and ending position (lend, en rend) of the genes. These indicate the nucleotide positions in the FASTA files (of course without counting new line characters etc.). Furthermore, counting starts at position 1.
Therefore, to derive the location and sequence of the genes, you have to write some simple code [2, 3] to read all the codes given in the FASTA file into a character array and do the appropriate counting. This also gives an easy way to print statistics, tables, and special formatted files that can be handled by the Hidden Markov Toolkits. (See also the given GHMM DNA example.)

References

[1] http://www.ncbi.nlm.nih.gov/blast/fasta.shtml
[2] http://bioweb.pasteur.fr/docs/seqio/seqio.html
[3] Sample code for FASTA-I/O and statistics: http://www.liacs.nl/~erwin/dbdm2008/fasta_io.zip