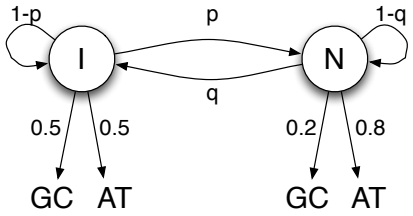
Series 4 - solutions

Genomics and bioinformatics - Week 5 - October 16, 2012

1 Hidden Markov Model

1. There are two states, say I (isochore) and N (normal). We observe sequences of bases A, T, G and C. For this exercise one can group G and C in one variable "GC", and similarly A and T in "AT". Note that from each state, the outgoing probabilities must sum to 1.



2. In state I (isochore), the probability to see GC is 0.5, the same for AT. From state N, the probabilities are 0.2 for GC and 0.8 for AT.

3. The isochore is 7000 bases long, the genome 23'000'000, so the probability for a random base in the genome to belong to the isochore is $x = P(I) = 7'000/23'000'000 = 3 \cdot 10^{-4}$.

4. We have

$$P(N|I) = p;\ P(I|N) = q;\ P(I|I) = (1-p);\ P(N|N) = (1-q)$$

$$P(I) = P(I|N)P(N) + P(I|I)P(I)$$
$$P(N) = P(N|I)P(I) + P(N|N)P(N)$$

so in terms of x, p and q:

$$x = q(1-x) + (1-p)x$$
$$1-x = px + (1-q)(1-x)$$

Note that the two equations are equivalent, so one cannot solve directly for p and q.

5. From state I, one can consider the event "staying in I" as a fail, with probability $1-p$, and "going to N" as a success, with probability p. The number X of failures before the first success is given by a geometric distribution:

$$P(X = k) = (1-p)^k p.$$

Its mean is $E[X] = \frac{1-p}{p}$ (another formulation, taking X as the time of the first success, leads to $E[X] = \frac{1}{p}$). [1]

[1] `http://en.wikipedia.org/wiki/Geometric_distribution`

6. If the isochore is generated from a geometric process as in the previous point, its length is most probably the mean of the distribution. So $7000 = E[X] = \frac{1-p}{p} \Rightarrow p = \frac{1}{7001}$ (or $\frac{1}{7000}$ with the alternative formulation), confirming what one could expect intuitively. Taking $p = \frac{1}{7000}$, one deduces from point 4 that $q = x/(1-x)p = \frac{1}{11496500}$. One may also compute q as follows: Exchanging the role of I and N in point 5, writing Y for the corresponding random variable and L for the average length of the two non GC-rich regions, one obtains $L = \frac{23000000-7000}{2}$, $E(Y) = L_{\text{total}} = 2L$ and $E(Y) = \frac{1-q}{q}$, so that $q = \frac{1}{11496500}$ as before.

2 Reading frame

See the program "series5_solution.py".

3 BLAST

(Results here may change with the evolution of sequencing databases).

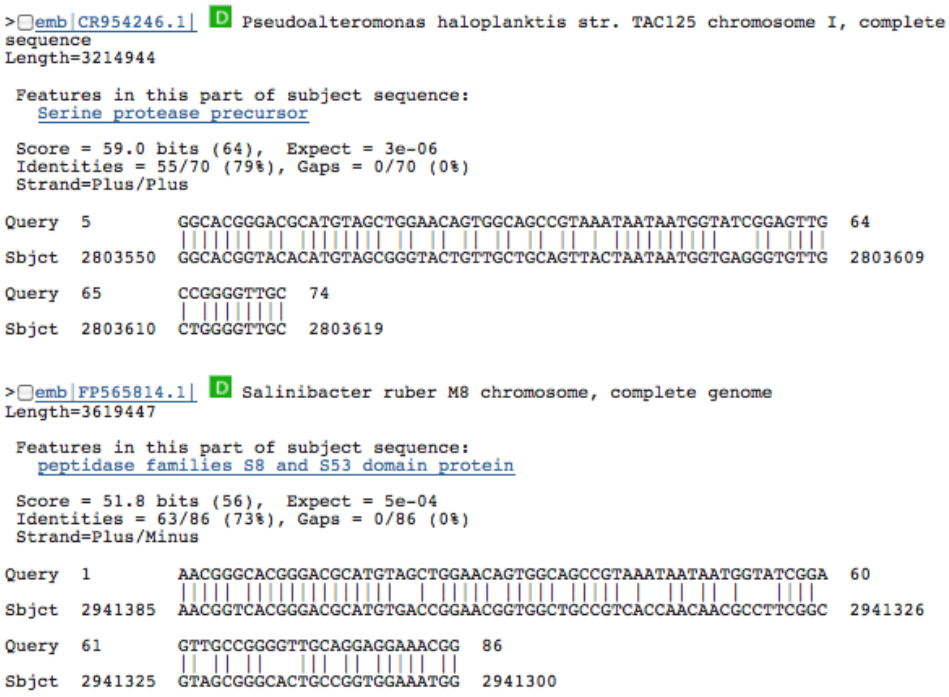
3.1 Nucleotide BLAST

1. Do you get any matches to `fragment_007`? Which parameters did you use? Record the alignment statistics for the top hits.

   - Using the Nucleotide Collection (nr/nt) Database and optimizing for “more dissimilar sequences” (discontiguous megablast)

```
>emb|CR954246.1|  Pseudoalteromonas haloplanktis str. TAC125 chromosome I, complete
sequence
Length=3214944

 Features in this part of subject sequence:
   Serine protease precursor

 Score = 59.0 bits (64),  Expect = 4e-05
 Identities = 55/70 (79%), Gaps = 0/70 (0%)
 Strand=Plus/Plus

Query  5        GGCACGGGACGCATGTAGCTGGAACAGTGGCAGCCGTAAATAATAATGGTATCGGAGTTG  64
                |||||||  || |||||||| ||  ||  ||  ||  ||  | ||||||||||    || ||||
Sbjct  2803550  GGCACGGTACACATGTAGCGGGTACTGTTGCTGCAGTTACTAATAATGGTGAGGGTGTTG  2803609

Query  65       CCGGGGTTGC  74
                 | ||||||||
Sbjct  2803610  CTGGGGTTGC  2803619


>emb|FP565814.1|  Salinibacter ruber M8 chromosome, complete genome
Length=3619447

 Features in this part of subject sequence:
   peptidase families S8 and S53 domain protein

 Score = 51.8 bits (56),  Expect = 0.005
 Identities = 63/86 (73%), Gaps = 0/86 (0%)
 Strand=Plus/Minus

Query  1        AACGGGCACGGGACGCATGTAGCTGGAACAGTGGCAGCCGTAAATAATAATGGTATCGGA  60
                |||||  |||||||||||||||   | |||||  |||||  |||||  |   ||  ||  |    ||||
Sbjct  2941385  AACGGTCACGGGACGCATGTGACCGGAACGGTGGCTGCCGTCACCAACAACGCCTTCGGC  2941326

Query  61       GTTGCCGGGGTTGCAGGAGGAAACGG  86
                 ||  ||  ||   |||  ||  ||||  ||
Sbjct  2941325  GTAGCGGGCACTGCCGGTGGAAATGG  2941300
```

2. Extract the sequence of the hit with the highest query coverage (this may not necessarily be the top hit) and perform another nucleotide BLAST, using the same parameters. Record the alignment statistics for the top hits.



3. What changes do you observe in the E-values? To which parameter could you attribute the these changes?

   - Improvement in the E-values. Parameter - Query coverage.

4. What is the default threshold for the E-value on NCBI BLAST?

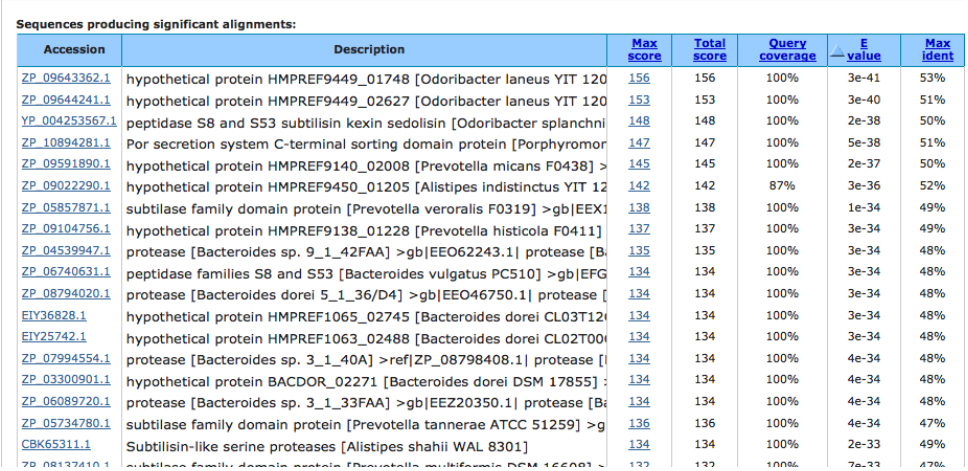
   - 10

5. Do you have any significant hits suggesting a possible function for `fragment_007`?

   - The E-values are not significant.

3.2 Protein BLAST

1. Using your custom function from exercise 2, one can extract the following nucleotides sequence from the translation of the forward strand with shift 0 (must start with 'M'; incomplete):
   `MSTQIFNSDGDYTNSETLVYRAIVYGADNGAVISQNSWGSQSLTIKELQKAAIDYFIDYAGMDETGEIQT GPMRGGIFIAAAGNDNVSTPNMPSAYERVLAVASMGPDFTKASYSTFGTWTDITAPGGDIDKFDLSEYGV LSTYADNYYAYGEGTSMACPHVAGAA`.
   Copy it into a file `aa_007.fasta`, or directly into the BLASTp interface, and run the alignment. After a few seconds, you get the following matches of the peptidases S8 S53 superfamily:



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
ZP_09643362.1	hypothetical protein HMPREF9449_01748 [Odoribacter laneus YIT 120	156	156	100%	3e-41	53%
ZP_09644241.1	hypothetical protein HMPREF9449_02627 [Odoribacter laneus YIT 120	153	153	100%	3e-40	51%
YP_004253567.1	peptidase S8 and S53 subtilisin kexin sedolisin [Odoribacter splanchni	148	148	100%	2e-38	50%
ZP_10894281.1	Por secretion system C-terminal sorting domain protein [Porphyromor	147	147	100%	5e-38	51%
ZP_09591890.1	hypothetical protein HMPREF9140_02008 [Prevotella micans F0438] >	145	145	100%	2e-37	50%
ZP_09022290.1	hypothetical protein HMPREF9450_01205 [Alistipes indistinctus YIT 12	142	142	87%	3e-36	52%
ZP_05857871.1	subtilase family domain protein [Prevotella veroralis F0319] >gb|EEX1	138	138	100%	1e-34	49%
ZP_09104756.1	hypothetical protein HMPREF9138_01228 [Prevotella histicola F0411]	137	137	100%	3e-34	49%
ZP_04539947.1	protease [Bacteroides sp. 9_1_42FAA] >gb|EEO62243.1| protease [B	135	135	100%	3e-34	48%
ZP_06740631.1	peptidase families S8 and S53 [Bacteroides vulgatus PC510] >gb|EFG	134	134	100%	3e-34	48%
ZP_08794020.1	protease [Bacteroides dorei 5_1_36/D4] >gb|EEO46750.1| protease [	134	134	100%	3e-34	48%
EIY36828.1	hypothetical protein HMPREF1065_02745 [Bacteroides dorei CL03T12	134	134	100%	3e-34	48%
EIY25742.1	hypothetical protein HMPREF1063_02488 [Bacteroides dorei CL02T00	134	134	100%	3e-34	48%
ZP_07994554.1	protease [Bacteroides sp. 3_1_40A] >ref|ZP_08798408.1| protease [	134	134	100%	4e-34	48%
ZP_03300901.1	hypothetical protein BACDOR_02271 [Bacteroides dorei DSM 17855] >	134	134	100%	4e-34	48%
ZP_06089720.1	protease [Bacteroides sp. 3_1_33FAA] >gb|EEZ20350.1| protease [Ba	134	134	100%	4e-34	48%
ZP_05734780.1	subtilase family domain protein [Prevotella tannerae ATCC 51259] >g	136	136	100%	4e-34	47%
CBK65311.1	Subtilisin-like serine proteases [Alistipes shahii WAL 8301]	134	134	100%	2e-33	49%
ZP_08137410.1	subtilase family domain protein [Prevotella multiformis DSM 16608] >	132	132	100%	7e-33	47%

2. `fragment_007` encodes for a subtilase family domain protein. It is a member of the peptidases S8 (subtilisin and kexin) and S53 (sedolisin) family. These include endopeptidases and exopeptidases.

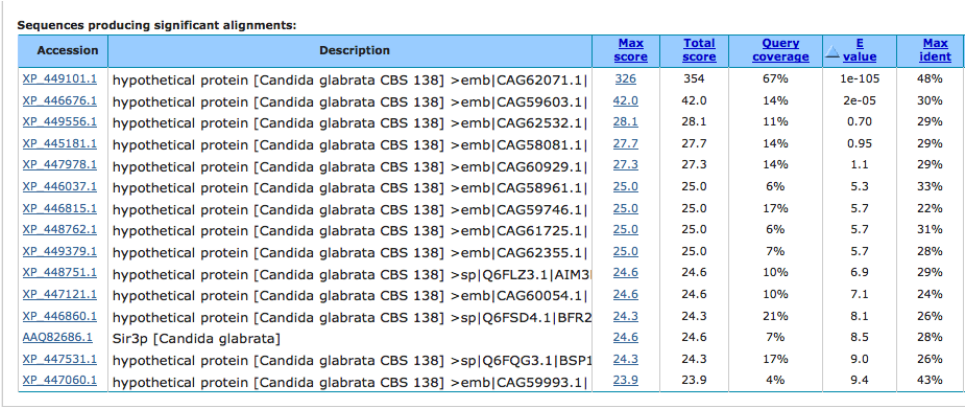
3. *Odoribacter, Prevotella, Porphyromonas* and *Alistipes* species are predominant. Note that *Alistipes* is the one you found with the nucleotide BLAST, and it is not the top match.

4. BLASTx

5. Amino acid sequences are more conserved than nucleotide sequences. Often even the highest-scoring subject sequences retrieved using the nucleotide sequence will cover only small regions of the query sequence, while quite often the corresponding sequences retrieved using the amino acid sequence will cover more of the gene.

3.3 Finding orthologs

Specify in the *Organism* section of the BLASTp interface that you want to align on species *Candida glabrata*. Consistently with the publication, the best match indicates `GENE ID: 2890989 CAGL0L07436g`:



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
XP_449101.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG62071.1|	326	354	67%	1e-105	48%
XP_446676.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG59603.1|	42.0	42.0	14%	2e-05	30%
XP_449556.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG62532.1|	28.1	28.1	11%	0.70	29%
XP_445181.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG58081.1|	27.7	27.7	14%	0.95	29%
XP_447978.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG60929.1|	27.3	27.3	14%	1.1	29%
XP_446037.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG58961.1|	25.0	25.0	6%	5.3	33%
XP_446815.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG59746.1|	25.0	25.0	17%	5.7	22%
XP_448762.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG61725.1|	25.0	25.0	6%	5.7	31%
XP_449379.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG62355.1|	25.0	25.0	7%	5.7	28%
XP_448751.1	hypothetical protein [Candida glabrata CBS 138] >sp|Q6FLZ3.1|AIM3	24.6	24.6	10%	6.9	29%
XP_447121.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG60054.1|	24.6	24.6	10%	7.1	24%
XP_446860.1	hypothetical protein [Candida glabrata CBS 138] >sp|Q6FSD4.1|BFR2	24.3	24.3	21%	8.1	26%
AAQ82686.1	Sir3p [Candida glabrata]	24.6	24.6	7%	8.5	28%
XP_447531.1	hypothetical protein [Candida glabrata CBS 138] >sp|Q6FQG3.1|BSP1	24.3	24.3	17%	9.0	26%
XP_447060.1	hypothetical protein [Candida glabrata CBS 138] >emb|CAG59993.1|	23.9	23.9	4%	9.4	43%