Introduction to the game

Recently an interesting protein with the amino acid sequence SARAND was found in the bacteria *S. Equencia*. It is now to be determined if a homologue exists in the species *B. Ionformatica*.

To determine this a lab amplificied a relevant part of the DNA of *B. Ionformatica* using PCR primers flanking the gene in *S. Equencia* which are believed to be highly conserved also in *B. Ionformatica*, allthough the sequence of *B. Ionformatica* is currently not known. The amplified DNA was sequenced using Ullamini LoSeq next generation sequencing technology yielding 11 reads. The quality of the reads are not perfect – read errors resulting in random read "mutations" are expected in one out of twenty bases.

As a bioinformatician you are given the task to find out if *B. Ionformatica* has a homologue of the protein SARAND and determine how its amino acid sequence differs in *B. Ionformatica*. However, the high performance moon grid engine supercluster is currently down (as it sometimes is) and you have to do it all by hand. Fortunely, you have printed all the reads. You task is as follows:

- Perform de-novo assembly of all the reads
- Find open reading frames that may contain a gene
- Find the amino acid sequence of any such gene to determine if it could be a homologue to SARAND
- Report your finding and claim eternal fame

But you have to hurry! Many other competing research groups have also gotten a hold of the reads and they will scoop you on this important discovery if you are not fast.

Detailed instructions

Cut the reads with a scissor so you have them as paper strips. Place these paper strips horizontally on the board in the read alignment area so that they overlap each other, matching base by base. Above the read alignment area is a row which have six predefined nucleotides in each side (the PCR primers). Your reads should also match these. Once you have aligned all the reads, you can fill the empty cells with the consensus sequence obtained from the alignment. It is possible to map all the reads, but not all reads may not map to the same strand. If you end up with reads that do not fit anywhere, then most likely your alignment is wrong.

Once you have the entire consensus sequence you look for the hidden protein. In the three top of the board you can write down amino acids (and start and stop codons) for the forward strand corresponding to your consensus sequence. In the bottom you do the same for the reverse strand.

An open reading frame consist of a start codon, some intermediate codons and a stop codon. Identify open reading frames and translate intermediate codons using the supplied amino acid table. Find the protein that looks the most homologue.

Winning the game: The team that first correctly reports the amino acid sequence[1] of the protein homologue and also its nucletide sequence wins the game.

Feedback during the game: During the game a team can hand in the IUPAC consensus codes of their alignment (see attached table). The judge of the game will then determine if the sequence is correct. Similarly you may make hand it a protein sequence and the judge will check its correctness (if correct, you win). However, this service comes at a price. The first time you ask it will cost you one of your reads. The second time it cost you two reads, the third time four reads and so on. So, use your reads wisely.

[1] Report amino acid sequence without the methionine encoded by the start codon

Solution:

														A	T	C	T	C	T	A	C	T	C	G	T	G	C	T	A	A	C	G	A	C	T	A	G		
A	C	C	C	C	G	C	G	C	T	A	C	G	C																										
		C	C	C	G	C	G	C	T	A	C	G	C	A	T																								
			C	C	G	C	G	C	T	A	C	G	C	A	T	C																							
										A	C	G	C	A	T	C	T	C	T	G	C	T	C																
															T	C	T	C	T	A	C	T	C	G	T	G	C	T											
																		C	T	A	C	T	C	G	T	G	C	T	A	A	C								
																				A	C	T	C	G	T	G	C	T	A	A	C	G	A						
																					C	T	C	A	T	G	C	T	A	A	C	G	A	C					
																										G	C	T	A	A	C	G	A	C	T	A	G	G	C
																										G	A	T	A	A	C	G	A	C	T	A	G	G	C
IUPAC Consensus sequence:																																							
A	C	C	C	C	G	C	G	C	T	A	C	G	C	A	T	C	T	C	T	R	C	T	C	R	T	G	M	T	A	A	C	G	A	C	T	A	G	G	**C**

Scaffold

<table>
<tr><td colspan="40">Amino acid sequence in each three reading frames (forward strand)</td></tr>
<tr><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td></td></tr>
<tr><td></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td colspan="2"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="2"></td></tr>
<tr><td colspan="40">Nucleotide sequence (forward strand):</td></tr>
<tr><td><b>A</b></td><td><b>C</b></td><td><b>C</b></td><td><b>C</b></td><td><b>C</b></td><td><b>G</b></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td><b>C</b></td><td><b>T</b></td><td><b>A</b></td><td><b>G</b></td><td><b>G</b></td><td><b>C</b></td></tr>
</table>

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

<table>
<tr><td colspan="40">Nucleotide sequence (forward strand):</td></tr>
<tr><td><b>T</b></td><td><b>G</b></td><td><b>G</b></td><td><b>G</b></td><td><b>G</b></td><td><b>C</b></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td><b>G</b></td><td><b>A</b></td><td><b>T</b></td><td><b>C</b></td><td><b>C</b></td><td><b>G</b></td></tr>
<tr><td colspan="40">Amino acid sequence in each three reading frames (forward strand)</td></tr>
<tr><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td></td></tr>
<tr><td></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td colspan="2"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td><td colspan="2"></td></tr>
</table>

Reads - cut out

C	T	A	T	T	G	C	T	G	A	T	C	C	G
T	G	G	G	G	C	G	C	G	A	T	G	C	G
G	A	T	G	A	G	C	A	C	G	A	T	T	G
G	A	G	T	A	C	G	A	T	T	G	C	T	G
T	G	A	G	C	A	C	G	A	T	T	G	C	T
T	G	C	G	T	A	G	A	G	A	C	G	A	G
T	G	G	G	G	C	G	C	G	A	T	G	C	G
C	G	A	T	T	G	C	T	G	A	T	C	C	G
G	G	G	C	G	C	G	A	T	G	C	G	T	A
A	G	A	G	A	T	G	A	G	C	A	C	G	A
G	G	C	G	C	G	A	T	G	C	G	T	A	G

Amino Acid translation table

Second base in codon

First base in codon	T	C	A	G	Third base in codon
T	TTT Phe F	TCT Ser S	TAT Tyr Y	TGT Cys Y	T
	TTC Phe F	TCC Ser S	TAC Tyr Y	TGC Cys Y	C
	TTA Leu L	TCA Ser S	TAA Stop *	TGA Stop *	A
	TTG Leu L	TCG Ser S	TAG Stop *	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

Standard genetic code. The table shows how triplets of nucleic acid bases correspond to different amino acids. Besides the codon ATG with always codes for methionine, alternatively TTG, CTG, ATT, ATC, ATA and GTG can serve as initiation codons, in which case they are translated as methionine rather than the amino acid indicated. However – in this game – we do not consider the first methionine as part of the solution amino acid sequence.

Nucleotide Code:	*Base:*
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base

IUPAC consensus codes.