Tutorial for "Stop codon reassignment in the wild"

Learning Objectives

This tutorial has two learning objectives:

1. Finding evidence of stop codon reassignment on DNA fragments.
2. Detecting and confirming infection of a specific bacterium by a specific virus.

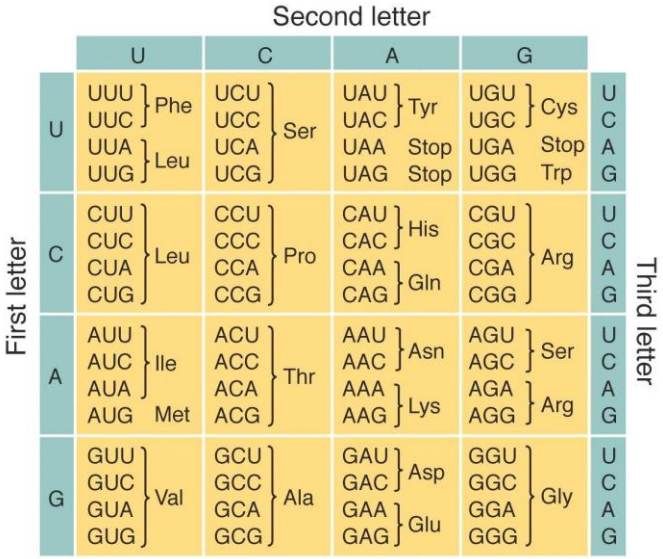
Biological Background

The user of this tutorial needs some biological background to complete it.

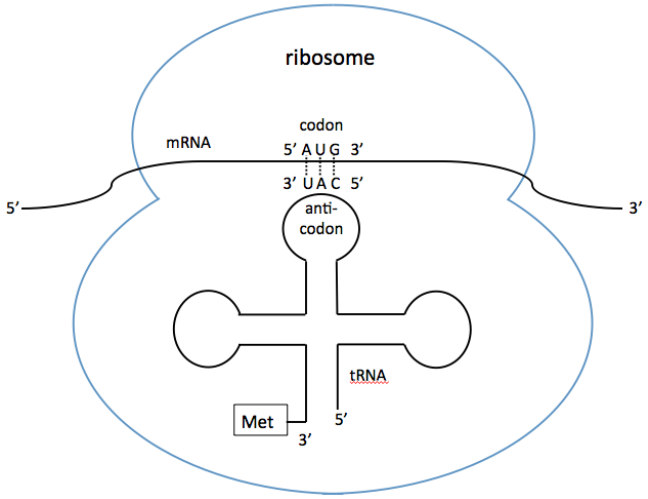
Detecting and Confirming Stop Codon Reassignment

1. The "central dogma" of molecular biology states that the DNA nucleotides of protein coding genes are transcribed to mRNA and then translated to amino acid chains, which form functional proteins.

2. Translation proceeds according to a nearly universal genetic code, which specifies the amino acids coded by triplets of nucleotides known as codons. Translation normally starts with the codon for methionine (Met) and terminates when any of the three stop codons are encountered. The standard genetic code appears below. The single letters of the code stand for nucleotides on mRNA, copied from DNA, with the exception that "T" in DNA is "U" in RNA. The three-letter abbreviations stand for amino acids.



3. Translation of nucleotide codons to amino acids requires transfer RNAs (tRNAs) that have two parts: an "anticodon" that recognizes mRNA codons, and an "acceptor stem" that attaches an amino acid to the tRNA. As mRNA is fed through a molecular machine called a ribosome, tRNAs are brought in and paired to triplets of nucleotides on the mRNA, where G pairs with C and A pairs with U, shown below. In this example, an mRNA triplet of "AUG" is being translated to the amino acid methionine (Met) by means of a tRNA molecule with an anticodon of "CAU." Note that nucleotide sequences have different ends, called 5' and 3', and that by convention nucleotide sequences are written in the 5' to 3' direction. Therefore, although the anticodon below reads left to right as "UAC" it is written as "CAU," because the C is on the side leading to the 5' end of the tRNA and U is on the side leading to the 3' end of the tRNA.



4. It is possible to reassign the meaning of a stop codon from "stop" to "code for an amino acid" by changing the anticodon of a normal tRNA to an anticodon that pairs with a stop codon?

Detecting and Confirming Infection of a Specific Bacterium by a Specific Virus

1. Bacteria have a method of storing pieces of DNA taken from the viruses that have infected them. The storage system is called a Clustered Regularly Interspaced Sequences of Palindromic Repeats (CRISPR).
2. In a bacterial CRISPR array, viral DNA fragments are separated from each other by the same DNA sequence, which serves as a divider.

a. Before a viral sequence is removed from its genome, it is called a "protospacer."

b. After a protospacer has been included into a CRISPR array, it is called a "spacer."

Bioinformatic Conventions and Terminology

1. A standard way of storing DNA sequences is in "fasta format." A fasta file consists of identification information for DNA sequences followed by the letters of the DNA sequence itself, in 5' to 3' order.

2. The foundation of bioinformatics is sequence similarity, which is used to infer evolutionary relationships. The closer the similarity of two sequences, the closer their evolutionary relationship. If two protein coding regions of DNA have very similar sequences, they probably come from the same gene in the same organism. A fundamental computer tool for comparing sequence similarity is BLAST (Basic Local Alignment Search Tool). This tool searches for matches between one sequence (known as the "query" sequence) and a database of one or more sequences (known as "subject" sequences). Good matches between the query and subject are aligned on the matching letters, with gaps inserted where necessary to achieve the best alignment. A sample protein alignment is shown below. Due to gaps from insertions and deletions, there are 53 positions in the alignment, even though the query is 51 characters and the subject is 50. Exact matches (identities) are indicated by printing the matching letters, matches of chemically similar amino acids (positives) are indicated by plus signs, and gaps are indicated by hyphens.

Score	Expect	Method	Identities	Positives	Gaps
81.6 bits(200)	3e-21	Compositional matrix adjust.	42/53(79%)	47/53(88%)	5/53(9%)

```
Query  1   MAVTKLV--KHGDSQWNRENRYTAWGGGYDVDLTEKGVSEAKAAGKLLKEEGY  51
           MAVTKLV  +HG+SQWN+ENR+T W   YDVDL+EKGVSEAKAAGKLLKEEGY
Sbjct  1   MAVTKLVLVRHGESQWNKENRFTGW---YDVDLSEKGVSEAKAAGKLLKEEGY  50
```

Required Computer Skills

1. Familiarity with web browsers, downloading data files, and use of the clipboard to transfer data.
2. Familiarity with web-based server programs that employ a browser-based graphical user interface to enter or upload data and employ drop-down menus to change program parameters.

TUTORIAL

Detecting Possible Stop Codon Reassignment

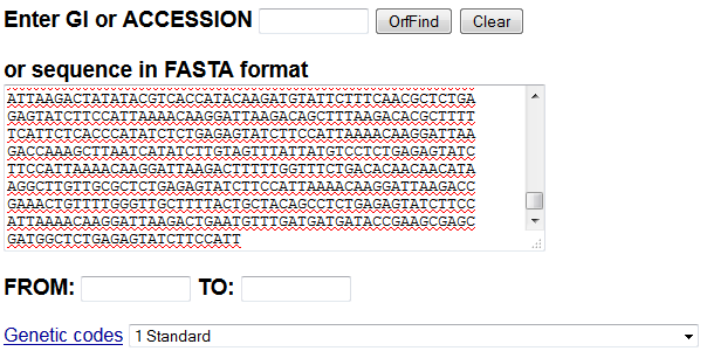
Tool: This exercise makes use of an NCBI web tool called "ORF Finder (Open Reading Frame Finder)" available at http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi. Each gene exists within an "Open Reading Frame" or ORF, which is bound by a start codon (as we said above, this is usually ATG for the standard code) and a stop codon (as we said above, this is usually TAA, TGA, or TAG for the standard code). Because codons consist of triplets, there are three reading frames on each strand of DNA, depending on whether the codons are grouped starting with the first, second, or third nucleotide on the strand. That makes six reading frames, three for the "plus" strand and three for the "minus" strand. Given a fragment of DNA, the ORF finder tool finds all possible open reading frames of a minimum length on each of the six possible reading frames on the given fragment using the genetic code selected.

Method: For comparison purposes, we will first consider a DNA fragment from an organism that uses the standard genetic code. We will submit the DNA sequence to the ORF Finder tool and confirm that it has good potential for coding a protein, assuming use of the standard genetic code. We will then submit another DNA fragment to ORF Finder and observe a poor coding potential assuming use of the standard genetic code and improved coding potential assuming an alternate genetic code. In principal, this is the method used in the paper for the initial detection of DNA fragments that might be recoded.

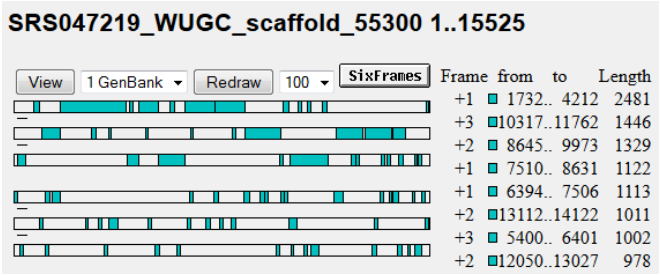
Procedure:

1. Download the tutorial file Prevotella.fasta.

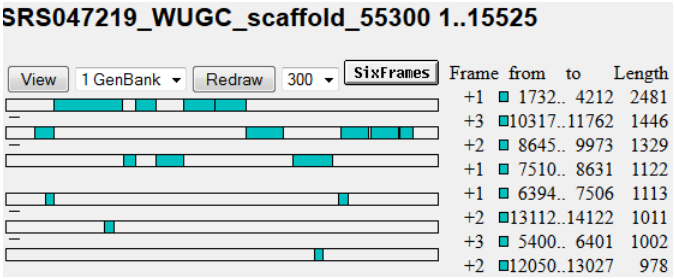
2. Go to http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi and paste the DNA sequence from Prevotella.fasta into the ORF finder window as shown. Note that Genetic Code 1 (Standard) has been selected for you.



3. Press the “OrfFind” button. The following screen will appear.

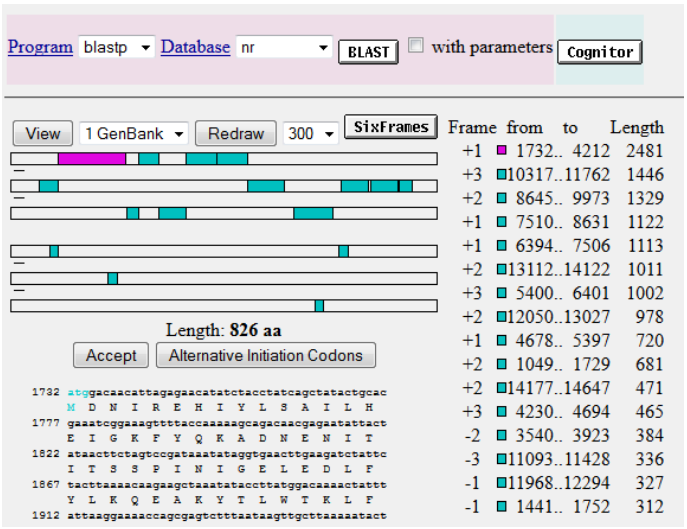


The screen depicts the location and length of each ORF that is at least 100 nucleotides (33 codons) long on each of the six reading frames of the DNA fragment, assuming no stop codon reassignment. Longer ORFs are likely to code for proteins and shorter ones are not. To filter out shorter ORFs, select “300” from the pull-down list (meaning “show only ORFs longer than 300 nucleotides”) to the right of the “Redraw” button, then press the “Redraw” button. You should see the screen below.

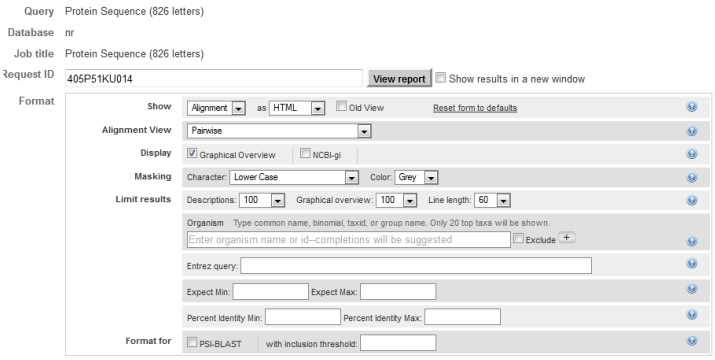


Notice how the remaining longer ORFs nearly continuously cover the sequence from left to right. This means that nearly all of the DNA is likely to code for a protein in one of its six reading frames on its two strands: It has high coding potential assuming use of the standard genetic code. Now, let’s confirm this by using BLAST to see if these longer ORFs code for proteins that are similar to known ones in known organisms. If so, this confirms use of the standard genetic code.

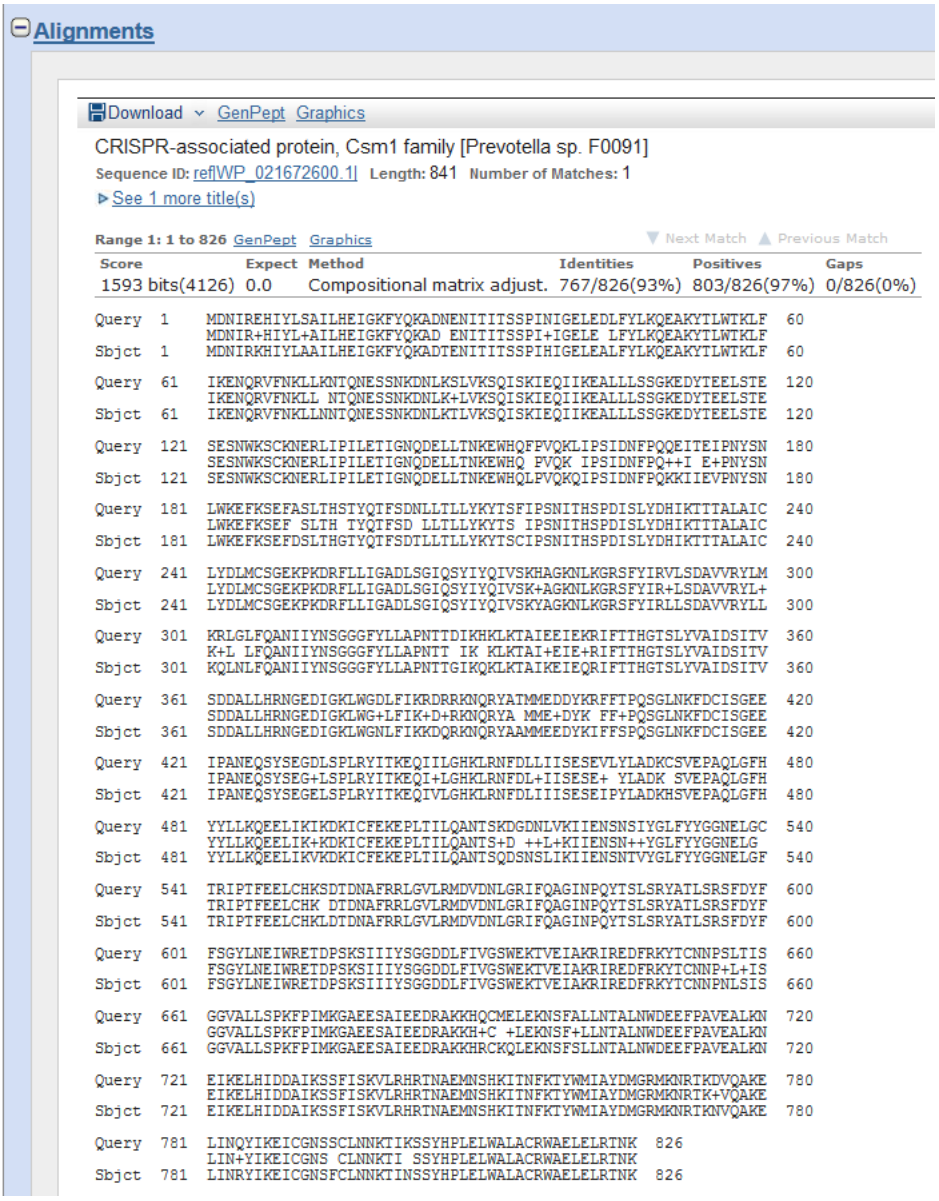
4. Click on the long ORF in the upper left and the screen now shows this:



The amino acid translation for this ORF appears at the bottom left. Now press the BLAST button to search the nonredundant (nr) protein database at GenBank to search for similar amino acid sequences. You will now see this screen:



Press the “View report” button. When the results screen appears, scroll down to the first sequence in the “Alignment” section of the screen and you will see this:



Alignments

Download GenPept Graphics

CRISPR-associated protein, Csm1 family [Prevotella sp. F0091]

Sequence ID: ref|WP_021672600.1| Length: 841 Number of Matches: 1

See 1 more title(s)

Range 1: 1 to 826 GenPept Graphics Next Match Previous Match

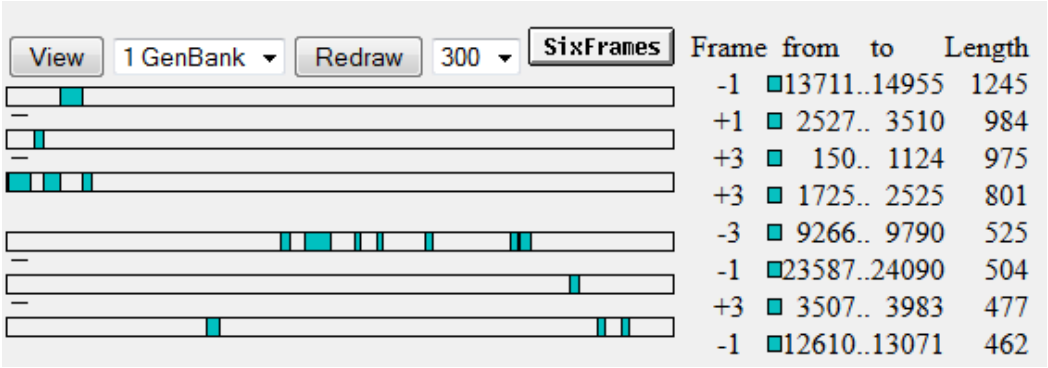
Score	Expect	Method	Identities	Positives	Gaps
1593 bits(4126)	0.0	Compositional matrix adjust.	767/826(93%)	803/826(97%)	0/826(0%)

```
Query  1    MDNIREHIYLSAILHEIGKFYQKADNENITITSSPINIGELEDLFYLKQEAKYTLWTKLF  60
            MDNIR+HIYL+AILHEIGKFYQKAD ENITITSSPI+IGELE LFYLKQEAKYTLWTKLF
Sbjct  1    MDNIRKHIYLAAILHEIGKFYQKADTENITITSSPIHIGELEALFYLKQEAKYTLWTKLF  60

Query  61   IKENQRVFNKLLKNTQNESSNKDNLKSLVKSQISKIEQIIKEALLLSSGKEDYTEELSTE  120
            IKENQRVFNKLL NTQNESSNKDNLK+LVKSQISKIEQIIKEALLLSSGKEDYTEELSTE
Sbjct  61   IKENQRVFNKLLNNTQNESSNKDNLKTLVKSQISKIEQIIKEALLLSSGKEDYTEELSTE  120

Query  121  SESNWKSCKNERLIPILETIGNQDELLTNKEWHQFPVQKLIPSIDNFPQQEITEIPNYSN  180
            SESNWKSCKNERLIPILETIGNQDELLTNKEWHQ PVQK IPSIDNFPQ++I E+PNYSN
Sbjct  121  SESNWKSCKNERLIPILETIGNQDELLTNKEWHQLPVQKQIPSIDNFPQKKIIEVPNYSN  180

Query  181  LWKEFKSEFASLTHSTYQTFSDNLLTLLYKYTSFIPSNITHSPDISLYDHIKTTTALAIC  240
            LWKEFKSEF SLTH TYQTFSD LLTLLYKYTS IPSNITHSPDISLYDHIKTTTALAIC
Sbjct  181  LWKEFKSEFDSLTHGTYQTFSDTLLTLLYKYTSCIPSNITHSPDISLYDHIKTTTALAIC  240

Query  241  LYDLMCSGEKPKDRFLLIGADLSGIQSYIYQIVSKHAGKNLKGRSFYIRVLSDAVVRYLM  300
            LYDLMCSGEKPKDRFLLIGADLSGIQSYIYQIVSK+AGKNLKGRSFYIR+LSDAVVRYL+
Sbjct  241  LYDLMCSGEKPKDRFLLIGADLSGIQSYIYQIVSKYAGKNLKGRSFYIRLLSDAVVRYLL  300

Query  301  KRLGLFQANIIYNSGGGFYLLAPNTTDIKHKLKTAIEEIEKRIFTTHGTSLYVAIDSITV  360
            K+L LFQANIIYNSGGGFYLLAPNTT IK KLKTAI+EIE+RIFTTHGTSLYVAIDSITV
Sbjct  301  KQLNLFQANIIYNSGGGFYLLAPNTTGIKQKLKTAIKEIEQRIFTTHGTSLYVAIDSITV  360

Query  361  SDDALLHRNGEDIGKLWGDLFIKRDRRKNQRYATMMEDDYKRFFTPQSGLNKFDCISGEE  420
            SDDALLHRNGEDIGKLWG+LFIK+D+RKNQRYA MME+DYK FF+PQSGLNKFDCISGEE
Sbjct  361  SDDALLHRNGEDIGKLWGNLFIKKDQRKNQRYAAMMEEDYKIFFSPQSGLNKFDCISGEE  420

Query  421  IPANEQSYSEGDLSPLRYITKEQIILGHKLRNFDLLIISESEVLYLADKCSVEPAQLGFH  480
            IPANEQSYSEG+LSPLRYITKEQI+LGHKLRNFDL+IISESE+ YLADK SVEPAQLGFH
Sbjct  421  IPANEQSYSEGELSPLRYITKEQIVLGHKLRNFDLIIISESEIPYLADKHSVEPAQLGFH  480

Query  481  YYLLKQEELIKIKDKICFEKEPLTILQANTSKDGDNLVKIIENSNSIYGLFYYGGNELGC  540
            YYLLKQEELIK+KDKICFEKEPLTILQANTS+D ++L+KIIENSN++YGLFYYGGNELG
Sbjct  481  YYLLKQEELIKVKDKICFEKEPLTILQANTSQDSNSLIKIIENSNTVYGLFYYGGNELGF  540

Query  541  TRIPTFEELCHKSDTDNAFRRLGVLRMDVDNLGRIFQAGINPQYTSLSRYATLSRSFDYF  600
            TRIPTFEELCHK DTDNAFRRLGVLRMDVDNLGRIFQAGINPQYTSLSRYATLSRSFDYF
Sbjct  541  TRIPTFEELCHKLDTDNAFRRLGVLRMDVDNLGRIFQAGINPQYTSLSRYATLSRSFDYF  600

Query  601  FSGYLNEIWRETDPSKSIIIYSGGDDLFIVGSWEKTVEIAKRIREDFRKYTCNNPSLTIS  660
            FSGYLNEIWRETDPSKSIIIYSGGDDLFIVGSWEKTVEIAKRIREDFRKYTCNNP+L+IS
Sbjct  601  FSGYLNEIWRETDPSKSIIIYSGGDDLFIVGSWEKTVEIAKRIREDFRKYTCNNPNLSIS  660

Query  661  GGVALLSPKFPIMKGAEESAIEEDRAKKHQCMELEKNSFALLNTALNWDEEFPAVEALKN  720
            GGVALLSPKFPIMKGAEESAIEEDRAKKH+C +LEKNSF+LLNTALNWDEEFPAVEALKN
Sbjct  661  GGVALLSPKFPIMKGAEESAIEEDRAKKHRCKQLEKNSFSLLNTALNWDEEFPAVEALKN  720

Query  721  EIKELHIDDAIKSSFISKVLRHRTNAEMNSHKITNFKTYWMIAYDMGRMKNRTKDVQAKE  780
            EIKELHIDDAIKSSFISKVLRHRTNAEMNSHKITNFKTYWMIAYDMGRMKNRTK+VQAKE
Sbjct  721  EIKELHIDDAIKSSFISKVLRHRTNAEMNSHKITNFKTYWMIAYDMGRMKNRTKNVQAKE  780

Query  781  LINQYIKEICGNSSCLNNKTIKSSYHPLELWALACRWAELELRTNK  826
            LIN+YIKEICGNS CLNNKTI SSYHPLELWALACRWAELELRTNK
Sbjct  781  LINRYIKEICGNSFCLNNKTINSSYHPLELWALACRWAELELRTNK  826
```

There is exceptional similarity in amino acid length (841 versus 826) and amino acid similarity (93% exact matches, with 0 gaps) between the selected ORF and a known sequence in GenBank. This confirms use of the standard genetic code for our fragment of DNA. It also tells us that our DNA fragment comes from a bacterium of the genus *Prevotella.* This will be important later.

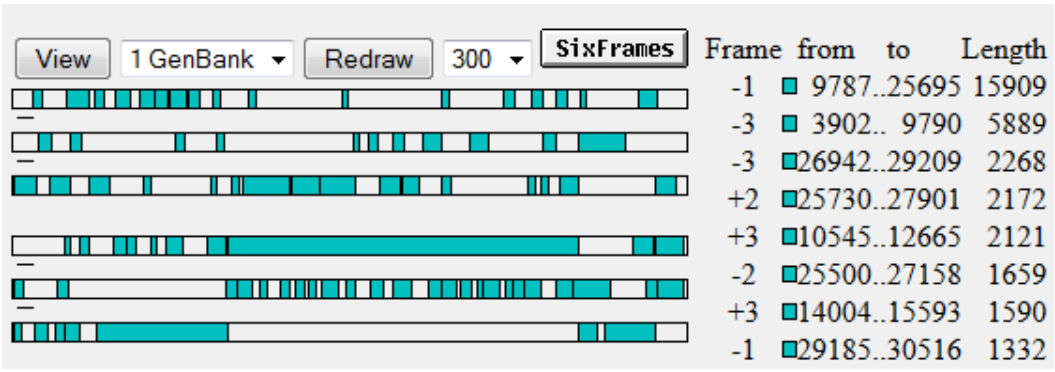
5. Now download the tutorial file called Phage_6_a.fasta.

6. Paste the DNA sequence from Phage_6_a.fasta into the ORF finder window, press the “OrfFind” button, then redraw the results with ORF length 300 selected, just as you did above. Now you will see this:

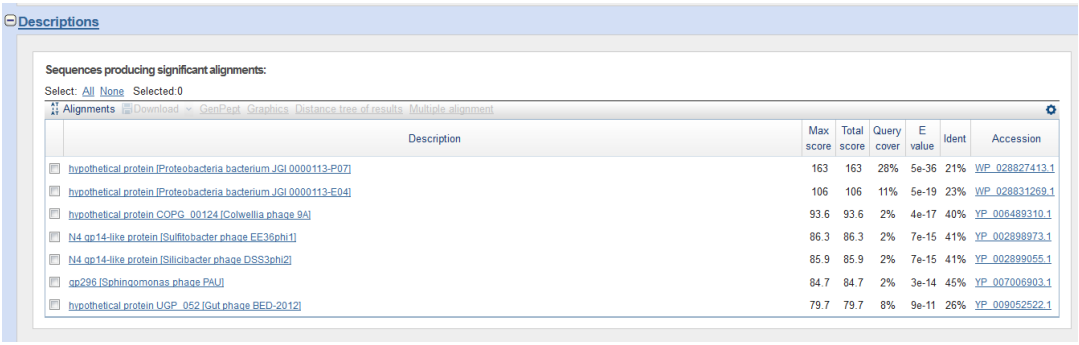


Notice that there are long stretches of sequence where no ORFs are predicted in any of the six reading frames. This means that the coding potential for this DNA fragment is low assuming use of the standard genetic code. Now let's go back and check the coding potential assuming stop codon reassignment.

7. Press the back button on your browser until you get to the ORF finder screen and select genetic code 6 (ciliate, Dasycladacean, and Hexamita Nuclear) from the pull down menu near the bottom of the screen. Again, redraw at 300.
8.



There is a dramatic increase in coding potential assuming an alternate genetic code. If we click on the longest ORF and run BLAST, as we did above for the *Prevotella* fragment, we get these results:



The descriptions of the sparse partial matches provide evidence that the DNA might come from a phage. However, because they do not cover the entire sequence at high similarity, they do not provide clear evidence that the stop codons in the sequence really are being assigned to amino acids. We need more evidence. Let's see if we can find a tRNA that recognizes stop codons in this phage.

Confirming Predicted Stop Codon Reassignment Using tRNA Similarity

Tool: This exercise makes use of the tRNAscan-SE Search Server maintained by the Lowe Lab at the University of California, Santa Cruz, at http://lowelab.ucsc.edu/tRNAscan-SE/ . Given a DNA fragment, this web service searches for tRNAs on the fragment and predicts their anticodon and (if possible) their amino acid.

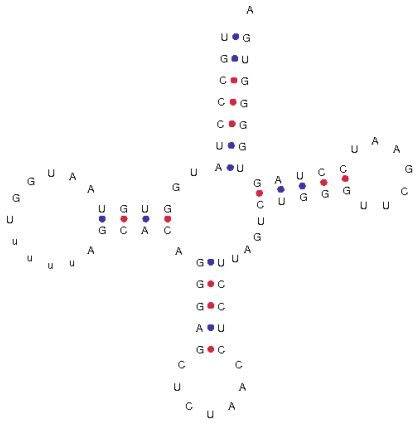
Method: We will submit a DNA fragment from Phage 6 to the tRNAscan-SE Search Server, run it with default parameters, and see if it predicts the presence of a tRNA that recognizes stop codons.

Procedure:

1. Go to http://lowelab.ucsc.edu/tRNAscan-SE/ and paste the DNA sequence for Phage_6_b_dna.fasta into the window that says "Paste your query sequence(s) here:" or submit it using the "Browse" button.
2. Press the "Run tRNAscan-SE" button. Within a few seconds, you will see these results:

Sequence Name	tRNA #	tRNA Begin	Bounds End	tRNA Type	Anti Codon	Intron Bounds Begin	Intron Bounds End	Cove Score
phage_6_2	1	7695	7624	Met	CAT	0	0	69.20
View tRNA								
phage_6_2	2	7460	7386	Sup	CTA	0	0	49.94
View tRNA								

The second tRNA is type "Sup" for "Suppressor" and its anticodon is CTA. That means it "suppresses" the stop message for the TAG (amber) codon, because TAG is the reverse complement of CTA. If you press the "View tRNA" button, you can confirm the tRNA structure that its anticodon—which appears in the bottom loop of the tRNA—is CTA (the diagram shows CUA, because it is RNA).



The presence of a suppressor tRNA is additional evidence that the phage from which this DNA was recovered has reassigned a stop codon. Now let's try to find the bacterial host that this phage infects. We'll start by looking for evidence that this phage has infected the *Prevotella* host we encountered earlier in this tutorial.

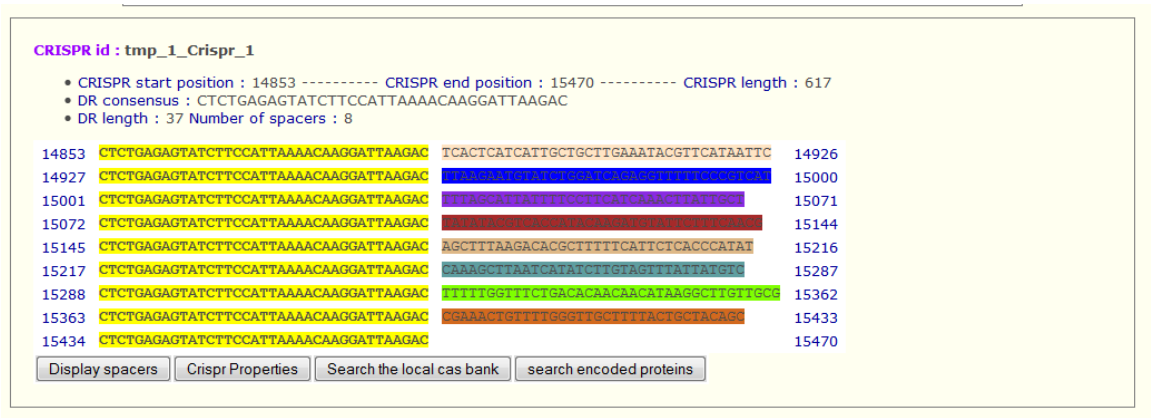
CRISPR Evidence of Phage Infection Crossing a Genetic Code Barrier

Tools: This exercise makes use of the CRISPRfinder tool at http://crispr.u-psud.fr/crispr/CRISPRdatabase.php?page=own . Given a DNA fragment, this web service searches for CRISPR arrays on the fragment and predicts their "spacer" and "repeat" sequences. This exercise also makes use of the "BLAST 2 sequences" tool at http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&BLAST_SPEC=blast2seq, which searches for matches between two DNA fragments. This will allow us to search for matches between spacer sequences in the bacterial host and the DNA sequence of our phage.

Method: We will submit the same *Prevotella* DNA fragment we used before with CRISPRfinder, which will find the spacer sequences in the CRISPR array found on that fragment. We will then search for the spacer sequences on one of the phage DNA fragments we have already used above to demonstrate reassignment of the amber stop codon by a phage.

Procedure:

1. Go to http://crispr.u-psud.fr/crispr/CRISPRdatabase.php?page=own and paste the DNA sequence for Prevotella.fasta into the window that says "Paste your query sequence(s) here:" or submit it using the "Browse" button below the window just mentioned.
2. Press the FindCRISPR button. You will see this screen:

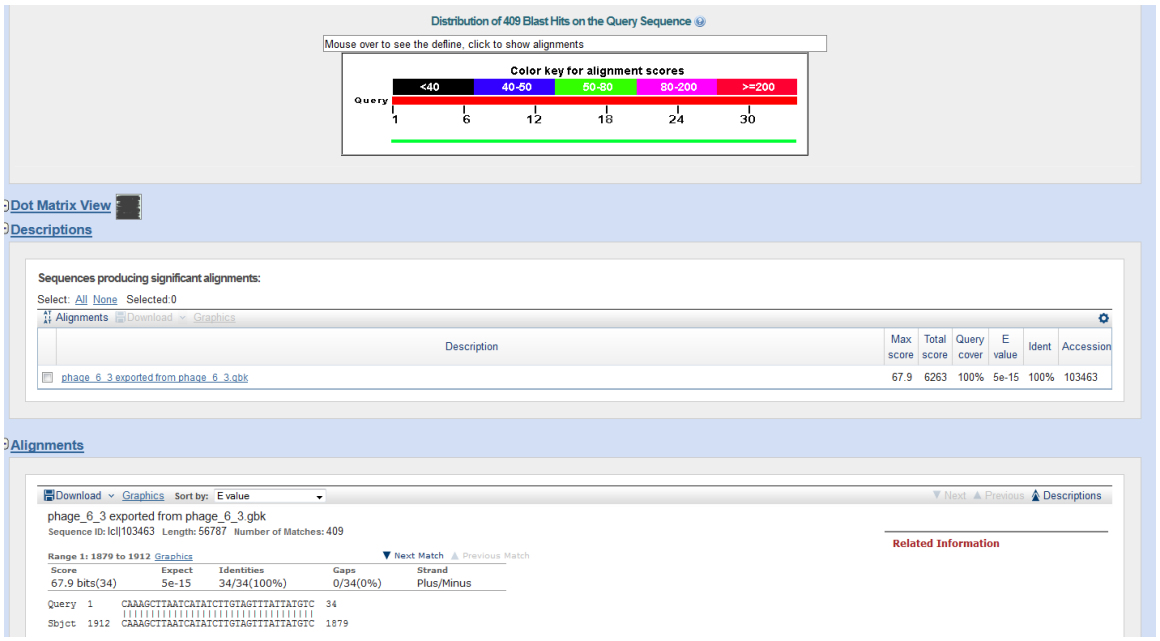


The common repeat sequence in the CRISPR array found on this DNA fragment is shown in yellow on the left. The different spacer sequences of the CRISPR array are multicolored to the left. Press the "Display spacers" button to get a list of the spacers for copying to the clipboard. The screen you see should look like this:

```
>spacer1
TCACTCATCATTGCTGCTTGAAATACGTTCATAATTC
>spacer2
TTAAGAATGTATCTGGATCAGAGGTTTTTCCCGTCAT
>spacer3
TTTAGCATTATTTTCCTTCATCAAACTTATTGCT
>spacer4
TATATACGTCACCATACAAGATGTATTCTTTCAACG
>spacer5
AGCTTTAAGACACGCTTTTTCATTCTCACCCATAT
>spacer6
CAAAGCTTAATCATATCTTGTAGTTTATTATGTC
>spacer7
TTTTTGGTTTCTGACACAACAACATAAGGCTTGTTGCG
>spacer8
CGAAACTGTTTTGGGTTGCTTTTACTGCTACAGC
```

3. Open a new browser window without closing the one that has the spacer sequences shown above. In the new browser window, go to http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&BLAST_SPEC=blast2seq . Use the clipboard to copy and paste spacer sequence #6 into the data entry area that says "Enter Query Sequence." Download the Phage_6_c.dna.fasta file, then press the Browse button under the area that says "Enter Subject Sequence" and navigate to Phage_6_c.dna.fasta for submission. The Phage_6_c.dna file is the long fragment of the Phage 6 genome from which the Phage_6_a.dna file was extracted.

4. Press the BLAST button. The website will search for the spacer sequence on the phage sequence. When it has finished, it will show these results:



At the bottom left, the results under the heading “Identities” show an exact match between all 34 letters of the spacer sequence and 34 letters in the phage genome fragment. The indicator “Plus/Minus” under the heading “Strand” means that the match of the spacer sequence (the query) occurred on the implied opposite strand of the phage genome fragment (the subject). Notice from the alignment of letters at the bottom that the 34 letters of the query from positions 1 to 34 match the reverse complement of the subject sequence from positions 1879 to 1912. The exact match of 34 letters and the “Expect” score of 5e-15 means that there is a high probability that the match is not likely a random one: A phage with amber reassignment has almost certainly infected the host bacterium. Because the phage and host have been shown to use two different genetic codes, this is strong evidence that genetic codes do not represent an impermeable barrier to host infection by a phage.

What Have We Learned?

1. Two indicators of alternate genetic coding in DNA samples are poor coding potential assuming standard coding and the presence of tRNAs that recognize stop codons.

2. CRISPR arrays can be used to link phages and their hosts.

3. There is evidence that stop codon reassignment is not a barrier to infection of bacteria by phage.