

Tutorial on MORF analysis

Outline of document:

1. Questions
2. Codon Analysis – section 1
3. Codon Distribution
4. Codon Distribution/ codon usage table
5. FABG homologs
6. DOG Processing – section 2

Multiple Open Reading Frames (MORF) are exactly what it sounds like. Any gene can be read in 6 different frames. An open reading (ORF) is defined as any region of codons seen between 2 stop codons (TAA, TAG and TGA). These regions in theory could code for a viable protein. There are many examples of ORFs that have more than 1 frame open, they have multiple open reading frames. Genes having MORFs are potentially proteins with ancient ancestry being able to code for multiple proteins. They also are a good way to gain an understanding of the proteins of interest but also give a deeper understanding of the coding system itself.

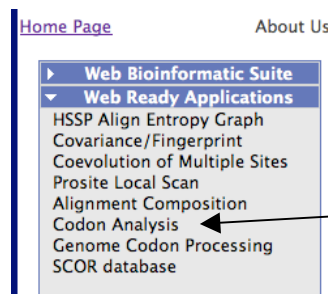
1. Questions:

- a. What is the codon distribution of the *M. tuberculosis* FABG?
- b. What is the main codon use in Beta Keto ACPR proteins?
- c. What is the MORF distribution of Beta Keto ACPR proteins?

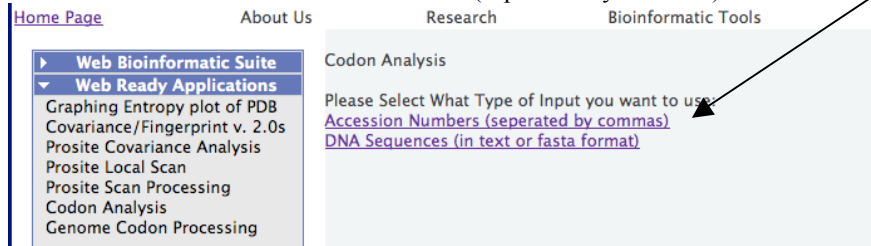
2. Codon Analysis section 1

The first thing we will do is generate a codon table of the FABG from *M. tuberculosis* and also see if it contains MORF for further analysis.

- a. Go to <http://labs.hwi.buffalo.edu/duax/home.php?p=wra>
- b. Click on the link to “codon analysis”



- c. Click on the link “Accession Numbers (seperated by commas)”



MORF Tutorial

This like will bring us to a text box. Enter the swiss protein accession number from *M. tuberculosis* FABG. “P0A5Y4”. Click “Go”

Please Enter an Accession Number: (for multiple accession numbers insert commas, eg A3DFK9, Q12345, etc):

P0A5Y4

Go

- d. The program will run for a bit and will say it is looking for your sequence. It is accessing the swiss prot database and downloaded the sequence for your protein. It is then doing some calculations in the background. Eventually it will stop and will give you an output that look like this:

ACCESSION ID	SPECIES	GC %	MORF TYPE (STOPS SEEN BY FRAME 1-6)	STATUS
P0A5Y4 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium; Mycobacterium tuberculosis complex.	64.40	SORF 1 8 3 2 12 5	Codon Analysis complete. View Now

Codon Analysis output webpage.

- Accession ID
 - Left click on the accession ID will bring up another window that contains the nucleotide sequence of all 6 frames.
 - This is an effective way to get the nucleotide sequence of a protein or many proteins quickly.
- Species
 - Each of the labels under the species are links to <http://www.uniprot.org/>. Each link will present a table with all proteins in the swissprot database that have that title.
 - Clicking “Bacteria” will present a table of all bacterial proteins. Where clicking on “Mycobacterium tuberculosis complex” will present all proteins with this title.
- GC %
 - Nucleotide compositions of G and C.
- MORF type (stops seen by frame1-6)
 - MORF type tells you how many open reading frames are present in this sequence. SORF =1 ORF, DORF=2 ORF, TORF=3 ORF, etc.
 - The numbers after the MORF type indicate how many STOP codes are present in each frame.

Number	Frame
1	Sense (5'3')
8	Sense+1
3	Sense+2
2	Antisense (3'5')
12	Antisense+1
5	Antisense+2
- Status
 - As the program is running, status gives an update of what the program is doing “Looking for P0A5Y4”
 - Once done, the status indicates Codon Analysis Complete with a link “view now”
 - Click “view now” will bring up the codon usage table from that protein.
 - The codon analysis table is set up to represent all the codons used in all 6 frames. This gene has a high GC content. As such the GC rich and GC only codon are used more often then the AT rich and only.

MORF Tutorial

add a caption to title attribute or leave blank												close or Esc Key	
CODON	COMPLEMENT	F1: A	F1: B	F1: T	F2: A	F2: B	F2: T	F3: A	F3: B	F3: T	TOTAL A	TOTAL B	TOTAL A+B
TTA-L	TAA-*QY	0	0	0	2	0	2	0	0	0	2	0	2
TAT-Y	ATA-IM	1	2	3	3	2	5	4	0	4	8	4	12

e. Codon Analysis Table

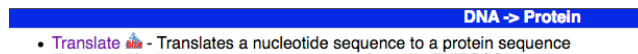
- i. Codon: 32 codons
- ii. Complement: complements of the 32 codons to make a total of 64
- iii. F1:A – Codon counts of 32 codons
- iv. F1:B – Codon count of complement 32 codons
- v. F1:T – Total count of 32 codons and their complements.
- vi. F2:A – sense strained +1 Codon counts of 32 codons
- vii. F2:B – sense strained +1 Codon count of complement 32 codons
- viii. F2:T – sense strained +1 Total count of 32 codons and their complements.
- ix. F3:A – sense strained +2 Codon counts of 32 codons
- x. F3:B – sense strained +2 Codon count of complement 32 codons
- xi. F3:T – sense strained +2 Total count of 32 codons and their complements
- xii. Total A – Counts in 32 codons from all 6 frames
- xiii. Total B - Counts in 32 complementary codons from all 6 frames
- xiv. Total A+B – Counts in triple codons codons.

Note: To see the antisense codons just reverse the Codon and Complement columns.

What are the least used codons FABG?

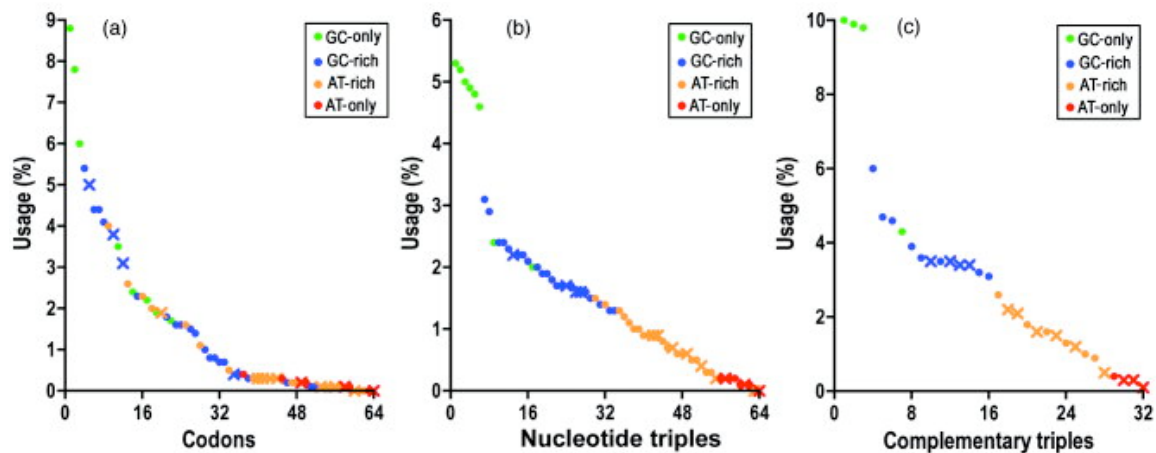
What are the most used codons in FABG?

- f. Where are the stops seen in the sequence? Are they at the beginning of the sequence or at the end or in the middle?
- i. Go to expasy.org and click on “DNA-> Protein”
 - ii. Click on the link at the top of the page “Translate”



- iii. Get the FABG nucleotide sequence from the Accession ID link. Paste it into the text box. Click “Translate” sequence”.
- iv. *Where are the stop codons located in all 6 frames? Is there any frame where you can put together long sequences of amino acids (>100 amino acids)?*
- v. For those who did not answer the last question but jumped ahead, there is a long sequence of ~150 amino acids on the antisense strand. *Is this a protein? Is there a protein on the antisense strand of FABG from M. Tuberculosis?*
- vi. Go back to the tutorial on Identifying homologs of a sequence “How to Blast a Sequence.” Blast the ~150aa sequence from the antisense strand of FABG. *What proteins are identified? Looking at the sequence overlap, what can we say about this stretch of Amino acids?*

3. Codon Distribution



Proteins Vol.61, 4 Pages: 900-906.

FABG belongs to the family of Short Chain Oxido reductase enzymes. This is a very large family of protein >16,000 in the gene bank. Plotted above is the distribution of 84 SCOR genes having three open reading frames. (a) Codon use (b) nucleotide triple frequency, and (c) complementary pairs of nucleotide triples. Codons that have different definitions in different species (TTG, CTT, CTC, CTA, CTG, ATA, ATG, GTG, TAA, TAG, AAA, TGA, AGA, and AGG) are identified by X.

- a. How does the FABG SORF compare to these 84 TORFs? Compare the codon use, nucleotide triple frequency and complementary pairs of FABG.
- b. All the data we need is present in the codon analysis table.
 - i. 64 codons plotted in (A). Using your favorite graphing program plot out the codons in F1:A and F1:B.
 - ii. 64 nucleotide triples in (B) plot columns Total A and Total B.
 - iii. 32 complementary triples are plotted from column Total A+B.
- c. Answer question in tab a.

4. Codon Distribution/ codon usage table

The codon distribution for FABG can be elucidated from the codon analysis table. This is only one protein of many in the genome. *How does this compare to the overall codon usage of mycobacterium tuberculosis genome?*

- a. Go to the Codon usage database <http://www.kazusa.or.jp/codon/>
- b. In the query box type "tuberculosis" and then click submit.

QUERY Box for search with Latin name of organism

tuberculosis

Case: ☒ sensitive ☐ insensitive

- c. This should bring up a list of those entries in the codon database with tuberculosis in their titles.

Answer for your query "tuberculosis" (case: sensitive search).

[Yersinia pseudotuberculosis \[gb|ct\]: 271](#)
[Corynebacterium pseudotuberculosis \[gb|ct\]: 6](#)
[Mycobacterium avium subsp. paratuberculosis \[gb|ct\]: 91](#)
[Mycobacterium tuberculosis \[gb|ct\]: 107](#)
[Mycobacterium avium subsp. paratuberculosis K-10 \[gb|ct\]: 4353](#)
[Yersinia pseudotuberculosis IP 32953 \[gb|ct\]: 4038](#)
[Mycobacterium tuberculosis F11 \[gb|ct\]: 3941](#)
[Yersinia pseudotuberculosis IP 31758 \[gb|ct\]: 4324](#)
[Mycobacterium tuberculosis H37Ra \[gb|ct\]: 4043](#)
[Mycobacterium tuberculosis CDC1551 \[gb|ct\]: 4189](#)
[Mycobacterium tuberculosis H37Rv \[gb|ct\]: 3998](#)

[Homepage](#)

- d. Click on the link to "*M. tuberculosis H37Ra*"
- e. The codon table that comes up is a compilation of all *M. tuberculosis H37Ra* genes sequenced.

Mycobacterium tuberculosis H37Ra [gb|ct]: 4043 CDS's (1347530 codons)

fields: [triplet] [frequency: per thousand] [(number)]

UUU 6.2(8324)	UCU 2.2(2991)	UAU 6.1(8217)	UGU 2.2(2994)
UUC 23.3(31382)	UCC 11.5(15542)	UAC 14.7(19799)	UGC 6.7(9014)
UUA 1.6(2286)	UCA 3.6(4807)	UAA 0.5(636)	UGA 1.6(2213)
UUG 17.9(24111)	UCG 19.4(26090)	UAG 0.9(1194)	UGG 14.7(19824)
CUU 5.5(7373)	CCU 3.4(4614)	CAU 6.4(8680)	CGU 8.5(11429)
CUC 17.3(23375)	CCC 17.0(22958)	CAC 15.8(21357)	CCG 28.5(38368)
CUA 4.8(6458)	CCA 6.2(8323)	CAA 8.1(10926)	CGA 7.2(9759)
CUG 50.4(67850)	CCG 31.3(42235)	CAG 22.7(30644)	CGG 24.7(33291)
AUU 6.5(8697)	AUC 3.7(4955)	AUA 5.3(7125)	AUG 3.6(4798)
AUC 33.9(45700)	ACC 35.1(47359)	AAC 19.8(26700)	AGC 14.5(19487)
AUA 2.2(2972)	ACA 4.6(6186)	AAA 5.3(7143)	AGA 1.3(1777)
AUG 18.4(24812)	ACG 15.7(21103)	AAG 15.0(20247)	AGG 3.2(4369)
GUU 8.1(10898)	GCU 11.0(14836)	GAU 15.8(21284)	GGU 18.8(25324)
GUC 32.7(44105)	GCC 59.8(80539)	GAC 42.1(56762)	GGC 50.4(67879)
GUA 4.8(6436)	GCA 12.9(17379)	GAA 16.2(21845)	GGA 10.0(13412)
GUG 40.0(53959)	GCG 48.6(65466)	GAG 30.6(41207)	GGG 19.1(25785)

- f. Now we can plot this data and compare it to the gene data. We can then tell if there is a bias usage of a codon with FABG. *Is there a usage of a codon in FABG that is not used widely in the tuberculosis genome?*

5. FABG homologs

- a. The codon analysis program can handle more than one protein for analysis. *Is the codon usage and MORF characterization for FABG from *M. tuberculosis* seen in other homologous FABG's?*
- b. From past tutorials we have identified ~20 FABGs from other species. Since the BLAST search from NCBI does not give us swiss prot ID, we are going to have to get those by re-blasting through swissprot.
- c. Go To www.expasy.org
- d. Type into the search "P0A5Y4" and click search.
- e. Scroll down the webpage of P0A5Y4 to the sequences tab. Under tools it should say Blast. If it does click "go"

Sequences				
Sequence	Length	Mass (Da)	Tools	
<input type="checkbox"/> P0A5Y4-1 [UniParc]. FASTA Last modified March 15, 2005. Version 1. Checksum: 70F6254B0FFCD47	247	25,697	Blast	go

- f. Wait for the Blast program to complete. This will give a collection of FABG homologs from other species. Instead of searching the Nonredundent database, which we searched through NCBI, this searched the swissprot database. Get the accession number from 10 or so hits of FABG from different species.

MORF Tutorial

None	Accession	Entry name	Status	Local alignment	Protein names	Organism
<input type="checkbox"/>	A5U2I7	A5U2I7_MYCTA	★		3-oxoacyl-(Acyl-carrier-protein)	Mycobacterium tuberculosis (strain ATCC 25177 / H37Ra)
<input type="checkbox"/>	A1KIS3	A1KIS3_MYCBP	★		Probable 3-oxoacyl-[acyl-carrier protein] reductase fabG1	Mycobacterium bovis (strain BCG / Pasteur 1173P2)
<input type="checkbox"/>	A4KH28	A4KH28_MYCTU	★		3-oxoacyl-[acyl-carrier protein] reductase	Mycobacterium tuberculosis str. Haarlem

- g. Copy the acc # in to the codon analysis online form. Separate each acc # by a comma. Click “go”.

Please Enter an Accession Number: (for multiple accession number eg A3DFK9, Q12345, etc):

Each acc# is separated by a comma.

- h. The results will tell a story of FABG evolution.

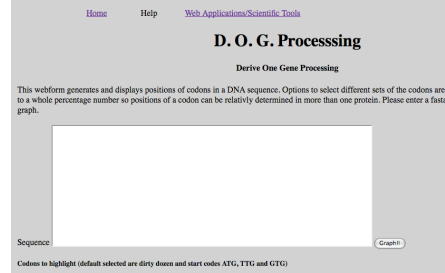
ACCESSION ID	SPECIES	GC %	MORF TYPE (STOPS SEEN BY FRAME 1-6)	STATUS
P0A5Y4 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium; Mycobacterium tuberculosis complex.	64.40	SORF 1 8 3 2 12 5	Codon Analysis complete. View Now
A1KIS3 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium; Mycobacterium tuberculosis complex.	64.40	SORF 1 8 3 2 12 5	Codon Analysis complete. View Now
B2HPD4 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium.	65.78	SORF 1 8 3 1 11 5	Codon Analysis complete. View Now
A1T8Q9 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium.	68.93	SORF 1 7 1 1 11 4	Codon Analysis complete. View Now
A0QHT9 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium; Mycobacterium avium complex (MAC).	69.04	DORF 1 10 1 0 9 4	Codon Analysis complete. View Now
P71534 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium.	68.54	DORF 1 6 4 0 13 6	Codon Analysis complete. View Now
B8ZS81 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Mycobacteriaceae; Mycobacterium.	62.61	SORF 1 9 5 3 14 2	Codon Analysis complete. View Now
Q0S0F9 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Nocardiaceae; Rhodococcus.	68.85	SORF 1 9 3 1 13 3	Codon Analysis complete. View Now
Q5YU16 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Corynebacterineae; Nocardiaceae; Nocardia.	69.07	DORF 1 8 2 0 15 5	Codon Analysis complete. View Now
Q47NU3 (translation)	Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales; Streptosporangineae; Nocardiopepsaceae; Thermobifida.	68.30	DORF 1 5 1 0 14 4	Codon Analysis complete. View Now

- i. From the run above we can see 4 of the FABG homologs have double open reading frames. Both the sense and the antisense strand are open. We can do the same type of analysis we did with the single FABG to all of these new FABG's.
- j. Do any of the above FABGs have any protein homologs on one of their other strands?

6. Tutorial on DOG Processing Section 2

Genomes today are sequenced at a very fast rate. This is a lot of information that needs to be annotated and analyzed. Several rules are followed to increase the chances of correct annotation. We have developed a tool that would enhance the correct selection of a coding frame (MORF) and correct selection of the Start and Stop codons.

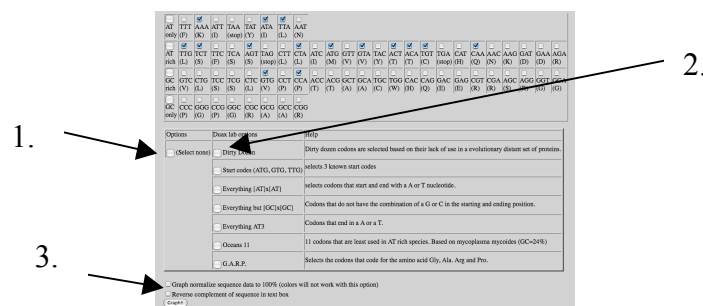
- For an example we are going to analyze a SCOR protein from the genome of *Anaeromyxobacter dehalogenans* 2CP-C.
- Get the nucleotide sequence with the NCBI ID: YP_467052
 - Go to ncbi.nih.gov, put the ID in the search box, search and click gene.
 - Ensure the gene is from *A. dehalogenans*.
 - Follow instructions from the database tutorial.
- Go to: http://labs.hwi.buffalo.edu/duax/phpmysql/webapplications/genome_a/



- Copy and paste in Fasta format the sequence into the above text box.
 - You should paste something like this

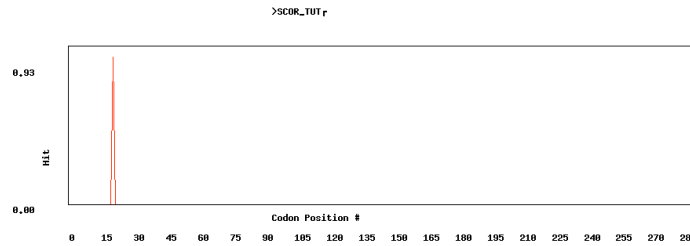
```
>SCOR_TUT
tcatgcatcc tccgtcgcc ggtggaccgc gcccgagacg cggcgcagga tcgccgggg
61 gaccaggcgc gacgcggccg cggtagctt gtatgcggg tcggcaccc acaccacetc...
```

- We are searching for the dirty dozen codons in this sequence. Default does not select the dirty dozen. So click under options “Select none” (1). This should remove all selected codon boxes.
- Click the button that says “Dirty Dozen” (2) Dirty dozen codons are selected based on their lack of use in an evolutionary distant set of proteins. These are potential stop codes in *A. dehalogenans* genome.



- This will select 12 codons that were described in the lecture portion of the course.
- For those that were observant, the nucleotide sequence is the complement of the coding gene. To the sequence we pasted in we need to select the “Reverse complement of sequence in text box” (3) button so that the program identify the correct frame.
- Click “Graph”
- You should get a graph that looks like this:

MORF Tutorial



- k. Where the sequence length runs along the bottom the x-axis and the codon position runs along the y-axis. Below the graph there is a table that describes what codon was selected.
- l. Codon TTA was found 21 codons in from the 5' end.
- m. If we plot this codon on to the SCOR sequence, we find that the program that was used to locate genes in this genome was wrong in identifying the start of this gene. The first 21 amino acids are GARP rich and contain a stop code. The start should have been the Leu residue directly next to the Stop Leu or the first Met that appeared.

>SCOR_TUT_protein sequence

MRGPAAPPRPGEPGLPGPGL | LHSPVMAPPGAPEPLAVVTGAS.....

- n. Note: The rest of the SCOR sequence does not contain any Dirty dozen (potential stop codes). This type of analysis can be done with any gene. There was significant analysis that went into the *A. dehalogenans* genome to develop the dirty dozen codons. What is seen as a potential stop in one genome could code for an amino acid in another. This is one reason if we ran this analysis on *M. tuberculosis* FABG we would incorrectly assume what we find to be true.