Laboratory VI
Codon Bias

Buffalo State
State University of New York

Genetics
Biology 303
Spring 2007
Dr. Wadsworth

Introduction

The genetic code is described as degenerate because there is more than one codon for most amino acids. For example, there are four codons corresponding to the amino acid valine, GUU, GUC, GUA, GUG. All four of the valine codons effectively code for valine in the polypeptide chain. Therefore, we might expect each of the valine codons to be used in about equal proportions. However, this is not the case for many species. For example, analysis of genes in *E. coli* shows that some valine codons are used more frequently than others are. The GUU codon is used in 25%, GUC is used in 21%, GUA is used in 15%, and GUG is used in 38% of the time. This phenomenon of unequal use of codons with identical functions is referred to as **codon bias**.

The biological significance of codon bias is unclear. The uneven use of codons is too extreme to be accounted for by chance deviation. Additionally, different species show different codon biases. For example *E. coli* prefers the GUG codon for valine (38%) while *H. sapiens* uses this codon only 10% of the time and instead prefer the GUC codon for valine (40%).

Many different explanations have been proposed to explain codon bias. Some researchers have hypothesized that codon bias is a genetic adaptation to the slight difference in translational machinery found in different species. For example, some species might not contain equal amounts of all the cognate tRNA's for a particular amino acid. Consider the codons for valine. As mentioned before there are several different codons for the amino acid valine. Therefore the cell could have several different tRNA for valine with differing anticodons. It is possible that these different leucine tRNA are not all equally abundant in the cell. For example, it is possible that in *E. coli* there are more valine-tRNA's that bind to GUG codons than there are valine-tRNA's that bind to GUA codons. Therefore, there may be selection for alleles that use the GUG codon for valine versus the GUA codon. After thousands of generations of selection this will result in codon bias in the genome.

Others have proposed that codon bias is not a response to selective pressure caused by biases in tRNA populations. Alternative factors proposed to lead to codon bias include sequences related to patterns of mutation bias in the DNA, secondary structure in mRNA, sequences that promote stability of mRNA and sequences that facilitate subcellular localization of the mRNA and the protein products.

Laboratory Exercise

For this laboratory exercise you will investigate genes from a single phylum to determine if that phylum demonstrates codon bias. Specifically, you will identify 5 different protein-encoding genes for your phylum. You will then translate the ORF of the genes and identify all the valine encoding codons. You will then prepare a summary table reporting the frequency at which each of the codons is used. You will use Chi-Square to test whether your phylum shows random codon usage.

The phylum you will investigate is:

______________________________________.

DNA Databases

To conduct this analysis you will need to find DNA sequences of different genes within your assigned phylum. The world's largest collection of gene sequences is available in the DNA databases of the National Center for Biotechnology Information. The easiest way to access this database is on the internet interface known as "Entrez". Entrez operates like internet search engines such as "Google" except it searches DNA databases.

Instructions for identifying genes on Entrez

1. Access the Entrez site on your internet browser using the following URL.

   http://www.ncbi.nlm.nih.gov/Entrez/

2. Click on "Nucleotide" on the black bar to search for gene sequences.

3. Enter your search parameters in to the search window. So that you can identify complete gene sequences from your phylum, I suggest that you use the following terms for your search parameters.

"**your phylum"   complete   cds**

4. To examine the first database entry click on the first blue accession number.

5. Review the components of the database entry. Particularly look at the "organism" line to confirm that this gene comes from your phylum. Because of the way the database is organized, it is possible your search will identify genes from other phylum. If the gene is from another phylum return to your search results to examine the next database entry. Additionally, it is important to not use any genes from the mitochondrial genome. The mitochondrion has a different codon bias than the nucleus.

6. To identify the sequence corresponding to the ORF, find the **CDS** link under "Features". Click this link to get a database entry for the portion of the gene corresponding to the ORF.

7. Next to the display button there is a drop down window. Select "Fasta" on the drop down window and click the display button. This will generate and ORF sequence with no notations or numbers. You can use the mouse to select and copy the DNA sequence in this FASTA format.

Translation of ORF

NCBI offers a number of on-line programs for analyzing DNA sequences. One of the programs, **ORF Finder**, identifies the open reading frames in a gene sequences and displays the predicted amino acid sequence above the nucleotide sequence.

Instructions for translating ORF on ORF Finder.

1. Go the NCBI home page. Click on the NCBI link in the upper corner of the entrez page or go to the following ORF.

   http://www.ncbi.nlm.nih.gov/

2. Click on the "ORF Finder" link on the right hand side of the page listed under "Hot Spots".

3. Paste your FASTA format sequence of the CDS you copied from entrez in the FASTA window. Click the ORFFind button.

4. A diagrammatic representation of the six possible reading frames will be displayed. Each bar represents one of the possible frames. Green regions identify the location of the ORF,s with start and stop codons. Typically, the top bar will have the longest green region. Click on the bar to obtain the sequence of this frame.

5. Scroll down and exam the DNA and Polypeptide sequence. The Start and Stop Codons are indicated in Green and Purple respectively.

6. Copy the DNA/Polypeptide Sequencing into a word document. Make sure the document is in a Courier Font and adjust the margins to align the DNA sequences.

Analysis of Codon Usage

There are four codons for the amino acid valine, GTT, GTC, GTA, GTG. On the translated sequences, identify all of the valines in the predicted protein. Use a highlighter to mark all of the valine codons in the gene. Record the number of times each valine codon is used.

Chi Square Analysis of Codon Bias.

Use Chi square to test the hypothesis that your phylum displays no codon bias. Your null hypothesis is that the four valine codons are used in equal frequencies in your phylum.

Laboratory Assignment

1. Use Entrez to identify five different genes from your phylum. Your genes must come from at least 3 different species within your phylum. You should use genes with CDS's that are between 300 and 3000 nucleotides long.

2. Identify and translate the ORF of the gene using the ORF Finder program.

3. Print and save a word version of the translated sequence. See the attached example.

4. Identify all the valine codons and record the number of times each codon is used in each gene. Report codon usage in a table format. See attached format

5. Use Chi square to test the hypothesis that codon usage is random in your phylum. Use the combine data from all five genes.

6. Write a brief discussion section of lab report concerning your findings. Follow the format described in **Pechenick**. The discussion section must be typed double spaced. You must also include a literature citation section for all literature you cite in your report.

   Some issues your might address?

   A. What were the general expectations in the study and why?
   B. How did your results compare with the expected results?
   C. Was the general pattern of codon bias consistent for all gene/species tested?
   D. Discuss any unexpected element of your experimental results.
   E. How does your pattern of bias compare with what others have found?
   F. What might this pattern of bias tell you about the biology of your phyllum?
   G. Based on your results, what is the next question that should be addressed? What experiment might be used to address it?

7. Attach your five translated sequences, your summary table and your chi square analysis to your discussion.

8. Report will be due at the beginning of the laboratory period the week of March 13, 2003.

Table I: Valine Codon Usage in Chordata Genes

	Number of Valine Codons			
Accession Number	GTG	GTC	GTT	GTA
AF525460	9	3	1	0
XX12345	12	8	0	3
XX12346	11	7	4	1
ZZ12345	8	5	1	0
QQ12345	25	12	4	1
Total	65	35	10	5

Raw Data Sheet Example

Phylum: Chordata
Species: *Homo sapiens*
Gene Name: alpha-1-globin gene
Accession Number: AF525460

```
  1 atggtgctgtctcctgccgacaagaccaacgtcaaggccgcctgg
    M  V  L  S  P  A  D  K  T  N  V  K  A  A  W
 46 ggtaaggtcggcgcgcacgctggcgagtatggtgcggaggccctg
    G  K  V  G  A  H  A  G  E  Y  G  A  E  A  L
 91 gagaggatgttcctgtccttccccaccaccaagacctacttcc
    E  R  M  F  L  S  F  P  T  T  K  T  Y  F  P
136 cacttcgacctgagccacggctctgcccaggttaagggccacggc
    H  F  D  L  S  H  G  S  A  Q  V  K  G  H  G
181 aagaaggtggccgacgcgctgaccaacgccgtggcgcacgtggac
    K  K  V  A  D  A  L  T  N  A  V  A  H  V  D
226 gacatgcccaacgcgctgtccgccctgagcgacctgcacgcgcac
    D  M  P  N  A  L  S  A  L  S  D  L  H  A  H
271 aagcttcgggtggacccggtcaacttcaagctcctaagccactgc
    K  L  R  V  D  P  V  N  F  K  L  L  S  H  C
316 ctgctggtgaccctggccgcccacctccccgccgagttcacccct
    L  L  V  T  L  A  A  H  L  P  A  E  F  T  P
361 gcggtgcacgcctccctggacaagttcctggcttctgtgagcacc
    A  V  H  A  S  L  D  K  F  L  A  S  V  S  T
406 gtgctgacctccaaataccgttaa 429
    V  L  T  S  K  Y  R  *
```