Expression Patterns of Codon Usage and Homopolymers in the 13-Lined Ground Squirrel

Amy Virta

Advisor: Marshall Hampton
Introduction

Why Squirrels?

The 13-lined ground squirrel, scientifically labeled *Ictidomys tridecemlineatus*, is a furry squirrel abundant in Minnesota and much of the Midwest. Named for the alternating stripes down its back, the 13-lined squirrel is the namesake of the University of Minnesota Golden Gophers [1]. It lives in tunnels underground rather than in trees like its cousins the tree squirrels. The 13-lined squirrel is slightly larger than a chipmunk in size. Most interestingly to scientific researchers, the 13-lined ground squirrel is a true hibernator that sleeps from October to March in their underground den [2].

The work of Dr. Matt Andrews, Ph. D, former department head of the Biology department at the University of Minnesota Duluth, has long centered on hibernating mammals, especially the 13-lined ground squirrel. Dr. Andrews’ lab is currently looking at identifying genes that are responsible for regulating the physiological characteristics of hibernation in the heart, skeletal muscle, and white adipose tissue (WAT) of the thirteen-lined ground squirrel [3]. The hope is that learning what genes are responsible for hibernation can benefit humans in areas such as hypothermia, ischemia/reperfusion injury, cardiac function, and organ preservation [3].

Squirrels are easy to study because of their relative abundance in the area and their size. They are particularly interesting to study because of the extremes their bodies go through during hibernation. Right before going to sleep (torpor), they pack on fat at a rate as high as 4 grams per day [2]. During torpor, their body temperature drops to around 0°C, oxygen consumption is 1/30 to 1/50 of the aroused condition and heart rate can be as low as 3-10 beats/minute, compared to 200-400 beats/minute when awake and active [3]. When they awake in March, they have lost as much as 1/3 of their body weight [2].
These extreme conditions would be deadly to humans, making their study of great interest. Since many of the genes identified in the squirrel also exist in humans [3], comparing the DNA that makes up these genes could shed light on what allows the squirrels to go through such massive physiological changes. For biologists or those in a medical field, this could lead to the development of new drugs or procedures to aid humans when bodies undergo similar physical stresses.

**Biology Background**

Even though 13-lined ground squirrels posses many of the same genes as humans, it is thought that they are expressed differently to allow the animals to hibernate [3]. Gene expression refers to the “detectable effects of a gene” while differential gene expression is “gene expression that responds to signals or triggers” [4]. It has been proposed [3] that the reason squirrels can have the same genes as humans yet have the ability to undergo and endure hibernation is because of differential gene expression; something or some situation triggers the genes to turn on and off differently in the squirrel.

**Codon Bias**

One factor that could be important in gene expression is codon bias. In DNA, a codon is a group of three nucleotide bases that codes for an amino acid [5]. The four possible nucleotides are A-Adenine, T-Thymine, C-Cytosine, and G-Guanine. The double-helix of DNA is made up of two strings of these nucleotides. The order in which the nucleotides appear is not random, as they are read in groups of three (codons). Each codon codes for one amino acid. However, there are 64 codons (3 of which signal a strand to stop coding for a protein and don’t actually code for an amino acid) and only 20 amino acids so most have more than one associated codon (see figures).

When many of these codons are strung together, they form long DNA chains (some as long as 73 million codons long) that contain all the genetic information in our body. Codon bias occurs when one of these codons is preferred by the gene over the others. There are many biological reasons debated for this happening, such as gene expression level, G-C composition of the gene, RNA stability, or transcriptional selection [6].

Codon bias was recently found to have a profound impact on the role of a protein by Ivana and Michael [7]. They found that changing one single codon in two otherwise 98% identical actin proteins changed their properties and allowed the proteins to exist in two different structural parts of a cell. The key finding was that this change of codon occurred within the first 18 codons and affected the speed of

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
</tr>
<tr>
<td>Aspartate</td>
<td>Asp</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
</tr>
<tr>
<td>Glutamate</td>
<td>Glu</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
</tr>
<tr>
<td>Selenocysteine</td>
<td>Sec (TGA)</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
</tr>
</tbody>
</table>

Table: List of all amino acids and three-letter and one-letter abbreviations, including selenocysteine[7]

Figure: The genetic code. Names of amino acids and chain termination codons are on the periphery of the circle. The first base of the codon is identified in the center ring; the second base of the codon is in the middle ring; and the third base(s) of the codon is in the outer ring of the circle. http://www.currentprotocols.com/protocol/mba01c
translation enough to change the way the ribosome folds. This causes it to engage in different pathways of modification and results in two very different roles [7]. The following research was designed to determine if an explanation of differentiated codon usage can shed light on the squirrel’s unique bodily functions during hibernation. Is it possible that codon bias aids the highly expressed squirrel genes during hibernation? This paper will show that this is not the case.

**Homopolymeric Tracts**

Codon bias could play a role in another area of the squirrel’s genetic code. Namely, the second topic studied here; the avoidance of homopolymeric tracts in the genetic code. A homopolymeric tract is a segment of one strand of DNA that includes repeats of the same dinucleotide [9]. For example, notice the following string of codons contains a section of five Alanine (A): TGAAAAACTGTC.

Previous studies have shown that the occurrence of the A-tracts of length greater than five, as well as G-tracts of length greater than two, is rare in certain regions of DNA in five commonly studied organisms [9]. Five species, including *S. cerevisiae* (fungi), *Arabidopsis thaliana* (plants), *Caenorhabditis elegans* (invertebrates), *Homo sapiens* (mammals), and *Danio rerio* (non-mammal vertebrates), were studied and all were found to universally avoid long A-tracts in the exon (or coding) regions of DNA [9]. Since the findings were unanimous across all five species, determining if squirrel DNA follows that same rule could help eliminate one possible explanation of hibernation differences.
Part I: Gene Expression and Codon Bias

Methods: MILC

Due to the complexity of DNA sequences, a plethora of methods exist for quantifying codon usage in genes. Many are designed by individual researchers to meet their specific needs and are not usable in all circumstances. In choosing a method for this project, it was important to find one that was minimally biased and that could be used universally and with varying lengths of sequences. Fran Supek and Kristian Vlahovicek developed a unique method that sought to eliminate many of the context-dependent issues of other measures, and primarily create a method that is not affected by sequence length [10]. It is called the Measure Independent of Length and Composition, or MILC [10]. This method was shown to be resistant to overall composition of nucleotides as well as gene length. This measure is based on a log-likelihood score and gives each individual gene a score that compares it to the expected distribution of the genes that are expressed at the same time. The MILC formula is described later on in the paper as it is important to first understand the data that was used.

Raw Data

Looking at the genome of the 13-lined ground squirrel at the DNA level has, in recent years, been made simpler by techniques such as 454 Life Science Sequencing. 454 Sequencing is a method of sequencing DNA that was purchased by the Roche company in 2007 [11]. It is capable of producing 400-600 million base pairs per run of a sample with read lengths of 400-500 base pairs [11]. Because the techniques used by Roche are very new technologically, generating data for the 13-lined ground squirrel has never been done in this magnitude before. The data from Dr. Andrews’ lab returned from the Roche 454 Company with over 3.7 million “reads” of sequences. It is helpful to know that a “read” is an instance of a given gene that was, at the time of the sampling, turned on. A gene being turned on implies it has been activated by the cell and is in use. This is precisely the information that is necessary to do useful study of the squirrel DNA during hibernation.
The samples sent to Roche 454 were taken from 6 squirrels, three female and three male, at each of six time-points for each of three tissues (see table at right). The time-points were chosen representative of very different physiological stages for the squirrel. The three tissues sampled have very different properties, functions and needs, helping make the data set expansive and impressively covering the squirrel’s genetic information.

Dr. Marshall Hampton took all 3.7 million reads and grouped them into contigs, “sets of overlapping DNA segments derived from a single genetic source” [12]. Contigs help the identification process for genes since they provide a much longer sequence than the shorter ones obtained from the 454 data, making identification easier. Once genes were identified and mapped to a named human gene, they were cut down to only contain the coding domain portion of each contig. In the end, the 3.7 million reads were condensed down to 6,351 sequences that code for proteins. The sequences range in length from 200-20,000 base pairs approximately.

### Background Distributions and Control Sequences

Expected frequencies of codons were then calculated using the entire data set of 6,351 genes. Since much of the intended analysis consisted of comparing one gene, or a group of genes, to the entire data set this was a necessary step. Then, three separate background distributions were calculated. Creating the distributions required the ability to store large amounts of data. Sage was relied upon to do the calculations. Each background distribution was eventually used to create control sequences, thus they were given names that implied their future use. The distributions are called CS1- Codon, CS2 - Uniform, and CS3 - Letter (CS for Control Sequence). Only CS1 – Codon is necessary for the first part of the project and will be discussed here. The others will be described in greater detail later in the project.
Calculating CS1 – Codon

Since DNA is read in codons, it makes sense to form a background distribution for our data that reflects this. Determining how likely each codon is to appear in a given sequence is dependent upon the protein sequence those codons represent. Take this random example of a short protein sequence and notice the two Phenylalanine (F): PFSDMTKFEEYMR. When we see these, we know they were obtained by one of the two codons: TTT or TTC (See previous figure). Therefore, to calculate a background distribution based on codon usage we first must determine the number of occurrences of each amino acid in a protein sequence. Then within each amino acid, it is possible to calculate the frequency of each codon using its number of occurrences. Using the previous example of Phenylalanine codons TTT and TTC, the probability of seeing each in the entire data set of squirrel genes would be calculated as follows:

\[ P(TTT) = \frac{\#TTT}{\#TTT + \#TTC} \quad P(TTC) = \frac{\#TTC}{\#TTT + \#TTC} \]

Probabilities for the remaining codons were calculated similarly. (Of the 64 possible codons, 3 are used as stop codons for a coding region and do not code for an amino acid (ochre, amber, UGA). These were removed from the end of the sequences, as they do not affect protein creation. One of these codons, TGA, at times occurs within sequences and codes for Selenocysteine (U), thus a total of 62 codon probabilities were calculated). This background distribution is expected to be the most similar to the actual codon usage of the data set. A bar chart that shows the frequencies for all three of the calculated background distributions can be found at the beginning of the second part of this project on page 26.

MILC Calculation

The codon frequencies used to calculate CS1 were used as the expected distribution of codons for the entire data set. This is an imperative first step to have completed in order to have a basis to determine if any particular gene’s codon usage is unique and therefore interesting for biological reasons. The MILC statistic provides a score that quantifies the distance any particular gene is from the expected distribution [10].
Supek and Vlahovicek compared many methods for quantifying codon usage in genes, and in their 2005 paper introduced a new measure called the Measure Independent of Length and Composition, MILC \[10\]. This new method was designed to be universally useful as it corrected for some of the negative issues of previous methods. The two main issues the MILC score addresses are noted in the name itself. It attempts to quantify codon usage free of dependence on sequence length and “codon bias discrepancy”, or composition, in the sequences \[10\].

Their formula takes a basic log-likelihood ratio score and incorporates correction factors to account for changing sequence length and the presence of codon bias \[10\]. This was a major factor in the selection of this method for this project. The method was shown by Supek and Vlahovicek to be a “generally applicable method;” it suits the purposes of this project well.

**The MILC Formula**

The scores generated from the MILC formula will be a descriptor of how different a particular gene’s codon usage is from the background distribution that includes all of the genes found in the squirrel. Low scores indicate the particular sequence’s usage of codons varies little to the usage of the entire set, while high scores indicate sequences with codon usage that differs most from the entire group of genes.

The complete formula for finding the MILC (Measure Independent of Length and Composition) score for a given sequence is \[10\]:

$$MILC_{sequence} = \frac{\sum_{a} \left[ 2 \sum_{c} O_{c} \ln \frac{f_{c}}{g_{c}} - (r_{a} - 1) \right]}{L} + 0.5$$

In the formula, \(a\) denotes amino acid and \(c\) codon. The symbol \(O_{c}\) represents the observed number of counts of an individual codon within the sequence. The natural log term includes \(f_{c}\) and \(g_{c}\) denoting the observed frequency and expected frequency of the codon respectively. The length of the sequence in
codons is designated by the $L$ in the denominator while $r_a$ denotes the number of codons that code for amino acid $a$ [10].

As previously mentioned, the MILC score is designed to account for differences in the lengths of the sequences being studied. Broken down, the formula is more easily studied [10]:

$$
MILC = \frac{\sum_a M_a}{L} - C
$$

where $M_a = 2 \sum_c O_c \ln \frac{f_c}{g_c}$ and $C = \frac{\sum_a (r_a - 1)}{L} - 0.5$

The log likelihood-based score ($M_a$) that is the basis of the MILC causes problems when used on its own with DNA sequences. Due to varying lengths of sequences and the dependence of codon usage on the presence of certain amino acids, the absence of a codon affects the ratio and a correction factor ($C$) is necessary. Supek and Vlahovicek found that individual sequence scores varied from the expected scores by approximately $r_a - 1$ due to small sample sizes and sampling errors.

The amount of calculated bias was elevated in smaller sequences due to the absence of codons. Dividing, then, by $L$ (the length of the sequence in codons) levels the playing field for the expecting increased score for larger sequences. The correction factor, while necessary to correctly compare gene sequences of different lengths, caused the MILC score to be negative when the observed and expected frequencies are similar. Supek and Vlahovicek added the 0.5 to the correction factor to make all final scores positive for easier comparison [10]. Since the importance of the score is its relative distance from the other scores and from the average and not its distance from zero, this addition does not compromise the scoring process.
MILC Results

With the scoring method chosen, the next necessary step was using Sage and its biological capabilities to find our scores. Using the data files created by Dr. Hampton, a series of code was created to organize the data and calculate a score for each individual sequence. This process was made challenging by the fact that each gene sequence had to not only have a score, but also maintain a record of its biological data for future analysis. The following graph shows the approximately 6,300 genes with their MILC score on the y-axis and the number order in which the particular score was calculated on the x-axis:
The descriptive statistics for the MILC scores are given in the following table, along with the human gene that the maximum and minimum score are attached to:

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Score</th>
<th>Matched to Human Gene</th>
<th>Biological Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>1.6401</td>
<td>H32_HUMAN</td>
<td>Histone</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.3977</td>
<td>SYJ2B_HUMAN</td>
<td>Mitochondrial Outer Membrane</td>
</tr>
<tr>
<td>Mean</td>
<td>0.6560</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std. Dev. =</td>
<td>0.1399</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As a way to better see the range of the data visually, the graph below shows the ordered scores for the entire data set with lines showing standard deviation bars out to three deviations.

The following graph is a histogram of the MILC scores that gives us a general shape of their distribution.

The discussion following will use this data in combination with biological counts to attempt to show variations in codon usage.
The above chart shows the five genes with the lowest MILC scores, and the number of reads of that gene at each of the time points. A low MILC score means that the codon usage of these particular genes differs least from the entire data set of genes. Two of the genes that mention Mitochondria have higher reads in the heart, which makes sense because heart cells contain a lot of mitochondria [16]. Otherwise, the genes have uninterestingly similar read values.

<table>
<thead>
<tr>
<th>MILC</th>
<th>Uniprot Gene Name: Function or placement in cell</th>
<th>Tissue, Time point, and Number of Reads at that time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3977</td>
<td>SYJ2B_HUMAN: Mitochondrial Membrane</td>
<td>7 5 3 10 6 8 26 30 27 31 21 21 11 14 4 8 2 7 10</td>
</tr>
<tr>
<td>0.4305</td>
<td>TMM97_HUMAN: Regulates cellular cholesterol homeostasis</td>
<td>0 4 0 0 3 2 5 4 5 3 0 7 4 2 6 3 4 4 5</td>
</tr>
<tr>
<td>0.4392</td>
<td>BL1S2_HUMAN: Cell proliferation</td>
<td>3 3 2 0 0 1 2 2 0 3 4 9 1 2 3 8 5 1 0</td>
</tr>
<tr>
<td>0.4422</td>
<td>RM18_HUMAN: Transport in Mitochondria</td>
<td>12 11 3 5 5 2 11 8 12 23 9 13 7 5 4 2 4 3 1</td>
</tr>
<tr>
<td>0.4468</td>
<td>RN121_HUMAN: Transmembrane</td>
<td>5 7 3 3 0 2 2 1 4 7 0 2 1 3 1 2 2 1 2</td>
</tr>
<tr>
<td>0.4496</td>
<td>CL045_HUMAN: Phosphoprotein</td>
<td>0 0 0 2 0 2 4 3 4 3 2 0 1 3 1 3 4 4 5</td>
</tr>
</tbody>
</table>
The above chart shows the five genes with the highest MILC scores, and the number of reads of that gene at each of the time points. A high MILC score means that the codon usage of these particular genes differs most from that of the entire data set of genes. Most of the high scoring genes have uninteresting read patterns where the gene is either present everywhere or scarce everywhere. The NR2F6 gene has higher numbers in the Heart during IBA and March, which could be an interesting point. Another note-worthy gene could be the SODE gene, with an MILC score of 1.4520. Upon further inspection, the gene is not seen in skeletal tissue, rarely seen in heart tissue, but is especially prevalent in WAT in October, Torpor and IBA. Further study of this gene could prove interesting for biologists, but is beyond the scope of this project.
MILC Scores: Biological Connections

At this point, it is necessary to step back and gain perspective on the goal of this part of the project and insert additional biological explanation. So far, a score has been found for each gene sequence that describes its codon usage variation from the entire group. To make the scores more biologically useful it is helpful to consider them in conjunction with the number of reads used to generate the sequence. The biological reasoning behind this is that at any given time point, for example during an IBA (Inter-bout arousal), the genes that have the most number of reads in our sample were being used the most often at that time. These IBAs are one of the most interesting times to study because the squirrel wakes up briefly and its body temperature and heart rate zoom back up close to normal again for a short time before it slips back in to the torpor stage and its temperature and heart rate take a dive. This “heating up” process happens very quickly, and the hypothesis of this project is that in order for that to happen efficiently, squirrel genes are biased in the codons they choose to use.

For the purposes of this project, a gene is considered highly expressed if its read counts are at least one standard deviation higher than the average for that time point. If codon bias plays a part in the squirrel’s genetics during hibernation, we hope to see a difference between the distribution of scores of the highly expressed group of genes compared to the distribution of scores for the entire set.

MILC and Kolmogorov-Smirnov

Determining the distribution of the DNA data is complicated. The large data set we want to compare includes all of the values in the smaller, highly expressed data set. There is no known distribution for the use of codons in 13-lined ground squirrel DNA so while the plot of the data appears similar to a gamma distribution, attempting to fit it is not a good option. The method used to compare the two data sets is the two-sample Kolmogorov-Smirnov (KS) test. This gives a KS-statistic measuring the largest vertical distance between the two EDFs. The null hypothesis in this test is that the distributions are the same, with
the desired alternate hypothesis being that the distributions are unequal. The KS-statistic for a two-sided test of two samples, sizes \( m \) and \( n \), with empirical distribution functions \( F_1 \) and \( F_2 \) is \([13][14]\):

\[
D_{m,n} = \sup_x |F_{1,m}(x) - F_{2,n}(x)|
\]

We reject the null hypothesis if

\[
D_{m,n} > K_\alpha
\]

where \( K_\alpha \) is the critical value and

\[
K_\alpha = c(\alpha) \sqrt{\frac{m + n}{mn}}
\]

The coefficient \( c(\alpha) \) value of 1.36 was used for the calculations corresponding to \( \alpha = 0.05[13][14] \).

**MILC and KS Results**

The next step was to find the highly expressed genes at each time point and for each tissue using the cutoff of one standard deviation above. (Going one standard deviation below the mean was impossible since the chart moved to negative values in all cases) Using a tool created by Dr. Hampton in Sage, individual time point and tissue graphs showing the highly expressed genes in red and the rest of the genes in blue were created, and can be seen in the tables on the next several pages. The tables also show the histogram of the entire data set previously shown with the addition of a superimposed histogram of the highly expressed genes for each tissue/time point combination. The mean number of reads and standard deviation for the particular tissue/time point is also given on the table as well as the KS-test value and the \( K_\alpha \) critical value for each group. A discussion of the results for each tissue follows the individual tables.
<table>
<thead>
<tr>
<th>Skeletal Tissue</th>
<th>Plots of all reads for a given time point (scores above one s.d. in red) with number of reads on y-axis and scores on x-axis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>April</strong></td>
<td>Histogram of the scores for the highly expressed group (red) vs. the remainder of the data set (blue) for each tissue/time point with scores on the x-axis</td>
</tr>
<tr>
<td>Mean = 9.9871</td>
<td>$n_r$ = number of red  $n_b$ = number of blue  KS = Kolmogorov-Smirnoff statistic  $K_a = $ Critical value for KS statistic</td>
</tr>
<tr>
<td>Std = 41.0473</td>
<td>$n_r$ = 186  $n_b$ = 6070  KS = 0.0620  $K_a$ = 0.1012</td>
</tr>
<tr>
<td><strong>August</strong></td>
<td>$n_r$ = 202  $n_b$ = 6054  KS = 0.0881  $K_a$ = 0.0973</td>
</tr>
<tr>
<td>Mean = 8.8983</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td>Std = 35.6074</td>
<td>$n_r$ = 246  $n_b$ = 6007  KS = 0.0616  $K_a$ = 0.0884</td>
</tr>
<tr>
<td><strong>October</strong></td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td>Mean = 10.1090</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td>Std = 40.8920</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td><strong>Torpor</strong></td>
<td>$n_r$ = 246  $n_b$ = 6007  KS = 0.0616  $K_a$ = 0.0884</td>
</tr>
<tr>
<td>Mean = 8.7734</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td>Std = 30.9999</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td><strong>April</strong></td>
<td>$n_r$ = 186  $n_b$ = 6070  KS = 0.0620  $K_a$ = 0.1012</td>
</tr>
<tr>
<td>Mean = 9.9871</td>
<td>$n_r$ = 202  $n_b$ = 6054  KS = 0.0881  $K_a$ = 0.0973</td>
</tr>
<tr>
<td>Std = 41.0473</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td><strong>August</strong></td>
<td>$n_r$ = 202  $n_b$ = 6054  KS = 0.0881  $K_a$ = 0.0973</td>
</tr>
<tr>
<td>Mean = 8.8983</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td>Std = 35.6074</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td><strong>October</strong></td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td>Mean = 10.1090</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td>Std = 40.8920</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td><strong>Torpor</strong></td>
<td>$n_r$ = 246  $n_b$ = 6007  KS = 0.0616  $K_a$ = 0.0884</td>
</tr>
<tr>
<td>Mean = 8.7734</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
<tr>
<td>Std = 30.9999</td>
<td>$n_r$ = 200  $n_b$ = 6062  KS = 0.0900  $K_a$ = 0.0977</td>
</tr>
</tbody>
</table>

**Note:** The table and graphs show the distribution of read scores for different time points and tissues, with statistical measures such as mean, standard deviation, and Kolmogorov-Smirnoff test statistics.
### IBA
Mean = 8.2187  
Std = 30.1062  
\( n_r = 230 \)  
\( n_b = 6032 \)  
KS = 0.0398  
\( K_\alpha = 0.0913 \)

### March
Mean = 8.8049  
Std = 34.4648  
\( n_r = 190 \)  
\( n_b = 6071 \)  
KS = 0.0894  
\( K_\alpha = 0.1002 \)

#### Skeletal Muscle Discussion

For skeletal muscle, the KS-statistics are all smaller than their critical values telling us to not to reject our null hypothesis that the two data sets are from the same distribution. In all cases, the genes with the highest number of reads have scores that are distributed the same as the rest of the data set. This would lead to the conclusion that the differentiated codon bias does not have an impact on expression level at any time point in the skeletal muscles of squirrels.
Heart Tissue

Plots of all reads for a given time point (scores above one s.d. in red) with number of reads on y-axis and scores on x-axis

Histogram of the scores for the highly expressed group (red) vs. the remainder of the data set (blue) for each tissue/time point with scores on the x-axis

\[ n_b = \text{number of blue} \quad n_r = \text{number of red} \quad \text{KS}=\text{Kolmogorov-Smirnoff statistic} \]

\[ K_{\alpha} = \text{Critical value for KS statistic} \]

<table>
<thead>
<tr>
<th>Month</th>
<th>Mean</th>
<th>Std</th>
<th>( n_r )</th>
<th>( n_b )</th>
<th>KS</th>
<th>( K_{\alpha} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>April</td>
<td>14.6530</td>
<td>56.6185</td>
<td>187</td>
<td>6074</td>
<td>0.0391</td>
<td>0.1009</td>
</tr>
<tr>
<td>August</td>
<td>13.8698</td>
<td>48.0435</td>
<td>204</td>
<td>6055</td>
<td>0.0549</td>
<td>0.0968</td>
</tr>
<tr>
<td>October</td>
<td>13.8391</td>
<td>51.3470</td>
<td>187</td>
<td>6071</td>
<td>0.0612</td>
<td>0.1009</td>
</tr>
<tr>
<td>Torpor</td>
<td>14.0564</td>
<td>48.5635</td>
<td>210</td>
<td>6047</td>
<td>0.0362</td>
<td>0.0954</td>
</tr>
</tbody>
</table>
Heart Discussion

As you can see from the plots of the heart data, the data appear to be distributed very similarly to that of the skeletal data. The mean number of reads is slightly higher in all of the heart time points and many of the standard deviations are larger than in skeletal muscle, but the results are comparable. The KS-statistics for the heart time points are also unanimous in telling us not to reject the null hypothesis that the data come from the same distribution. Again, the histograms look remarkably similar to each other and leave us without verifying our hypothesis. The expression levels of genes in the heart are also not greatly affected by codon bias.
| WAT (White Adipose) Tissue | Plots of all reads for a given time point (scores above one s.d. in red) with number of reads on y-axis and scores on x-axis
| | Histogram of the scores for the highly expressed group (red) vs. the remainder of the data set (blue) for each tissue/time point with scores on the x-axis
| | \( n_r = \text{number of red} \quad n_b = \text{number of blue} \)
| | KS=Kolmogorov-Smirnoff statistic
| | \( K_a = \text{Critical value for KS statistic} \)
| **April** | Mean = 10.5937
| | Std = 41.5641
| | \( n_r = 140 \)
| | \( n_b = 6122 \)
| | KS=0.0957
| | \( K_a = 0.1162 \)
| **August** | Mean = 12.6204
| | Std = 125.1953
| | \( n_r = 39 \)
| | \( n_b = 6217 \)
| | KS=0.0842
| | \( K_a = 0.2185 \)
| **October** | Mean = 13.4360
| | Std = 111.6766
| | \( n_r = 56 \)
| | \( n_b = 6206 \)
| | KS=0.1436
| | \( K_a = 0.1825 \)
### WAT Discussion

The third set of tissues gives us results remarkably similar to those of the first two tissues. Even with varying standard deviation values and thus sample sizes, the graphs of WAT are very similar to those of both skeletal tissue and the heart. The KS-statistics here are also all less than the critical values telling us there is not enough evidence to suggest that the distributions are different. Again in WAT we have codon bias having very little effect on the expression levels at any of the time points.
Codon Bias/MILC Conclusions

The hope in analyzing the squirrel DNA using the MILC scores was to be able to show that codon bias aids in the ability of genes to be highly expressed at certain time-points. It is known that codon bias exists and allows for all sorts of differentiation in genes, but the findings of this project are that actual codon usage in highly expressed squirrel genes does not vary from codon usage in the minimally expressed genes at any given time point. The findings do not support codon bias as an influencing factor for aiding hibernation. It is, however, interesting to note the similarity among the scores from each of the tissues and time points. The data from this project supports the fact that genes and DNA work efficiently in the bodies of squirrels, but their codon usage is not necessarily the cause of this efficiency.
Part II: Homopolymeric Tracts

Materials and Methods

A second topic investigated in this project is the avoidance of lengthy homopolymeric tracts in the genetic code. As mentioned in the introduction, previous research has found that both the genetic code and the nucleosome positioning code coexist in the exon, or coding, regions of DNA. These particular parts of the coding regions, because of this dual functionality, have a strong codon bias that allows for the avoidance of long A-tracts.

As previously stated, the data used in this project was all taken from protein coding regions of the squirrel DNA. This makes it the perfect data set to attempt to verify the avoidance of not just long A-tracts, but G-tracts as well in squirrel genes. The intention is to calculate and compare the statistics around the squirrel DNA to that of accessible human data and cross reference it with the results presented in Cohanim and Haran’s paper. There, it was found that only the A and G-tracts are of any significance when looking at them as homopolymers.

In order to use the MILC formula, expected frequencies for each of the codons were found from the given data. To create the CS1 that was named in the first part of this project, the expected frequencies were then used in Sage to create random sequences to use to compare our original sequences to. In fact, for each original sequence, one hundred control sequences were created, stored and used for later calculations. This was repeated for the CS2 and CS3 described in the next section.

Calculating CS2 – Uniform

A second background distribution was created assuming all codons are uniformly distributed within each amino acid. It is known that this is not an accurate assumption to make as codon bias does exist in DNA, but a uniform distribution will aid in the comparisons of results[6]. Returning to the Phenylalanine example, it is easy to see that under an assumption of uniform codon distribution, we would have $P(TTT) = 0.5$ and $P(TTC) = 0.5$. As in CS1, this calculation was repeated for all 62 codons. While this may be
the easiest background distribution in terms of calculation, we expect it to differ from CS1 and the actual codon bias. It will be useful to look for codon usage that is different from uniform (thus verifying previous knowledge).

**Calculating CS3 – Letter**

Again, for the sake of comparison, a third background distribution was calculated. Since DNA is read in groups of three, any distribution that did not take this into consideration would widely miss the mark in terms of defining a DNA data set. However, since DNA is made up of only the four bases, A, T, C, and G, it is not completely faulty to consider a background distribution that assumes uniform base distribution. For CS3, The frequency of each base was calculated for the entire set of genes, independent of which amino acid was present. The results were 'A': 0.256, 'C': 0.257, 'T': 0.220, and 'G': 0.266. The probability of each codon appearing was then determined using these frequencies. For our Phenylalanine example, we would have:

\[
P(TTT) = \frac{f(T)f(T)f(T)}{f(T)f(T)f(C) + f(T)f(T)f(T)} \quad P(TTC) = \frac{f(T)f(T)f(C)}{f(T)f(T)f(C) + f(T)f(T)f(T)}
\]

The following histogram shows the frequencies for each of the calculated sets of control sequences. As noted below, the red bars represent our CS1, or the distribution closest to the actual codon usage in our genes, while the yellow is CS2 (the uniform distribution) and the blue shows CS3 (the distribution calculated from the letter frequencies). The bars represent the distribution of the codons within each amino acid. To continue with our Phenylalanine example, locating its symbol, F, on the x-axis shows us that \(TTC\) (CS1=.5400 and CS3=.5388) is more commonly used than \(TTT\) (CS1=.4600 and CS3=.4612). While in this case, the result for CS1 and CS3 appear similar to each other, and very different from the uniform frequencies of \(TTT=TTC=.5\), a glance at any other group of codons show that this trend is not consistent. One example of this is Leucine (L) and its 6 codons. From the histogram, we see that CS2 and CS3 are more similar to each other while the CS1 distribution is much more varied. This plot is
meant to provide a visual representation of the differences and similarities in the 3 different background distributions.

There are a few amino acids whose codon distribution looks interesting when compared to CS2 and CS3. Glutamine (Q), for example, has only two codons that code for it and according to the CS data CAA is much more rarely used than CAG in the squirrel genes. Leucine (L) is another place where one codon is obviously much preferred over the others. CTG is used 40% of the time while the other five codons used for Leucine are each used less than 20% of the time. Interestingly, three of the six Leucine codons have dinucleotide repeats as does the less used codon in Glutamine. This may not be significant biologically, but it is interesting to note as an attempt is made to verify that the squirrel genes avoid long homopolymeric tracts.

**Counting Homopolymers**

Once the background distributions were calculated, control sequences were needed to compare our data sequences to. To create these, several intermediate steps were necessary. Each individual sequence from the data set had to be first translated into an amino acid chain. To bring back our earlier example, all TTC and TTT codons (Phenylalanine) from an original sequence became an F in the amino acid chain and so on for all codons in the sequence (Ex: FMRKAFVGSILVIMDEFFAVGF). From this amino acid chain,
there were then one hundred random sequences created based on the frequencies of the codons in each of the three background distributions. So an individual sequence from the original data set had a total of three hundred control sequences created for it based on the codons (CS1), uniform (CS2), and on the distribution of the bases (CS3).

The next task was to count the occurrence of homopolymers of length 2-9 in each of these over 1.8 million sequences. For a given squirrel sequence, the actual counts of the homopolymers was stored and used for comparison. As each set of control sequences was created, the homopolymers of length 2 – 9 were counted and then averaged and stored to be compared to the original counts. The work of Cohanim and Haran used a Log ratio to compare the original (“natural”) sequence data to the control series data so, to be consistent, the comparison used in this project shows the Log ratio of the data as well [9]. The results will be shown in graph form using the following formula: \( \log \left( \frac{f_{nat}}{f_{CS}} \right) \).

where \( f_{nat} \) is the frequency of the homopolymer in the original data sequence and \( f_{CS} \) is the frequency of the homopolymer in the control sequence(s).

**Results and Discussion**

In each of the data graphs that follow, the x-axis shows the length of the homopolymer (tract) and the y-axis shows the Log score. The first three colored bars at any given x value represent the squirrel data compared to CS1, CS2 and CS3 and the second set of colored bars represent human data compared to CS1, CS2 and CS3. The comparison with human data was done in order to verify that the process used would give results consistent to that in the literature [9]. For this purposes of this project, only up to 8 bp of a repeated nucleotide are shown on the graphs of the results as opposed to 10 bp on the graphs from the comparison study [9].
**G-tracts**

The Log scores for the human G-tracts are consistent with the scores given in the paper. It also appears that the squirrel genes follow in the same pattern of the other species from the paper and after five repeated base pairs, the G-tracts become less common in the natural sequences than in the control sequences. This was the expected result.

![Diagram](image.png)

*Squirrel – first three bars after each number (pink group) in order of CS1, CS2, and CS3
Human – last three bars after each number (brown group) in order of CS1, CS2, and CS3*
A-Tracts

The results for A-tracts were also similar to those in the paper. As seen on the following graph, the human data (in orange) was corroborated by this project with similar Log scores and the squirrel data agrees with that of all species in that A-tract avoidance is prevalent. Looking at the darkest blue bars (corresponding to CS1), the natural squirrel sequences become more rare than the control sequences after 4 bp. The project data shows the squirrel genes are slightly more apt to avoid long A-tracts than the human data. This is also in agreement with the other species in the paper. It appears that in both squirrel and human genes, avoidance of homopolyer A-tracts in natural sequences is most different from the CS3 control sequences. These used the frequency of each letter to generate random sequences and were expected to differ most from the actual frequencies.
**T-Tracts and C-Tracts**

Neither T- nor C-tracts were significantly different enough to warrant the inclusion of plots in the literature [9]. Interestingly, the results from this project question whether, for the squirrel, all homopolymers are rare after 5 bp. From the graphs below, the T-tract results look similar to the A-tract results reported earlier and the G-tract and C-tract results look similar. This makes sense since A and T pair up in the double helix of DNA as well as G and C so we would expect to see similar behaviors of their repeats. Both graphs show the homopolymers of increasing length to be slightly less rare in the squirrel than in the human. The results also show that the C-tracts, more than any of the others, are actually more common in the natural sequences than in the control sequences until the repeated letters reach 5 bp. While research shows that A-tracts especially are avoided in parts of the coding region of DNA, the results shown here seem to say that C-tract may be interesting to study as well [9].
Conclusions

This project ends with the surprising results that my initial hypothesis was not correct. While codon bias does exist in squirrel genes, the bias cannot explain the high expression levels of different genes at time-points important to hibernation. The codon usage of these highly expressed genes mirrors that of the entire data set of genes. While this is not the interesting result hoped for, what the result DOES say about squirrel genes is that they are remarkably efficient with their codon usage. A similar story can be told of the search for homopolymers in squirrel exon regions. This project shows that the presence of these long tracts in squirrel DNA mimics that of the other species studied. Perhaps a future study of the specific genes that contain long C-tracts could be interesting. Looking into both the highly expressed groups of genes as well as the groups with exceptionally long homopolymers would be a possible extension of the work done here.
References


2. Spermophilus tridecemlineatus: thirteen-lined ground squirrel.

3. UMD Biology Department. Matt Andrews.


16. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists

17. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the
    [PubMed]
