HW2

Philipp Ross

2017-01-23

Last updated: 2017-01-24

Code version: 6805621

Problem 1

For this problem we know the following:

- Population size = 1000
- Haploid genome size = 10^9
- Mutation rate per organism per generation = $\frac{1}{10^6}$
- Percent neutral mutations = 20%
- Percent deleterious mutations = 80%

Random mating and equal fitness.

(a)

The number of total mutations per generation = $2e10^9 \cdot \frac{1}{10^6} \cdot 10^3 = 2e10^6$. We get this by multiplying the haploid genome size by 2 since we're looking at diploid individuals, multipying by the rate of mutation per genome, and finally by the population size. The number of neutral mutations would be $0.2 \cdot 2e10^6 = 4e10^5$ since 20% of the total number of mutations are neutral.

(b)

We can calculate this by multiplying the neutral rate of substitutions per locus (which is the same as the locus mutation rate) by the probability of fixation by the number of generations. Then we get the number of fixed sequence differences = $2e10^5 \cdot \frac{1}{10^3} \cdot 10^3 = 2e10^5$ haploid sequence differences.

(c)

The fraction of nucleotide sites in an average individual expected to tbe heterozygous is 0.11.

(d)

The probability is $\frac{1}{2N} = \frac{1}{2000}$

(e)

The probaility that a new beneficial allele with selection coefficient $s = 0.01$ is defined as $P_{fix} = \frac{1-e^{-2s}}{1-e^{-4Ns}}$

Where $P_{fix} = 0.0198013$. Thus, this allele is most likely going to be removed from the population. The probility of extinction is just one minus the previous result or 0.9801987.

Problem 2

First let's define some functions:

```
seq1 <- "AAGCCGCCTTTCTTATGGTACTA"
seq2 <- "AAACCACCTTACTAAGGGTGCTA"

# Poisson evolutionary distance correction
PoisK <- function(D) {
```

```
    return(-log(1 - D))
}

# Jukes-Cantor evolutionary distance correction
JCK <- function(D, s) {
    return(-((s - 1)/s) * log(1 - (s/(s - 1)) * D))
}
```

(a)

The amino acid sequences for both nucleotide sequences are:

```
# for seq1
paste(seqinr::translate(seqinr::s2c(seq1)), collapse = "-")
```

```
[1] "K-P-P-F-L-W-Y"
```

```
# for seq2
paste(seqinr::translate(seqinr::s2c(seq2)), collapse = "-")
```

```
[1] "K-P-P-Y-*-G-C"
```

```
# Observed differences
d <- 3/8
# Calculate K
PoisK(d)
```

```
[1] 0.4700036
```

The statistic K corrects for the evolutionary distance based on the assumption that the probability of mutation at any site is independent and follows a poisson distribution. This accounts for convergence between two sequences and reversal of substitutions.

(b)

```
# calculate D for nucleotides
d <- 6/24
PoisK(d)
```

```
[1] 0.2876821
```

```
JCK(d, s = 4)
```

```
[1] 0.3040988
```

The values here are different because the equations to correct for evolutionary distance are different. The Jukes-Cantor model takes into account what happens over long periods of time. That is, once enough time has passed, two sequences of a certain length with a certain number of states per site will contain a certain number of similarities just by chance. Another way of thinking about it is if you were to align two completely random nucleotide sequences of the same length with the same nucleotide disitributions, just by chance, they would be 25% similary to one another.

(c)

We can calculate the following for seq1:

- Number of synonymous sites = 6
- Number of nonsynonymous sites = 18

Next we can calculate the following:

- $D_a = \frac{3}{18}$
- $D_s = \frac{3}{6}$

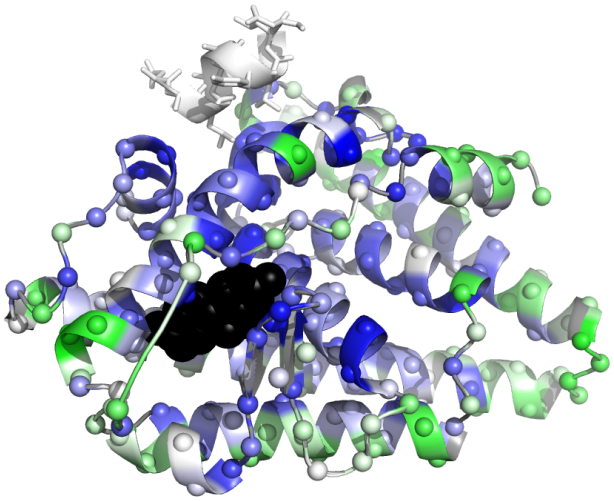
- $K_a = 0.1832585$
- $K_s = 0.7098537$
- JC-corrected $\frac{K_a}{K_s} ratio = 0.2581637$

These data are not consistent with the hypothesis that the sequences were evolving by drift alone. A deviation from 1 in this ratio indicates that selection is acting on these sequences. In this case, since the ratio is significantly less than 1, we assume it's puryfing selection.

(e)

These data are consistent with the Neutral Theory of molecular evolution as a much greater proportion of the fixed mutations come from synonymous sites as opposed to nonsynonymous sites.

Problem 3



We can see in the above structure that sites near the ligand and within the core of the molecule (i.e. those not exposed to the surface) are highly conserved, while most residues exposed to the surface are highly variable. Alpha helices packed in the core of the molecule tend to be highly conserved whereas those exposed to the surface of the moecule tend to be more highly variable. Residues near beta sheets tend to be highly conserved.

These data are consistent with the Neutral Theory as we can see from the large number of highly variable amino acid residues. This indicates that a large proportion of the mutations seen in this protein are netral mutations, while a smaller number mutations would be negatively selected for if they were to be within the core of the protein.

Problem 4

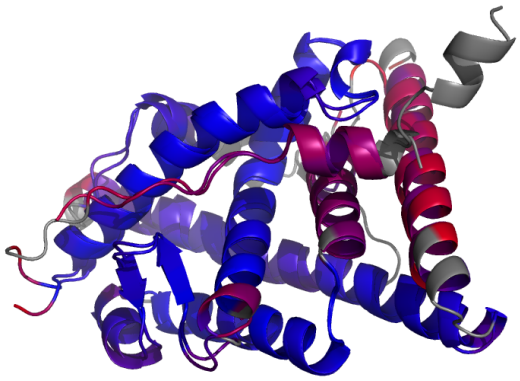
(a)

We included nitrogens, alpha carbons, and carbons while excluding oxygens and other atoms. We might exclude all others bceause they interfere with an effective alignment of the structures possibly due to their high degrees of freedom that allows the atoms to occupy different positions in space without providing structural information important for a comparison between two structures.

(b)

The RMSD for these atoms is 1.308. This is a measure of the quality of your alignment. The smaller the RMSD, the better your strucutral alignment. The goal is to minimize the RMSD between each pair of aligned atoms with the structures being aligned.

(c)



Colored and gray regions represent the distances between atoms. Blue is on the low end of the spectrum whereas red is on the high end, in terms of distance. Gray are perfectly aligned atoms. We can see a similar trend in regards to the regions of the protein that appear more structurally divergent from one another as we saw in the mutational profile of the alignment from question 3. That is, the structures that are packed within the core of the protein appear less structurally divergent, while those exposed to the surface appear more so.

(d)

I answered the last question before I read this one... but yes. I believe those structures that are less exposed to the surface and near the ligand binding site, are more structurally conserved based on the alignment within PyMol and we see similar things relative to unique amino acids at these sites. That is, we see a relatively low number of unique amino acids at these structurally conserved sites, suggesting that mutations in regions of high structural conservation will likely be deleterious to the function of the protein. This suggests a trend between mutationally variable versus conserved sites over evolutionary time and protein structure.