

## Lab 8

DNA sequencing is a method of determining the DNA sequence of a tissue sample (blood, hair, etc). At each position along the sequence the method determines whether the sample has an A, C, G or T by measuring the heights of peaks in a picture known as a “chromatogram”.

In humans, and most other large organisms, each individual has two copies of DNA at each position: one coming from the father, and one from the mother. Since the DNA of any two humans is typically very similar (differing at about 1 position in 1000), most samples will have two copies of the same base at most positions. Occasionally however, a sample will have inherited two different DNA bases from the father and the mother (eg an A from the father and a C from the mother). When this happens, the height of the peak in the chromatogram is about half as big as you would usually expect. These positions are called “heterozygotes”. Therefore we might hope to identify the heterozygotes by looking for heights that are particularly smaller than expected.

But what is expected? The expected height varies along the sequence, as we will see, which makes it more difficult to assess what is expected. Here we will try to regress height on position along the sequence to predict the expected height at any given position. The data set at

<http://www.stat.washington.edu/stephens/S423/DNAH1.txt>

gives data for one tissue sample.

1. Read the data file into R, and investigate how the height of the peak varies with position along the sequence. Does it look linear? Does the expected height also vary with the base (A,C,G or T?)

You might find the following color plot useful (assuming `d` is a dataframe with the data in it):

```
plot(d$position,d$height,col=as.numeric(d$base))
```

2. Ignore the base variable for the moment, and see if you can develop a regression to predict the height from position along the sequence. (Recall how in class we did an example where we let  $y$  depend on  $x$ , or powers of  $x$  up to 2, 3, 4, . . . . Try using a similar idea to fit a regression to predict height from position along the sequence.) Try several models, and see which seems to fit best. What does the AIC say? Try taking a log or square-root transformation of height, and see if it improves the behaviour of the residuals at all.

- Using your best model, try to identify some positions that might be heterozygotes. Can you quantify your certainty in whether the position is a heterozygote or not? (This second part is challenging.)
- Try adding “base” as a covariate to your best model. Note that “base” is not numeric, so R treats it as a “factor”, which means that each possible value (A, C, G or T in this case) is allowed to have a different mean. (We talked about this in class for the case of a 0-1 variable.) Try plotting the fitted values against position, for the different bases (ideally on one graph, in different colors). Here’s some R commands to get you started on this:

```
plot(d$position,d$height,col=as.numeric(d$base))
lines(d$position[d$base=="A"],d.lm$fitted[d$base=="A"],col=1)
lines(d$position[d$base=="C"],d.lm$fitted[d$base=="C"],col=2)
```

What do you notice? How does it differ from what you would get if you had fitted four regressions: one for the data on As, one for Cs, one for Gs and one for Ts?

- For those who are interested (you don’t need to hand anything in for this), the data on the secondary heights are also available (DNAH2.txt). These give the heights of the second biggest peak at each position. Those sites that are real heterozygotes should have both a smaller H1 than you would expect, and a bigger H2 than you would expect. It might be interesting to look at the points you identified as potential heterozygotes, to see if any of them also have a big value for H2.