Homework 10: Markov Chains and DNA

36-467/667, Fall 2020

Due at 6 pm on **Thursday, 12 November 2020**

The cellular slime mold *Dictyostelium discoideum* is, during most of its life-cycle, a single-celled amoeba, moving through the soil eating plant matter. When food grows scarce, however, these cells organize themselves into a multi-cellular body, with differentiated parts, which then crawls to the nearest available height and releases spores. When the spores land, they turn into new independent cells, beginning the cycle over again. *D. discoideum* is thus somewhere between a single-celled amoeba and a multi-cellular animal[1].

The data file `dicty-seq-1.dat`, derived from Eichinger *et al.* (2005), contains the genome for one chromosome of *D. discoideum*. This is a sequence of discrete values $X(t)$, where each $X(t)$ is either A, C, G or T, representing the four bases out of which DNA is assembled, or the wildcard symbol N, indicating variation in the DNA at that position.

Computing Note: The two data files here are quite large, and some parts of the assignment may take an excessive amount of time to run on many computers. It's OK to just use the first 100,000 observations in each data file, if you clearly note what you're doing.

1. (5) Load the data. How long is the sequence? How often does each of the four bases (A, C, G and T) occur in the data? *Hint:* `table()`.
2. (5) How often does each of the possible pairs of successive symbols occur? *Hint:* The following function may be helpful.

```
symbseq.to.blocks <- function(s, L) {
    n <- length(s)
    collapser <- function(i) {
        paste(s[i:(i + L - 1)], collapse = "")
    }
    max.index <- n - L + 1
    blocks <- sapply(1:max.index, collapser)
    return(blocks)
}
```

3. (10) Fit a first-order Markov chain to the data, and report the transition matrix.

[1] The best introduction to these fascinating, if not very cuddly, organisms is Bonner (2009).

4. (5) What is the log-likelihood of the data under your estimated Markov chain?

5. (10) Provide a 95% confidence interval for each entry in the transition matrix. (You can use any of the methods discussed in the lecture slides.) Do these intervals seem unusually large or small to you?

6. (5) Using your estimated transition matrix, find the invariant distribution of the Markov chain. How well does this agree with the distribution over the bases you found in Problem 1?

7. (10) Fit a second-order Markov model to the data. Report the transition matrix in the form of a table giving $P(X(t+1) = k|X(t-1) = i, X(t) = j)$ for all relevant combinations of i, j, k.

8. (2) What is the log-likelihood of the second-order Markov model?

9. (5) What is the p-value for a likelihood-ratio test of the null hypothesis of a first-order Markov model, against the alternative of a second-order Markov model? Report the test statistic, the number of degrees of freedom, the p-value, and your conclusion.

10. (5) The file `dicty-seq-2.dat` contains a second chromosome for *Dictyostelium*. Fit a first order Markov model to this chromosome, and report the transition matrix and 95% confidence intervals.

11. (3) What is sum of the log-likelihoods of the two chromosomes under their respective models?

12. (10) Fit *one* Markov model to *both* chromosomes. Report the transition matrix.

    Hint: You'll need to combine information from the two data sets somehow. There are multiple ways to do this. If you can't figure out a better way to do so, you *could* just concatenate one sequence on to the end of the other, but doing so will introduce a source of error, which you should explain if you take this route.

13. (4) What is the log-likelihood of the two chromosomes under the common, pooled model?

14. (10) Use a likelihood ratio test to whether the two chromosomes should be fit with the same Markov chain. Report the test statistics, the number of degrees of freedom, the p-value, and your conclusions. *Hint:* Think carefully about the number of degrees of freedom.

15. (1) How much time did you spend on this assignment?

Presentation Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized,

free of grammatical and other mechanical errors, and easy to follow. All plots, tables, etc., are generated automatically by code embedded in the R Markdown file. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant, without dangling or useless commands. All parts of all problems are answered with coherent sentences, and raw computer code or output are only shown when explicitly asked for.

References

Bonner, John Tyler (2009). *The Social Amoebae: The Biology of Cellular Slime Molds*. Princeton, New Jersey: Princeton University Press. URL `https://www.jstor.org/stable/j.ctt7s6qz`.

Eichinger, L., J. A. Pachebat, G. Glockner *et al.* (2005). "The genome of the social amoeba *Dictyostelium discoideum*." *Nature*, **435**: 43–57. doi:10.1038/nature03481.