Statistics 200B
Spring 1997

Lab IV: A Random Walk Model for Chromatin

A human chromosome is a very large molecule, about two or three centimeters long, containing of the order of 100 million base pairs (Mbp). The cell nucleus, where it is contained, is in contrast only about a thousandth of a centimeter in diameter. The chromosome is packed in a series of coils, called chromatin, in association with special proteins (histones), forming a string of microscopic beads. This mixture of DNA and protein constitutes the *chromatin*. In the G0/G1 phase of the cell cycle, between mitosis and the onset of DNA replication, the mitotic chromosomes diffuse into the interphase nucleus. At this stage a number of important processes related to chromosome function take place. For example, DNA is made accessible for transcription and is duplicated and repairs are made of DNA strand breaks. By the time of the next mitosis, the chromosomes have been duplicated. The complexity of these and other processes raises many questions about the large scale spatial organization of chromosomes and how this organization relates to cell function. Fundamentally, it is puzzling how these processes can unfold in such a spatially restricted environment.

At a scale of about 10^{-3} Mbp, the DNA forms a chromatin fiber about 30 nm in diameter; at a scale of about 10^{-1} Mbp the chromatin may form loops. Very little is known about the spatial organization beyond this scale. Various models have been proposed, ranging from highly random to highly organized, including irregularly folded fibers, giant loops, radial loop structures, systematic organization to make it readily accessible to transcription and replication machinery, and stochastic configurations based on random walk models for polymers.

A series of experiments (see references below) were conducted to learn more about spatial organization on larger scales. Pairs of small DNA sequences (size about 40 kbp) at specified locations on human chromosome four were flourescently labeled in a large number of cells. The distances between the members of these pairs were then determined by flourescence microscopy (the distances measured were actually two-dimensional distances between the projections of the paired locations onto a plane). The empirical distribution of these distances provides information about the nature of large scale organization.

There has long been a tradition in chemistry of modeling the configurations of polymers by the theory of random walks. As a consequence of such a model, the two dimensional distance should follow a Rayleigh distribution

$$f(r|\theta) = \frac{r}{\theta^2} \exp(-r^2/2\theta^2) \tag{1}$$

Basically, the reason for this is as follows: the random walk model implies that the joint distribution of the locations of the pair in R^3 is multivariate Gaussian; by properties of the multivariate Gaussian, the joint distribution of the locations of the projections onto a plane is bivariate Gaussian. As in Example A of Section 3.6.2 of the text, the distance between the points follows a Rayleigh distribution.

In this lab you will fit the Rayleigh distribution to some of the experimental results and examine the goodness of fit. The entire data set is comprised of 36 experiments in which the separation between the pairs of flourescently tagged locations ranged from 10 Mbp to 192 Mbp. In each such experimental condition, about 100-200 measurements of two dimensional distances were determined. This lab will just be concerned with the data from three experiments (short, medium, and long separation), but if you want to look at the data from some of the other experiments, the TA can tell you how to obtain it.

The assignment:

1. What are the maximum likelihood and method of moments estimates of θ? What is the asymptotic variance of the maximum likelihood estimate?

2. Estimate θ for each of the experiments. Does there appear to be any relationship between your estimates and the genomic separation of the points?

3. Estimate the asymptotic variances of your estimates.

4. For one of the experiments, compare the asymptotic variance to the result obtained from a parametric bootstrap. In order to do this, you

will have to generate random variables from a Rayleigh distribution with parameter θ.

Show that if X follows a Rayleigh distribution with $\theta = 1$, then $Y = \theta X$ follows a Rayleigh distribution with parameter θ. Thus it is sufficient to figure out how to generate random variables that are Rayleigh, $\theta = 1$. Show how Proposition D of Section 2.3 of the text can be applied to accomplish this.

Some care is needed when using Splus for bootstrapping. The basic problem is that Splus does not allocate and reclaim memory well, with the result that innocuous appearing calculations in loops can use very large amounts of memory. The TA will give you some tips on how to avoid catastrophes.

$B = 100$ bootstrap samples should suffice for this problem. Make a histogram of the values of the $\hat{\theta}^*$. Does the distribution appear roughly normal? Do you think that the large sample theory can be reasonably applied here? Compare the standard deviation calculated from the bootstrap to the estimate based on large sample theory.

5. Next examine the goodness of fit via histograms. For each of the three datasets, make a histogram of the data and superpose a curve showing the fitted density. Be careful of the scaling—in order to line up properly, the density curve and the histogram should both integrate to the same value. Try varying the number of bins until you find a display that is informative. Comment on the qualitative nature of the fits.

6. Display the deviations of the fitted density from the histogram counts as a hanging rootogram (see Section 9.8 of the text). Are there regions in which their appears to be systemmatic misfit? Are these the same in each of the three experiments?

7. Goodness of fit can also be examined via probability plots in which the quantiles of a theoretical distribution are plotted against those of the empirical distribution (Section 9.9 of the text). Following the discussion in Section 9.9, show that it is sufficient to plot the observed order statistics, $X_{(k)}$ versus the quantiles of the Rayleigh distribution with $\theta = 1$. Construct three such probability plots and comment on any systematic lack of fit that you observe. Compare to your answers to

the previous question. To get an idea of what sort of variability could be expected due to chance, simulate several sets of data from a Rayleigh distribution and make corresponding probability plots.

8. Finally, formally test goodness of fit by performing a chi-squared goodness of fit test, comparing histogram counts to those predicted from the Rayleigh model. You may need to combine cells of the histograms so that the expected counts in each cell are at least 5.

References:

R.K Sachs, G. van den Engh, B. Trask, H. Yokota, and J.E. Hearst (1995). A random-walk/giant-loop model for interphase chromosome. *Proc. Nat. Acad. Sci. USA* **92**, 2710-2714.

H. Yokota, G. van den Engh, J.E. Hearst, R.K. Sachs, and B. Trask (1995). Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. *J. Cell Biol.* **130**, 1239-1249.