Strong Limit Theorems

MATH285K - Spring 2010 *Presenter: Forrest Crawford*

Reference: [DK91].

1 Introduction

In biology and bioinformatics, it is often necessary to identify homologous sites in two or more molecular sequences of DNA or amino acid residues. There exist many deterministic, probablistic, and heuristic methods for finding a gapless alignment that achieves a good match given a sceme for giving each alignment a **score**. Choosing a "good" alignment depends on a probablistic interpretation of the serial scoring process. To conduct a statistical test of significance, it is necessary to determine the (asymptotic) distribution of alignment scores under a null model.

2 The Main Result

Suppose two aligned sequences give rise to an independent score X_i at each site i with $\mathbb{E}(X) < 0$ and $\Pr(X > 0) > 0$. Let $S_m = \sum_{i=1}^{m} X_i$ with $S_0 = 0$, and note that S has negative drift. Additionally let

$$M(n) = \sup_{0 \leq k < l \leq n} (S_l - S_k) = \sup_{0 \leq k < l \leq n} \sum_{j=k}^{l} X_j$$

be the maximum segmental score. Define the stopping times $K_0 = 0$ and generally $K_\nu = \min\{k \geq K_{\nu-1} + 1,\ S_k - S_{K_{\nu-1}} \leq 0\}$, with $\nu = 1, 2, \ldots$. K_ν is called a **ladder point**. Since the process S_m has negative drift, the random variables K_ν are finite-valued. Further define the stopping times

$$T_\nu(y) = \inf\{k :\ k > K_{\nu-1} \text{ s.t. either } S_k - S_{K_{\nu-1}} \leq 0 \text{ or } S_k - S_{K_{\nu-1}} \geq y\}$$

for $y > 0$, so $T_\nu(y)$ is the time of first exit from the interval $(0, y)$ since $K_{\nu-1}$. The interval $(K_{\nu-1}, T_\nu(y))$ is called a **y-excursion**. Let $L_\nu(y) = T_\nu(y) - K_{\nu-1}$ be the length of the νth y-excursion. Let $I_t(y) = 1$ if at time t, $S_k - S_{K_{\nu-1}} \geq y$ be the indicator that an excursion attains height y or more. We seek to characterize the asymptotic distribution of $L_v(y)$ for large y-excursions during the epoch of the **maximal $M(n)$-excursion**.

Theorem 1.

$$\frac{L_\nu(y)}{y} \to \frac{1}{w^*}$$

almost surely as $y \to \infty$, where $w^* = \mathbb{E}(Xe^{\theta^* X})$, and θ^* is the unique positive root of the equation $\mathbb{E}(e^{\theta X}) = 1$.

The proof is very long, so we present a rough sketch. First, there is a unique $\theta^* > 0$ satisfying $\mathbb{E}(e^{\theta^* X}) = 1$ by the convexity of $\sum \Pr(X = x)e^{\theta x}$. Furthermore, $w^* = \mathbb{E}(Xe^{\theta^* X}) > 0$.

Let $\psi(\theta) = \mathbb{E}\left(e^{\theta X}\right)$ be the moment-generating function of X. The family of random variables

$$P_m = \frac{e^{\theta S_m}}{[\psi(\theta)]^m},\ m = 0, 1, \ldots$$

is a Wald martingale [KT75]. Note that this does not mean that S is a martingale. The **optional sampling theorem** implies that for $m = L$,

$$\mathbb{E}[e^{\theta S_L - L\log(\psi(\theta))}] = \mathbb{E}[P_1] \equiv 1. \tag{1}$$

This equation for the expected state of the Wald martingale P at a time L is used to show that the probability an excursion reaches a height greater than y has roughly exponential tail decay:

$$0 < \delta \leq e^{\theta^* y} \Pr(I(y) = 1) \leq 1. \tag{2}$$

Differentiating Eq 1 with respect to θ with $\theta = \theta^*$ (recall $\psi(\theta^*) = 1$ and $\psi'(\theta^*) = w^*$) yields

$$\mathbb{E}\left[(S_L - w^* L)e^{\theta^* S_L}\right] = 0$$

and differentiating again gives

$$\mathbb{E}\left[(S_L - w^* L)^2 e^{\theta^* S_L}\right] = \frac{d}{d\theta}\left[\frac{\psi(\theta)}{\psi'(\theta)}\right]_{\theta^*} \mathbb{E}[Le^{\theta^* S_L}]$$

These yield the sum of expectations conditional on whether the walk is greater than y,

$$\begin{aligned} w^*\mathbb{E}\left[Le^{\theta^* S_L}\right] = e^{\theta^* y}\Pr(I(y) = 1)\mathbb{E}\left[S_L e^{\theta^*(S_L - y)}\,\middle|\, I(y) = 1\right] \\ + \Pr(I(y) = 0)\mathbb{E}\left[S_L e^{\theta^* S_L}\,\middle|\, I(y) = 0\right] \end{aligned} \tag{3}$$

Applying Eq 2, we obtain

$$\mathbb{E}[(S_L - w^* L)^2 e^{\theta^* S_L}] = O(y) \tag{4}$$

and

$$\mathbb{E}[S_L^2 e^{\theta^* S_L}] = O(y^2)$$

Expanding the square in Eq 4 and applying Eq 2, we eventually arive at

$$\mathbb{E}\left[(S_L - w^*L)^4 e^{\theta^*(S_L - y)} \Big| I(y) = 1\right] = O(y^2) \tag{5}$$

and substituting $S_L = y$, we get

$$\mathbb{E}\left[(y - w^*L(y))^4 \Big| I(y) = 1\right] = O(y^2) \tag{6}$$

Let $\tau_\nu(n)$ be the νth time at which the process S_k starting at $\kappa_\nu(y)$ first departs from $(0, y)$. Then $\tau_\nu(y) - \kappa_\nu(y)$ has the distribution of $L(y)|\{I(y) = 1\}$. Then rearranging terms in Eq 6, we find that

$$\mathbb{E}\left[\left(\frac{\tau_j(n) - \kappa_j(n)}{n} - \frac{1}{w^*}\right)^4\right] \leq \frac{C}{n^2} \tag{7}$$

Now consider the event that the quantity in parenthesis above is greater than ϵ. Applying Markov's inequality to Eq 7, we have for every $\epsilon > 0$ and some $C_\epsilon < \infty$,

$$\sum_{n=1}^{\infty} \sum_{j=1}^{[A\log(n)]} \Pr\left(\left|\frac{\tau_j(n) - \kappa_j(n)}{n} - \frac{1}{w^*}\right| > \epsilon\right) \leq C_\epsilon A \sum_{n=1}^{\infty} \frac{\log n}{n} < \infty. \tag{8}$$

Then since this sum is finite, by the Borel-Cantelli lemma,

$$\lim_{n\to\infty} \max_j \left|\frac{\tau_j(n) - \kappa_j(n)}{n} - \frac{1}{w^*}\right| = 0 \text{ almost surely.} \tag{9}$$

Replacing $\tau_\nu(n) - \kappa_\nu(n)$ by $L_n(y)$ completes the proof.

3 An Application

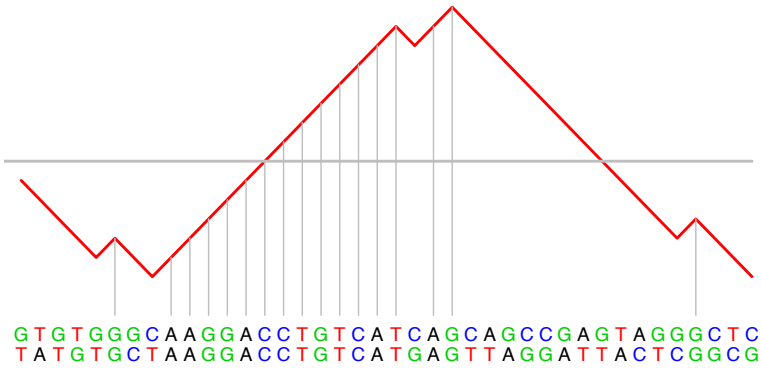
Suppose two sequences A and B are aligned so that the letter A_i is regarded as homologous to the letter B_i for all i. The sequences A and B are realizations of iid variables taking values in some set, such as $C = \{A, G, C, T\}$. Under our null model, the probability of observing letters a and b in the two sequences is $\Pr(A_i = a, B_i = b) = p_a p_b$. Let $X_i = X(A_i = a, B_i = b) = s_{ab}$ be the **score** at site i. We stipulate that mismatches have a negative score so that there is negative drift in the scoring process:

$$\mathbb{E}(X) = \sum_{\{a,b\}} s_{ab} p_a p_b < 0$$

but positive scores are possible: $\Pr(X > 0) > 0$. We define the partial sum of the scores as

$$S_m = \sum_{i=1}^{m} X(A_i, B_i)$$

In the diagram below, two aligned sequences provide a realization of the walk S. It reaches the ladder point $K_{\nu-1}$ at the C/T site and climbs beyond y, shown as a horizontal line. Vertical lines identify the matching sites which contribute a positive score. $T_\nu(y)$ is the C/C match where S reaches y.



Theorem 1 applies directly to the quantities introduced above. This, along with other results relating to the asymptotic distribution of maxima can be used to derive the asymptotic distribution function of $M(n)$ for use in statistical tests of alignments.

References

[DK91] Amir Dembo and Saumuel Karlin. Strong limit theorems of empirical functionals for large exceedances of partial sums of iid variables. *The Annals of Probability*, 19(4):1737–1755, 1991.

[KT75] Sanuel Karlin and Howard M. Taylor. *A first course in stochastic processes*. Academic Press, New York, 2nd edition, 1975.