Genome Assembly with Long Noisy Reads III

Seed Statistics

Amir Kadivar

February 2016

abstract *In an earlier report we considered the statistical properties of diagonal distances (aka "shifts") among seeds of overlapping sequences. A probabilistic model of diagonal distances is needed for two reasons: first, to find the width of diagonal strips to analyze when testing the overlapping hypothesis, and second, to find the radius of the banded overlap alignment. Here, we complete the statistical analysis of diagonal distances and introduce the corresponding analysis for seed distances along diagonals which is needed for chaining.*

Contents

Diagonal Distance

The diagonal distance between two seeds is precisely what we formerly refered to as *shifts*. We found a recurrence relation for the probability of an alignment of length T leaving a band of radius r and demonstrated, via a random walk model, that the recurrence relation is a stable finite difference scheme for a diffusion IBVP. The idea is to approximate the solution of the recurrence relation by the analytic solution of the IBVP (the reverse of what one does in numerical solution of differential equations). Furthermore, since we wish to optimize over the *domain* of the PDE we need solutions that are computationally easy to invert. To this end, we proposed an approximation of the IBVP by ignoring the boundary conditions. The goal of this section is to justify this approximation.

Original Problem Given some $r > 0$ let P_r be the following IBVP:

$$\begin{cases} u_t - pu_{xx} & = 0 \\ u(x,0) & = 1 \text{ over } (-r,r) \\ u(\pm r,t) & = 0 \end{cases}$$

For given $T > 0$ and $\epsilon \ll 1$, we wish to find the smallest r^* such that the solution u of P_{r^*} satisfies:

$$u(0,T) \geq 1 - \epsilon$$

Fourier transform is not applicable due to the boundary conditions and the solution found by separation of variables gives:

$$u(0,T) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)\pi} \exp\left(-pT\frac{(n+\frac{1}{2})\pi}{r}\right)$$

which, though exact, is not analytically invertible in r. For convenience, from here on we let $\mathscr{L}$ be the differential operator $\mathscr{L} = \partial_t - p\partial_{xx}$.

Proposition *(Approximate Problem):* Let $\hat{P}_r$ be the following IVP obtained by relaxing the boundary condition of P_r:

$$\begin{cases} \mathscr{L}\hat{u}_t & = 0 \\ \hat{u}(x,0) & = 1 \text{ over } (-r,r) \end{cases}$$

Define r^* to be:

$$r^* \equiv 2\sqrt{pT}\,\mathrm{erf}^{-1}(1-\epsilon)$$

Then the solution $\hat{u}$ of $\hat{P}_{r^*}$ satisfies:

$$\hat{u}(0,T) \geq 1 - \epsilon$$

proof: For $\hat{P}_r$ Fourier transform would be applicable yielding the solution: $\hat{u} = \Phi * \hat{u}(x,0)$ where the convolution is over x and Φ is the diffusion kernel:

$$\Phi(x,t) = \frac{1}{\sqrt{4\pi pt}} \exp(-\frac{x^2}{4pt})$$

At any time t, this is precisely the probability density function of $\mathscr{N}(0,\ \sqrt{2pt})$ with cumulative distribution:

$$F(x;t) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{2\sqrt{pt}}\right)\right]$$

Thus:

$$\hat{u}(x,t) = F(x+r;t) - F(x-r;t) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x+r}{2\sqrt{pt}}\right) - \mathrm{erf}\left(\frac{x-r}{2\sqrt{pt}}\right)\right]$$

implying:

$$\hat{u}(0,T) = \mathrm{erf}\left(\frac{r}{2\sqrt{pT}}\right)$$

Noting that erf is a monotonically increasing function completes the proof. ■

Remark The formula for $\hat{u}(x,t)$ simplifies at $x = 0$ and at $x = \pm r$:

$$\hat{u}(0,t) = \operatorname{erf}\left(\frac{r}{2\sqrt{pt}}\right) \quad \text{and} \quad \hat{u}(\pm r,t) = \frac{1}{2}\operatorname{erf}\left(\frac{r}{\sqrt{pt}}\right) \tag{1}$$

Proposition *(Accuracy of approximation):* Let r^* be such that the solution $\hat{u}(x,t)$ of the approximate IVP $\hat{P}_{r^*}$ satisfies:

$$\hat{u}(0,T) \geq 1 - \epsilon$$

Then the solution $u(x,t)$ of the exact IBVP P_{r^*} satisfies:

$$u(0,T) \geq 1 - \frac{3}{2}\epsilon$$

proof: Define the residue $v(x,t) = \hat{u}(x,t) - u(x,t)$. Clearly, we have $v(x,t) \geq 0$ for all x and $t > 0$ and we wish to show:

$$v(0,T) \leq \frac{1}{2}\epsilon$$

We know that v satisfies the IBVP:

$$\begin{cases} \mathscr{L}v & = 0 \\ v(x,0) & = 0 \text{ over } (-r,r) \\ v(\pm r,t) & = f(t) \end{cases}$$

where, using (1), we have defined:

$$f(t) \equiv \hat{u}(\pm r,t) = \frac{1}{2}\operatorname{erf}\left(\frac{r}{\sqrt{pt}}\right)$$

with $f(0) = 1/2$ to maintain continuity. Now define $h(x,t)$ for $t \geq 0$ by

$$h(x,t) = f(t) - v(x,t)$$

The function h satisfies the IBVP:

$$\begin{cases} \mathscr{L}h & = \dot{f} \\ h(x,0) & = \frac{1}{2} \text{ over } (-r,r) \\ h(\pm r,t) & = 0 \end{cases}$$

The solution is the sum of a particular solution and a homogenous solution:

$$h(x,t) = \frac{1}{2}\hat{u}(x,t) + [\Phi * \dot{f}]_{x,t}$$

where the convolution is over both x and t:

$$[\Phi * \dot{f}]_{x,t} = \int_0^t \int_{-\infty}^{\infty} \Phi(x-y, t-s)\dot{f}(s)dyds$$

Therefore:

$$\begin{aligned} h(0,T) &= \frac{1}{2}\hat{u}(0,T) + \int_0^T \dot{f}(s) \int_{-\infty}^{\infty} \Phi(y, T-s)dyds \\ &= \frac{1}{2}\hat{u}(0,T) + \int_0^T \dot{f}(s)ds = \frac{1}{2}\hat{u}(0,T) + f(T) - \frac{1}{2} \end{aligned}$$

where in the last step we have used the fact that $\Phi(\cdot, t)$ is a probability distribution over $\mathbb{R}$ for any t. Finally:

$$v(0,T) = f(T) - h(0,T) = \frac{1}{2}(1 - \hat{u}(0,T)) \leq \frac{1}{2}\epsilon$$

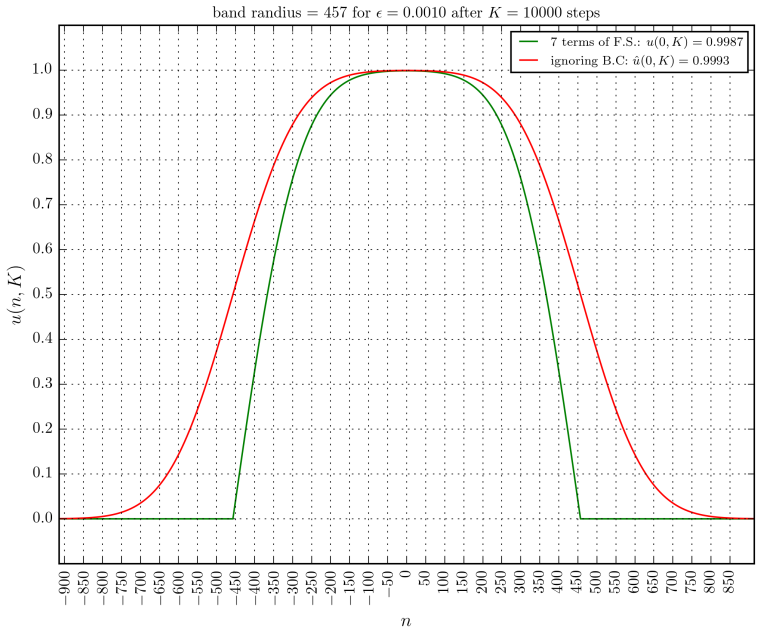
■

Corollary *(Approximation Algorithm):* It follows that if r^* is defined as:

$$r^* \equiv 2\sqrt{pT}\text{erf}^{-1}\left(1 - \frac{2\epsilon}{3}\right)$$

then the solution u of P_{r^*} satisfies:

$$u(0,T) \geq 1 - \epsilon$$

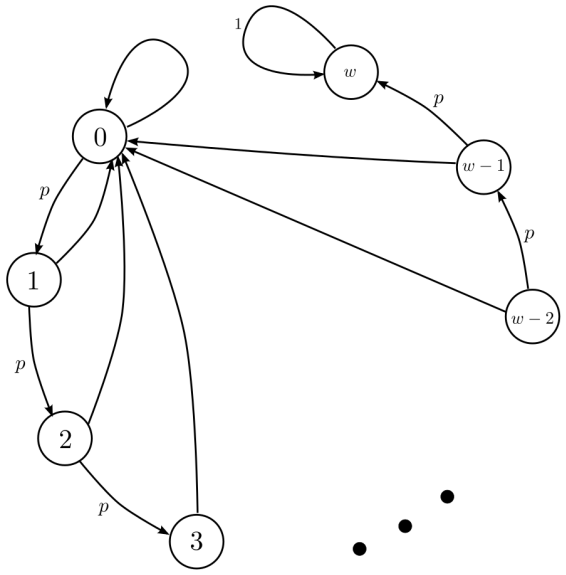


Seed Chaining

We now consider the following problem: given a set of seeds within a diagonal strip of the dynamic programming table, which partial overlap alignment maximizes the likelihood of the observed seeds, where a partial overlap alignment is a sequence of chainable seeds. This can be reduced (detailed analysis will follow in another report) to a heaviest-path problem over the DAG of all seeds in the given diagonal strip where the weight of edges connecting two seeds is obtained from calculating the the probability that a sub-alignment of length n does not contain any seeds of length k and. We noted earlier that this corresponds to the probability distribution of waiting times between runs of k successes in a sequence of i.i.d Bernoulli trials (where k is the word length). Some combinatorial properties of this distribution, known as the *k-th order geometric distribution*, are known but they are not computationally useful. We here propose an exact algorithm to calculate this distribution efficiently.

Computational Requirements We seek the probability distibution $f(n;w)$ of a sub-alignment of length n containing no seed of length w. However, as opposed to the diagonal distance case, we do not have the complication of dependence on diagonal position. Therefore, it suffices to solve a recurrence relation once and use the results for all sequence comparisons.

Probabilistic Model Consider the following Markov chain where w is the word length and p is the probability of an exact match at any given position of the alignment:



Let $u(n, k)$ be the probability of being at state n at time k. Since $u(n, w)$ is the probability of observing a seed by the n-th step of the alignment the desired distribution $f(n)$ is given by:

$$f(n; w) = 1 - u(n, w)$$

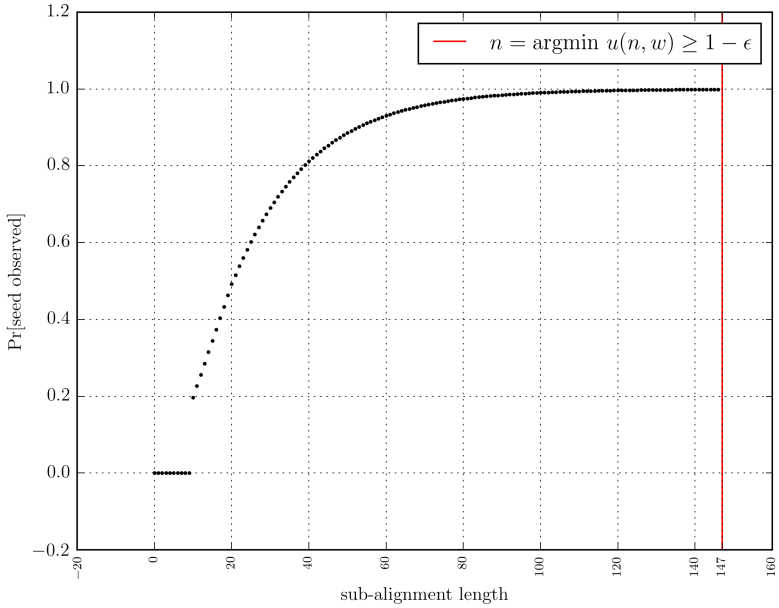
The recurrence relation of $u(n, k)$ for $0 < n < w$ is:

$$u(n, k) = pu(n - 1, k - 1)$$

with boundary and initial conditions:

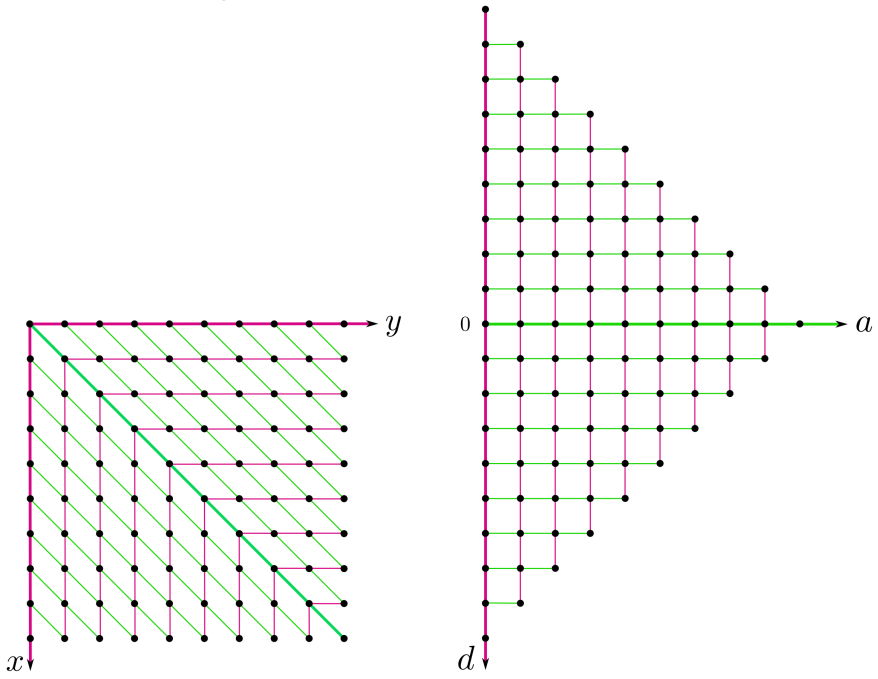
$$u(w, k) = pu(w - 1, k - 1) + u(w, k - 1)$$
$$u(0, k) = 1 - \sum_{n=1}^{w} p(n, k)$$

We know that $f(n; w)$ is decreasing and thus, in practice, we only need to calculate $f(n)$ upto some threshold, say for all n such that $f(n; w) > \epsilon$. We then solve the recurrence relation in increasing order of n and decreasing order of k as long as $u(n, w) < 1 - \epsilon$. The following is a plot of $u(n, w)$ for $w = 10$, $p = 0.85$, and $\epsilon = 10^{-3}$. Since the very same distribution is used for all sequence comparisons the computational cost is effectively zero.



Mapping Bands to Rectangular Grids

In order to only allocate the necessary memory for the banded overlap alignment we need to map bands (diagonal strips) to rectangular grids in memory. This requires a change of coordinates which maps parallelograms or trapezoids bound by the edges of the dynamic programming table to (roughly) rectangular regions. Two alternatives were previously discussed. Here, we present a refined version of the more convenient of the two: coordinates based on shift and distance from diagonal start cell:



Let (x, y) denote coordinates in the dynamic programming table and (d, a) denote the new

coordinates where d is the shift $x - y$ and a is the distance along the starting cell of the d-diagonal. The change of coordinates mapping is given by $\phi : \mathbb{Z}^2 \to \mathbb{Z}^2$:

$$(x, y) \xrightarrow{\phi} \big(x - y, \min(x, y)\big)$$
$$\big(a + \max(d, 0), a - \min(d, 0)\big) \xleftarrow{\phi^{-1}} (d, a)$$

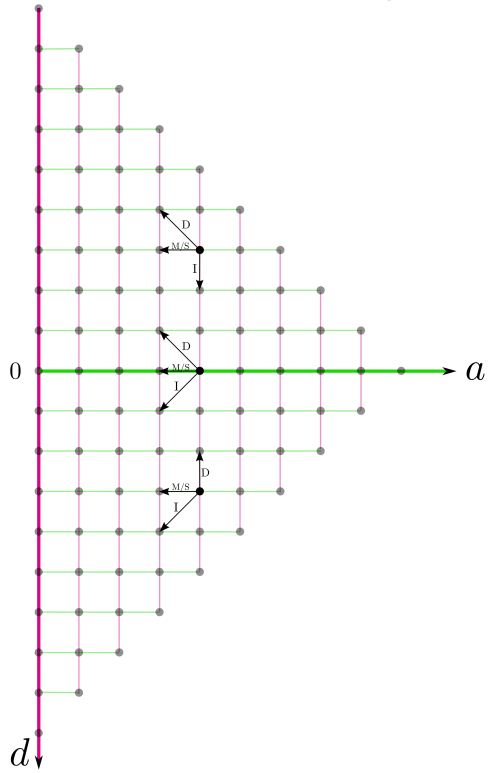
Furthermore, the length of the row at height d in the transformed coordinates is:

$$L(d) = \min(|S| - d, |T|) + \min(d, 0) + 1$$

The alignment band Ω is the subset of the grid that needs to be populated. We have:

$$\Omega_{xy} = \{(x, y); d_{\min} \le x - y \le d_{\max}\}$$
$$\Omega_{da} = \{(d, a); d_{\min} \le d \le d_{\max}\}$$

Dynamic programming dependence rules depend on the sign of d:



For simplicity, we populate the dynamic programming table in the natural order of (x, y)-coordinates (while memory is mapped in (d, a)-system). To avoid sweeping the entire (x, y)-grid for in-band cells we use the following bounds:

$$\forall (x, y) \in \Omega : \max(0, d_{\min}) \le x \le \min(|S|, |T| + d_{\max})$$
$$\forall (x, y) \in \Omega : \max(0, x - d_{\max}) \le y \le \min(|T|, x - d_{\min})$$