Sequence alignment through TASEP

MATH285K - Spring 2010 *Presenter: Alexander Vandenberg-Rodes*

Reference: [1].

The problem is as follows: suppose we have two DNA (or protein) sequences, and we want to measure the degree of similarity between them. For this purpose, computational biologists have invented a multitude of scoring systems. However, rigorous statistics on what constitutes a close match have been hard to come by. The most elementary scoring systems are termed *gapless*, and they account only for strict site-by-site comparisons. However, gapless scoring systems have well-known deviation behavior, given by the Gumbel distribution, which we now describe.

Let

$$s(a,b) = \begin{cases} 1 & \text{if } a = b \\ -\mu & \text{if } a \neq b \end{cases}$$

be the assigned score for the pair (a, b) of letters (taken from A,G,T, or C, for example). Here, $\mu > 0$ is the mismatch cost. Let

$$\Sigma(\vec{a}, \vec{b}) = \max_{\mathcal{A}} \sum_{(a,b)\in\mathcal{A}} s(a,b) \tag{1}$$

be the optimal local score of two sequences $\vec{a}, \vec{b}$, where the maximum is taken is over all gapless alignments – same-length chunks of the two sequences. For example, if

$$\vec{a} = CG\overbrace{ATGC}T$$
$$\vec{b} = T\underbrace{GCTC}GA,$$

where the two chunks are identified by the braces, then the sum in (4) is over the pairs $(A, G), (T, C), (G, T)$, and (C, C).

Assuming two independent sequences $\vec{a}, \vec{b}$ of lengths $L \equiv M$, with iid coordinates, and such that the score of a pair of coordinates has negative expected value but positive probability of a positive score, the optimal local score has approximately, for large L and M, the Gumbel distribution

$$\mathbb{P}(\Sigma < S) = e^{-\kappa LMe^{-\lambda S}}, \tag{2}$$

with λ the solution of

$$\mathbb{E}e^{\lambda s(a_1,b_1)} := \sum_{a,b} \mathbb{P}(a_1 = a, b_1 = b)e^{\lambda s(a,b)} = 1, \tag{3}$$

and κ a more complicated function of the scoring matrix and letter distribution.

If one wants to consider also insertions and deletions in the sequence, one can introduce *gapped* alignment. There, one also allows gaps in the chunks, but with a penalty $\delta > 0$ for each gap. Specifically, (4) becomes

$$\Sigma(\vec{a}, \vec{b}) = \max_{\mathcal{A}} \sum_{(a,b)\in\mathcal{A}} s(a, b) - \delta(\#gaps), \tag{4}$$

where the maximum is over all alignments with gaps. Unfortunately, rigorous statistics for this scoring system are still unknown. Now, it is assumed henceforth – with support from numerical studies – that gapped alignment also follows the Gumbel distribution (2). Then main question is then how to compute the tail decay parameter λ, which is what we now consider.

Global alignment scoring.

We define the global alignment scoring function $h(r, t)$ by the recursion

$$h(r, t + 1) = \max\{h(r, t - 1) + \sigma(r, t), h(r + 1, t) - \delta, h(r - 1, t) - \delta\}, \tag{5}$$

with the initial conditions $h(\pm k, k) = k\delta$. Here, $\sigma(r, t) = s(a_i, b_j)$ where $(r, t) = (i - j, i + j - 1)$. Then $h(0, k)$ is precisely the highest scoring alignment of the first k letters of the two sequences (see figure 2 in [1] for the lattice representation of the coordinates (r, t)).

Let

$$Z(\lambda, 2N) = \mathbb{E}e^{\lambda h(0,2N)} \tag{6}$$

be the generating function for $h(0, 2N)$. Some non-rigorous arguments indicate that the analogue of (3) is solving for

$$\lim_{N\to\infty} Z(\lambda, 2N) = 1. \tag{7}$$

Notice that if one reduces to the gapless case by taking $\delta = \infty$, then by relation (5), (6) reduces to

$$\mathbb{E}e^{\lambda \sum_{t=1}^{N} \sigma(0,2t)} = \left(\mathbb{E}e^{\lambda s(1,1)}\right)^N,$$

because the collection of scores $\{s(i, i)\}$ are iid random variables. Hence we are in agreement with (3).

We can now make a connection between gapped alignment and the current in the *totally asymmetric exclusion process* (TASEP); the latter having the distinction of being perhaps the most-studied model of stochastically interacting particles.

A very short intro to TASEP.

Imagine one has a collection of sites $\{1, 2, \ldots, 2N\}$, arranged in a ring. Each site either contains a particle (denoted by a 1), or is empty (denoted by a 0). With TASEP, as with any other interacting particle system, one starts with some initial condition, e.g. $(1001011\cdots010)$. We then apply the following dynamics[1]: at each odd time step ($t = 1, 3, 5, \ldots$), each particle at an *even* position with probability p attempts a jump to its "right", i.e. $2 \to 3, 4 \to 5, \ldots, 2N \to 1$. If there already is a particle to its right, the jump fails and the particle stays where it was. Similarly, at each even time step the particles on the odd sites attempt to move. One of the most important quantities associated with TASEP is the *current*. One picks a bond, say between site 1 and site 2, and calculates how many particles made the $1 \to 2$ jump after T time steps. We will denote this as $J_{12}(T)$, the current across the bond $1 \to 2$ at time T.

The connection with the height function is as follows. We consider the case where $\delta = \mu = 0$, which, although seemingly irrelevant to biological applications, applies to the slightly more interesting case of $\delta = \mu/2 > 0$ via the easy relation

$$h_\mu(0, 2N) = (1 + \mu)h(0, 2N) - N\mu$$

between the two global alignment scores. We also assume that the N^2 collection of scores $s(a_i, b_j)$ are iid, even though they are generated by only $2N$ variables. We then consider the discrete gradient

$$n(r, t) = \begin{cases} h(r + 1, t + 1) - h(r, t) & \text{for } r + t \text{ odd,} \\ h(r + 1, t) - h(r, t + 1) + 1 & \text{for } r + t \text{ even} \end{cases}$$

Through the relation (5), it turns out that the variables $n(r, t)$ are all binary valued, with the following dynamics: if $r + t$ is even, $n(r - 1, t - 1) = 1$, $n(r, t - 1) = 0$, *and* $\sigma(r, t) = 0$, then $n(r - 1, t) = 0$ and $n(r, t) = 1$. Otherwise, $n(r - 1, t) = n(r - 1, t - 1)$ and $n(r, t) = n(r, t - 1)$. In the above language of TASEP, $n(r, t)$ indicates whether there is a particle occupying site r at time t, and $\sigma(r, t)$ indicates whether a jump is *not* attempted. The condition that $r + t$ be even enforces that particles at even sites only jump at odd times, and vice-versa[2].

Finally, through a simple but notationally awkward computation, it can been seen that the global score $h(0, 2N)$ is the TASEP current $J_{12}(2N)$, defined above. While this latter quantity has been studied extensively under different space-time

[1] this is known as TASEP with discrete-time sub-lattice parallel update. Most of the work on TASEP has dealt with the continuous-time version.

[2] I have here completely ignored the problem of boundary conditions, which leads to the ring topology for TASEP with particles initially occupying every other site.

scaling regimes and jump dynamics, [1] devotes the rest of the paper to analyzing the generating function of the current in this particular setting, so that one can solve (7).

References

[1] R. Bundschuh, *Asymmetric exclusion process and extremal statistics of random sequences*, Phys. Rev. E **65** (2002).