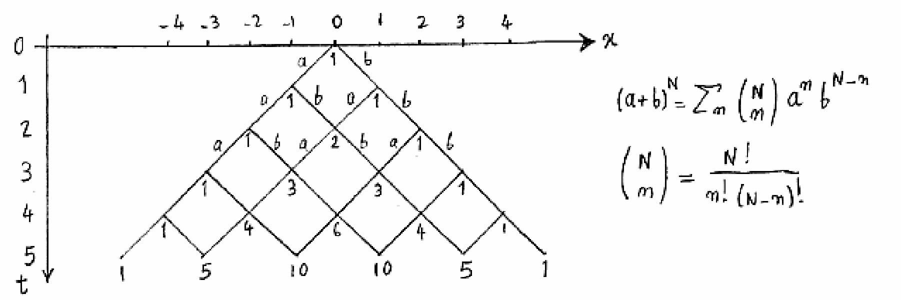
1.5 Sequence alignment

The dramatic increase in the number of sequenced genomes and proteomes has lead to development of various *bioinformatic* methods and algorithms for extracting information (data mining) from available databases. *Sequence alignment* methods (such as BLAST) are amongst the earliest and most widely used tools, essentially attempting to establish relations between sequences based on common ancestry, for example as a means of guessing function.

As discussed in the earlier lectures by Prof. Mirny, the *explicit inputs* are two (or more) sequences (or nucleotides for DNA/RNA, or amino-acids for proteins)

$$\{a_1, a_2, \ldots, a_m\} \text{ and } \{b_1, b_2, \ldots, b_n\},$$

for example, corresponding to a query (newly sequenced gene) and a database. *Implicit inputs* are included as part of the *scoring procedure*, e.g. by assigning a *similarity matrix* $s(a, b)$ between pairs of elements, and costs associated with initiating or extending *gaps* $s(a, -)$. *Global alignments* attempt to construct the single best match that spans both sequences, while *local alignments* look for (possibly) multiple subsequences than are represent good local matches. In either case, *recursive algorithms* enable scanning the exponentially large space of possible matches in polynomial time. Within bioinformatics these methods are referred to as *dynamic programming*, in statistical physics they appear as *transfer matrices*, and have precedent in early recursive methods such as in the construction of binomial coefficients with the *binomial triangle* (below).



In most implementations the output of the algorithm is an optimal match, and a corresponding score S. An important question is whether this output is due to a meaningful relation between the tagged sequence (e.g. due to common ancestry, or functional convergence), or simply a matter of chance (e.g. due to the large size of database). To rule out the latter, we need to know the probability that a score S is obtained randomly. This probability can be either obtained numerically by applying the same algorithm to randomly generated (or shuffled) sequences, or if possible obtained analytically. Analytical solutions are particularly useful as significant alignment scores are likely to fall in the tails of the random distribution; a portion that is hard to access by numerical means.

1.5.1 Significance of gapless alignments

We can derive some analytic results in the case of alignments that do not permit gaps. We again begin with two sequences

$$\vec{a} \equiv \{a_1, a_2, \ldots, a_m\}, \quad \text{and} \quad \vec{b} \equiv \{b_1, b_2, \ldots, b_n\}\,,$$

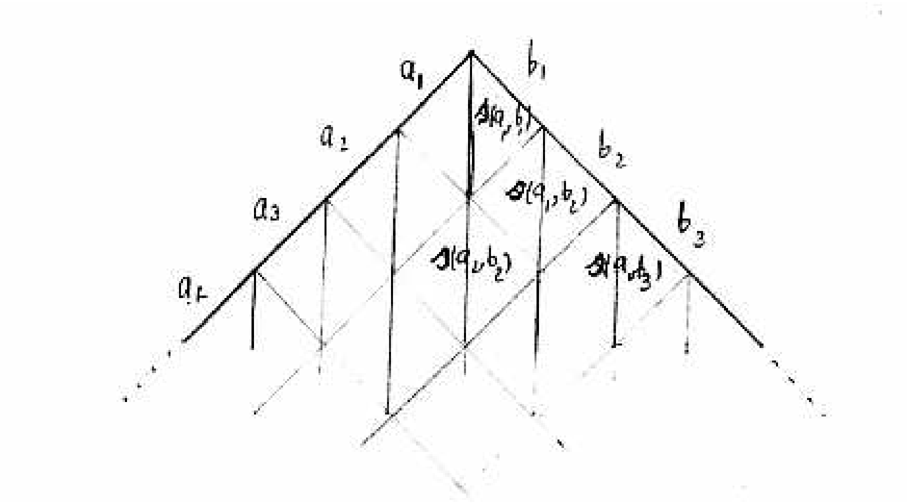
of lengths m an n respectively. We can define a matrix of alignment scores

$$S_{ij} \equiv \text{Score of best alignment terminating on } a_i, b_j. \tag{1.92}$$

For gapless alignments, this matrix can be built up recursively as

$$S_{ij} = S_{i-1,j-1} + s(a_i, b_j), \tag{1.93}$$

where $s(a,b)$ is the scoring matrix element assigned to a match between a and b. This alignment algorithm can be made to look somewhat like the binomial triangle if we consider the following representation: Place the elements of sequence $\vec{a}$ along one edge of a rectangle, characters of the sequence $\vec{b}$ along the other edge, an rotate the rectangle so that the point $(0,0)$ is at the apex, with the sides at ±45°. The square (i,j) in the rectangle is to be regarded as the indicator of a match between a_i and b_j, and a corresponding score $s(a_i,b_j)$ is marked along its diagonal.



To make a connection to dynamical system, we now introduce new coordinates (x,t) by

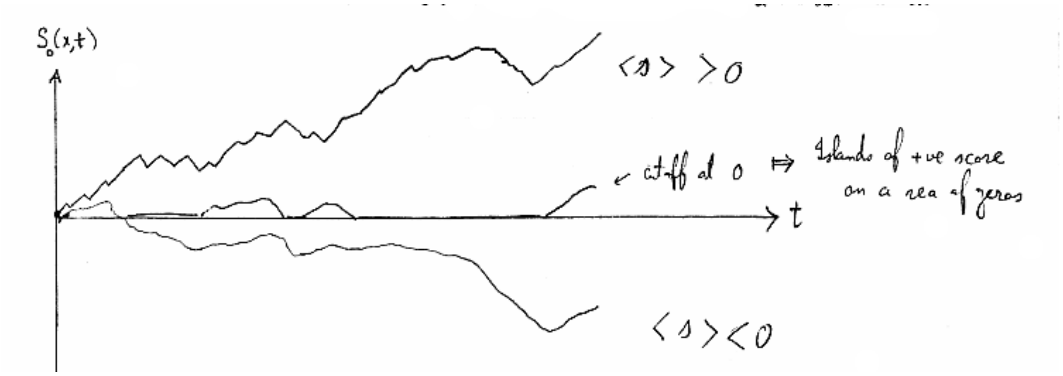
$$x = j - i, \;\; t = i + j. \tag{1.94}$$

The recursion relation in Eq. (1.93) is now recast as

$$S(x,t) = S(x,t-2) + s(x,t)\,, \tag{1.95}$$

describing the 'time' evolution of the score at 'position' x. In this gapless algorithm, the columns for different x are clearly evolve independently; as we shall point out later gapped alignments correspond to jumps between columns. For a global (*Needleman–Wunsch*) alignment, the column with the highest score at its end point is selected, and the two matching sub-sequences are identified by tracing back. If the two sequences are chosen randomly, the corresponding scores $s(x,t)$ will also be random variables. The statistics of random (global) alignment is thus quite simple: According to the central limit theorem the sum of a large number of random variables is Gaussian distribute, with mean and variances obtained by multiplying the number of terms with the mean and variance of a single step. By comparison with such a Gaussian PDF, we can determine the significance of a global gapless alignment score.

The figure below schematically depicts two evolving scores generated by Eq. (1.95). In one case the mean $\langle s\rangle$ of pairwise scores is positive, and the net score has a tendency to increase, in another $\langle s\rangle < 0$ and $S(t)$ decreases with increasing sequence length. In either case, the overall trends mask local segments where there can be nicely matched (high scoring) subsequences.



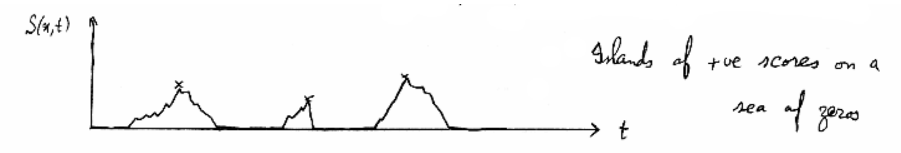
This issue is circumvented in a local (*Smith-Waterman*) alignment scheme in which negative scores are cut off according to

$$S_{ij} = \max\{S_{i-1,j-1} + s(a_i, b_j), 0\}, \tag{1.96}$$

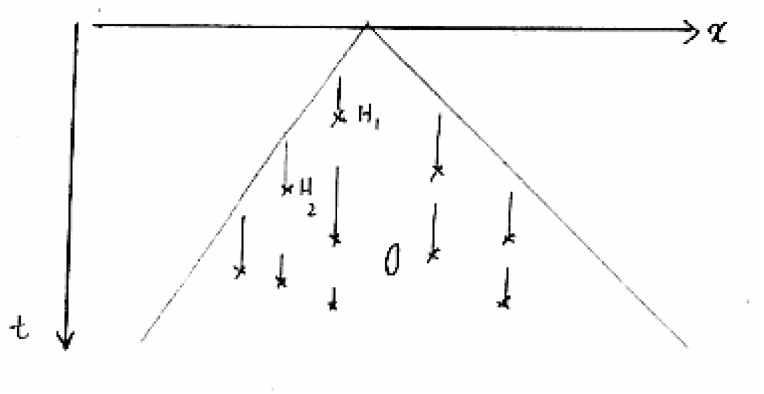
or in the notation of x and t,

$$S(x,t) = \max\{S(x,t-2) + s(x,t-1), 0\}. \tag{1.97}$$

This has little effect if $\langle s\rangle$ is larger than 0, but if $\langle s\rangle < 0$, the net result is to create "islands" of positive S in a sea of zeroes.



The islands represent potential local matches between corresponding sub-sequences, but many of them will be due to chance. Let us denote by H_α the peak value of the score on



island α. To find significant alignments we should certainly confine our attention to the few islands with highest scores. For simplicity, let us assume that there are K islands and we pick the one with the highest peak, corresponding to a score

$$S = \max_{\alpha=1,\cdots,K} \{H_\alpha\}. \tag{1.98}$$

To assign significance to a match, we need to know how likely it is that a score K is obtained by mere chance. Of course we are interested in the limit of very long sequences ($n, m \gg 1$), in which case it is likely that the number of islands grows proportionately to the area of our 'ocean'– the rectangle of sides m and n– i.e.

$$K \propto mn. \tag{1.99}$$

Equation (1.98) is an example of *extreme value statistics*. Let us consider a collection of random variables $\{H_\alpha\}$ chosen *independently* from some PDF $p(H)$, and the extremum

$$X = \max\{H_1, H_2, \ldots, H_K\}, \tag{1.100}$$

The *cumulative probability* that $X \leq S$ is the product of probabilities that any selected H_α is less than S, and thus given by

$$\begin{aligned} P_K(X \leq S) &= \text{Prob.}(H_1 \leq S) \times \text{Prob.}(H_2 \leq S) \times \cdots \times \text{Prob.}(H_K \leq S) \\ &= \left[\int_{-\infty}^{S} dH p(H)\right]^K \\ &= \left[1 - \int_{S}^{\infty} dH p(H)\right]^K . \end{aligned} \tag{1.101}$$

For large K, typical S are in the tail of $p(H)$, which implies that the integral is small, justifying the approximation

$$P_K(S) \approx \exp\left[-K \int_S^\infty dH p(H)\right]. \tag{1.102}$$

Assume, as we shall prove shortly, that $p(H)$ falls exponentially in its tail. We can then write

$$\int_S^\infty dHp(H) = \int_S^\infty dHae^{-\lambda H} = \frac{a}{\lambda}e^{-\lambda H}. \tag{1.103}$$

The cumulative probability function $P_K(S)$ is therefore

$$P_K(S) = \exp\left[-\frac{Ka}{\lambda}e^{-\lambda S}\right], \tag{1.104}$$

and a corresponding PDF of

$$p_K(S) = \frac{dP_K(S)}{dS} = ka\exp\left(-\lambda S - \frac{ka}{\lambda}e^{-\lambda S}\right). \tag{1.105}$$

The exponent

$$\phi(S) \equiv -\lambda S - \frac{ka}{\lambda}e^{-\lambda S}, \tag{1.106}$$

has an extremum when

$$\frac{d\phi}{dS} = -\lambda + kae^{-\lambda S^*} = 0, \tag{1.107}$$

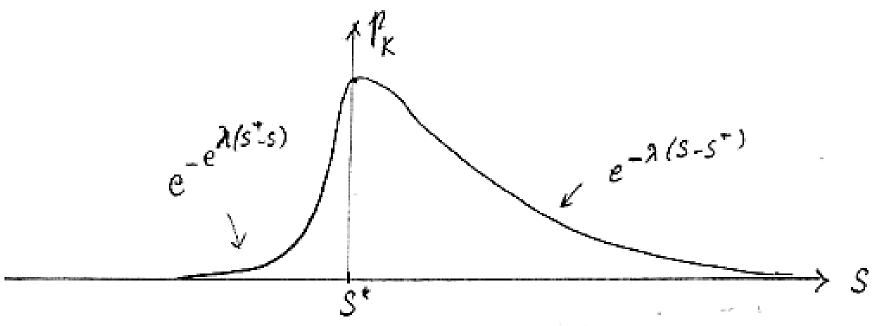
corresponding to a score of

$$S^* = \frac{1}{\lambda}\log\left(\frac{ka}{\lambda}\right). \tag{1.108}$$

Equation (1.108) gives the most probable value for the score, in terms of which we can re-express the PDF in Eq. (1.105) as

$$p_K(S) = \lambda\exp\left[-\lambda(S - S^*) - e^{-\lambda(S-S^*)}\right]. \tag{1.109}$$

Evidently, once we determine the most likely score S^*, the rest of the probability distribution is determined by the single parameter λ. Equation (1.109) is known as the *Gumbel* or the *Fisher-Tippett* extreme value distribution (EVD). It is characterized by an exponential tail above S^* and a much more rapid decay below S^*. It looks *nothing like* a Gaussian, which is important if we are trying to gauge significances by estimating how many standard deviations a particular score falls above the mean.



Given the shape of the EVD, it is clearly essential to obtain the parameter λ, and indeed to confirm the assumption in Eq. (1.103) that $p(H)$ decays exponentially at large H. The height profile of an island (by definition positive) evolves according to Eq. (1.95). For random variables s, this is clearly a Markov process, with a transition probability p_s for a jump of size s (with the exception jumps that render S negative). Thus the probability for a height (island score) h evolves according to

$$\begin{aligned} p(h,t) &= \sum_s p_s p(h-s,t-2) \\ &= \sum_{a,b} p_a p_b p[h-s(a,b),t-2], \end{aligned} \tag{1.110}$$

where the second equality is obtained by assuming that the characters are chosen randomly frequencies $\{p_a, p_b\}$ (e.g. 30% for an A, 30% for a T, and 20% for either G or C in a particular variety of DNA). To solve the precise steady steady solution $p^*(h)$ for the above Markov process is somewhat complicated, requiring some care for values of h close to zero because of the modification of transition probabilities by the constraint $h \geq 0$. Fortunately, we only need the probability $p^*(h)$ for large h (close to the peaks) for which, we can guess and verify the exponential form

$$p^*(h) \propto e^{-\lambda h}. \tag{1.111}$$

Substituting this ansatz into Eq. (1.110), we obtain

$$e^{-\lambda h} = \sum_{a,b} p_a p_b e^{-\lambda(h-s(a,b))}. \tag{1.112}$$

Consistency is verified since we can cancel the h-dependent factor $e^{-\lambda H}$ from both sides of the equation, leaving us with the implicit equation for λ

$$\boxed{\sum_{a,b} p_a p_b e^{\lambda s(a,b)} = 1.} \tag{1.113}$$

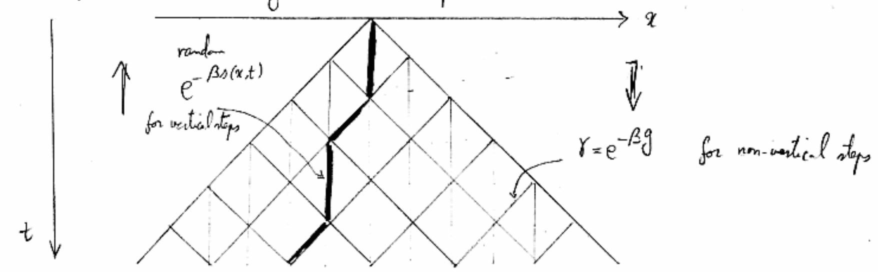
Equation (1.111) shows that the probability distribution for the island heights is exponential at large h. If we consider the highest peak H of the island, the corresponding distribution $p^*(H)$ will be somewhat different. However, as indicated by Eq. (1.109) the maximization will not change the tail of the distribution. Hence the value of λ in Eq. (1.113), along with S^* completely characterizes the Gumbel distribution for statistics of random local gapless alignments. This expression was first derived in the early 1990s by Karlin *et al.*

1.5.2 Gapped alignments

Comparison of biological sequences indicates that in the process of evolution sequences not only mutate, but also lose or gain elements. Consequently, useful alignments must allow for gaps and insertions (more so for more evolutionary divergent sequences). In the scoring process, it is typical to add a cost that is linear in the size of the gap (and sometimes extra costs for the end-points of the gap). Dynamic programming algorithms can be constructed (e.g. below) to deal with gapped alignments, but obtaining analytical results is now much harder. Empirically, it can be verified that the PDF for the score of random gapped *local* alignments is still Gumbel distributed. This result could again be justified by noting that local alignments rely of selecting the best score amongst a large number. The shape of the 'islands' is now slightly different, and their statistics is harder to obtain, as discussed below.

Gaps/insertation can be incorporated in the earlier diagrammatic representation by sideway steps from one column x to another. The sideway step does not advance along the coordinate corresponding to the sequence with gap, but progresses over the characters of the other sequence. As depicted below, these resulting trajectories are still pointed downwards, but may include transverse excursions. Such *directed paths* occur in many contexts in physics from flux lines in superconductors to domain walls in two-dimensional magnets.

• Simple model : The directed path introduced in the previous lecture with random energies $\{s(x,t)\}$ along the vertical steps.



In the spirit of statistical physics we may even introduce a fuzzier version of alignment corresponding to a finite temperature β^{-1}. We can then regards the scores as (negative) energies used to construct Boltzmann weights $e^{\beta S}$ to various paths. Now consider the constrained partition function

$$W(x,t) = \text{sum of all paths' Boltzmann weights from } (0,0) \text{ to } (x,t). \tag{1.114}$$

We can use a so-called *transfer matrix* to recursively compute this quantity by

$$W(x,t) = e^{\beta s(x,y)} W(x,t-2) + e^{-\beta g} \left[W(x+1,t-1) + W(x-1,t-1) \right]. \tag{1.115}$$

The first term is the contribution from the configuration that goes down along the same x, while the remaining two come from neighboring columns (at a cost g in gap energy).

The above transfer matrix is the finite temperature analog of dynamic programming, and indeed in the limit of $\beta \to \infty$, the sum in Eq. (1.115) is dominated by the largest term, leading to $(W(x,t) = \exp\left[\beta S(x,t)\right])$

$$S(x,t) = \max\left\{ S(x,t-2) + s(x,t), S(x+1,t-1) - g, S(x-1,t-1) - g \right\}. \tag{1.116}$$

This analogy has been used to obtain certain results for gapped alignment, but will not be pursued further here.

MIT OpenCourseWare
http://ocw.mit.edu

8.592J / HST.452J Statistical Physics in Biology
Spring 2011

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.