## 1.5 Backward Kolmogorov equation

When mutations are less likely, genetic drift dominates and the steady state distributions are peaked at $x = 0$ and 1. In the limit of $\mu_1 = 0$ (or $\mu_2 = 0$), Eq. (1.68) no longer corresponds to a well-defined probability distribution, as the $1/x$ (or $1/(1-x)$) divergence close to $x = 0$ (or $x = 1$) precludes normalization. This is the mathematical signal that our expression for the steady state is no longer valid in this limit. Indeed, in the absence of mutations a homogeneous population (all individuals $A_1$ or $A_2$) cannot change through random mating. In the parlance of dynamics homogeneous populations correspond to *absorbing states*, where transitions are possible into the state but not away from it. In the presence of a single absorbing state, the steady state probability is one at this state, and zero for all other states. If there is more than one absorbing state, the steady state probability will be proportioned (split) among them.

In the absence of mutations, our models of reproducing populations have two absorbing states at $x = 0$ and $x = 1$. At long times, a population of fixed number either evolves to $x = 0$ with probability $\Pi_0$, or to $x = 1$ with probability $\Pi_1 = 1 - \Pi_0$. The value of $\Pi_0$ depends on the initial composition of the population that we shall denote by $0 < y < 1$, i.e. $p(x, t = 0) = \delta(x - y)$. Starting from this initial condition, we can follow the probability distribution $p(x, t)$ via the *forward Kolmogorov equation* (1.42). For purposes of finding the long-time behavior with absorbing states it is actually more convenient to express this as a *conditional probability* $p(x, t|y)$ that starting from a state $y$ at $t = 0$, we move to state $x$ at time $t$. Note that in any realization the variable $x(t)$ evolves from one time step to the next following the transition rates, but irrespective of its previous history. This type of process with no memory is called *Markovian,* after the Russian mathematician Andrey Andreyevich Markov (1856-1922). We can use this property to construct evolution equations for the probability by focusing on the change of position for the last step (as we did before in deriving Eq. (1.42)), or the first step. From the latter perspective, we can decompose the conditional probability after a time interval $t + dt$ as

$$p(x, t + dt|y) = \int d\delta_y \, R(\delta_y, y)dt \times p(x, t|y + \delta_y) + \left(1 - \int d\delta_y \, R(\delta_y, y)dt\right) p(x, t|y) \, , \quad (1.70)$$

where we employ the same parameterization of the reaction rates as in Eq. (1.37), with $\delta_y$ denoting the change in position. The above equation merely states that the probability to arrive at $x$ from $y$ in time $t + dt$ is the same as that of first moving away from $y$ by $\delta_y$ in the initial interval of $dt$, and then proceeding from $y + \delta_y$ to $x$ in the remaining time $t$ [the first term in Eq. (1.70)]. The second term corresponds to staying in place in the initial interval $dt$, and taking a trajectory that arrives at $x$ in the subsequent time interval $t$. (Naturally we have to integrate over all allowed intermediate positions.) Expanding left side of Eq. (1.70)

in $dt$, and the right side in $\delta_y$ (assuming dominance of local changes), gives

$$\begin{aligned} p(x,t|y) + dt\frac{\partial p(x,t|y)}{\partial t} &= p(x,t|y) + \left(\int d\delta_y\, R(\delta_y,y)dt - \int d\delta_y\, R(\delta_y,y)dt\right) p(x,t|y) \\ &+ \left(\int d\delta_y\, \delta_y R(\delta_y,y)dt\right)\frac{\partial p(x,t|y)}{\partial y} \\ &+ \frac{1}{2}\left(\int d\delta_y\, \delta_y^2 R(\delta_y,y)dt\right)\frac{\partial^2 p(x,t|y)}{\partial y^2} + \cdots . \end{aligned} \tag{1.71}$$

Using the definitions of drift and diffusion coefficients from Eqs. (1.43) and (1.44), we obtain

$$\frac{\partial p(x,t|y)}{\partial t} = v(y)\frac{\partial p}{\partial y} + D(y)\frac{\partial^2 p}{\partial y^2}\,, \tag{1.72}$$

which is known as the *backward Kolmogorov equation.* If the drift velocity and the diffusion coefficient are independent of position, the forward and backward equations are the same– more generally one is the *adjoint* of the other.

### 1.5.1 Fixation probability

Let us consider a general system with multiple absorbing states. Denote by $\Pi^*(x_a, y)$, the probability that a starting composition $y$ is at long time *fixed* to the absorbing state at $x_a$, i.e. $\Pi(x_a, y) = \lim_{t\to\infty} p(x_a, t|y)$. For the case of two possible alleles, we have two such states with $\Pi_0(y) \equiv \Pi^*(0, y)$ and $\Pi_1(y) \equiv \Pi^*(1, y)$, but we shall keep the more general notation for the time being. The functions $\Pi^*(x_a, y)$ must correspond to steady state solutions of Eq. (1.72), and thus obey

$$v(y)\frac{d\Pi^*(y)}{dy} + D(y)\frac{d^2\Pi^*(y)}{dy^2} = 0\,. \tag{1.73}$$

After rearranging the above equation to

$$\frac{\Pi^*(y)''}{\Pi^*(y)'} = \frac{d}{dy}\log\Pi^*(y)' = -\frac{v(y)}{D(y)}, \tag{1.74}$$

we can integrate it to

$$\log\Pi^*(y)' = -\int^y dy'\frac{v(y')}{D(y')} = -\ln\left(D(y')p^*(y)\right)\,, \tag{1.75}$$

and obtain

$$\Pi^*(y)' \propto -\int^y dy'[D(y')p^*(y')]^{-1}\,. \tag{1.76}$$

The result of the above integration is related to an intermediate step in calculation of the steady state solution $p^*$ of the forward Kolmogorov equation in (1.65). However, as we

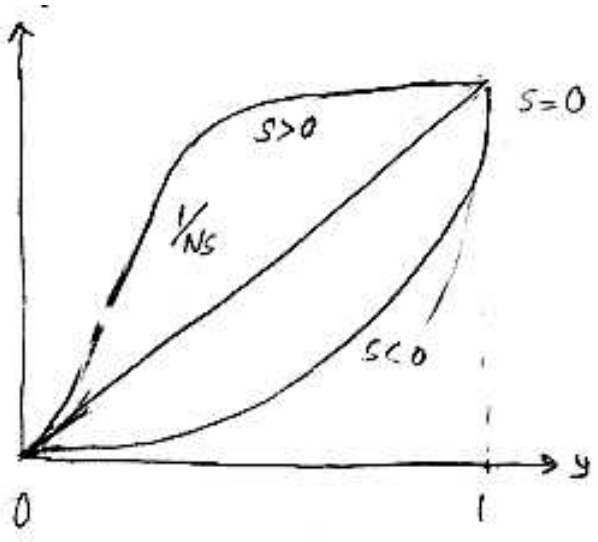



Figure 1: Fixation probability $\Pi_1(y)$ for different selection parameters $s$.

noted already, in the context of absorbing states the function $p^*$ is not normalizable and thus cannot be regarded as a probability. Nonetheless, we can choose to express the results in terms of this function. For example, the *probability of fixation*, i.e. $\Pi_1(y)$ is obtained with the boundary conditions $\Pi_1(0) = 0$ and $\Pi_1(1) = 1$, as

$$\Pi_1(y) = \frac{\int_0^y dy'[D(y)p^*(y')]^{-1}}{\int_0^1 dy'[D(y)p^*(y')]^{-1}}\,. \tag{1.77}$$

When there is selection, but no mutation, Eq. (1.59) implies

$$\log \Pi^*(y)' = -\int^y dy' \frac{v(y')}{D(y')} = -2\int^y (Ns) = -2Nsy + \text{constant}. \tag{1.78}$$

Integrating $\Pi^*(y)'$ and adjusting the constants of proportionality by the boundary conditions $\Pi_1(0) = 0$ and $\Pi_1(1) = 1$, then leads to the fixation probability of

$$\Pi_1(y) = \frac{1 - e^{-2Nsy}}{1 - e^{-2Ns}}\,. \tag{1.79}$$

The fixation probability of a neutral allele is obtained from the above expression in the limit of $s \to 0$ as $\Pi_1(y) = y$.

When a mutation first appears in a diploid population, it is present in only one copy and hence $y = 1/(2N)$. The probability that this mutation is fixed is $\Pi_1 = 1/(2N)$ as long as it is approximately neutral (if $2sN \ll 1$). If it is advantageous ($2sN \gg 1$) it will be fixed with probability $\Pi_1 = 1 - e^{-s}$ irrespective of the population size! If it is deleterious ($2sN \ll -1$) it will have a very hard time getting fixed, with a probability that decays with population size as $\Pi_1 = e^{-(2N-1)|s|}$. The *probability of loss* of the mutation is simply $\Pi_0 = 1 - \Pi_1$.

### 1.5.2 Mean times to fixation/loss

When there is an absorbing state in the dynamics, we can ask how long it takes for the process to terminate at such a state. In the context of random walks, this is known as the *first passage time*, and can be visualized as the time it takes for a random walker to fall into a trap. Actually, since the process is stochastic, the *time to fixation* (or loss) is itself a random quantity with a probability distribution. Here we shall compute an easier quantity, the mean of this distribution, as an indicator of a typical time scale for fixation/loss.

Let us consider an absorbing state at $x_a$, and the difference $p(x_a, t+dt|y) - p(x_a, t|y) = dt\partial p(x_a, t|y)/\partial t$. Clearly the probability to be at $x_a$ only changes due to absorption of particles, and thus $\partial p(x_a, t|y)/\partial t$ is proportional to the *probability density function (PDF) for fixation* at time $t$. *The conditional PDF* that the process terminates at $x_a$ must be properly normalized to unity, and must be divided by

$$\int_0^\infty dt \frac{\partial p(x_a, t|y)}{\partial t} = p(x_a, \infty|y) - p(x_a, 0|y) = \Pi^*(x_a, y) - 0 = \Pi^*(x_a, y)\,. \tag{1.80}$$

Thus the properly normalized conditional PDF for fixation at time $t$ at $x_a$ is

$$p_a(t|y) = \frac{1}{\Pi^*(x_a, y)} \frac{\partial p(x_a, t|y)}{\partial t}\,. \tag{1.81}$$

The *mean fixation time* is now computed from

$$\langle \tau(y) \rangle_a = \int_0^\infty dt\, t\, p_a(t|y) = \frac{1}{\Pi^*(x_a, y)} \int_0^\infty dt\, t\, \frac{\partial p(x_a, t|y)}{\partial t}\,. \tag{1.82}$$

Following Kimura and Ohta (1968)[3], we first examine the numerator of the above expression, defined as

$$T_a(y) = \lim_{T\to\infty} \int_0^T dt\, t\, \frac{\partial p(x_a, t|y)}{\partial t}\,. \tag{1.83}$$

(Rewriting $\lim_{T\to\infty} \int_0^T$ rather than simply $\int_0^\infty$ is for later convenience.) We can integrate this equation by parts to get

$$\begin{aligned} T_a(y) &= \lim_{T\to\infty} \left[ T p(x_a, T|y) - \int_0^T dt\, p(x_a, t|y) \right] \\ &= \lim_{T\to\infty} T\Pi^*(x_a, y) - \int_0^\infty dt\, p(x_a, t|y)\,. \end{aligned} \tag{1.84}$$

Let us denote the operations involved on the right-hand side of the backward Kolmogorov equation by the short-hand $\mathcal{B}_y$, i.e.

$$\mathcal{B}_y F(y) \equiv v(y) \frac{\partial F(y)}{\partial y} + D(y) \frac{\partial^2 F(y)}{\partial y^2}\,. \tag{1.85}$$

[3] M. Kimura and T. Ohta, Genetics **61**, 763 (1969).

Acting with $\mathcal{B}_y$ on both sides of Eq. (1.84), we find

$$\mathcal{B}_y T_a(y) = \lim_{T\to\infty} T\mathcal{B}_y \Pi^*(x_a, y) - \int_0^\infty dt\, \mathcal{B}_y p(x_a, t|y)\,. \tag{1.86}$$

But $\mathcal{B}_y \Pi^*(x_a, y) = 0$ according to Eq. (1.73), while $\mathcal{B}_y p(x_a, t|y) = \partial p(x_a, t|y)/\partial t$ from Eq. (1.72). Integrating the latter over time leads to

$$\mathcal{B}_y T_a(y) = -p(x_a, \infty|y) = -\Pi^*(x_a, y)\,. \tag{1.87}$$

For example, let us consider a population with no selection ($s = 0$), for which the probability to lose a mutation is $\Pi_0 = (1 - y)$. In this case, Eq. (1.87) reduces to

$$\frac{y(1-y)}{4N}\frac{\partial^2 T_0}{\partial y^2} = -(1-y), \Rightarrow \quad \frac{\partial^2 T_0}{\partial y^2} = -\frac{4N}{y}\,. \tag{1.88}$$

After two integrations we obtains

$$T_0(y) = -4Ny\,(\ln y - 1) + c_1 y + c_2 = -4Ny \ln y\,, \tag{1.89}$$

where the constants of integration are set by the boundary conditions $T_0(0) = T_0(1) = 0$, which follow from Eq. (1.83). From Eq. (1.82), we then obtain the mean time to loss of a mutation as

$$\langle \tau(y) \rangle_0 = -4N\frac{y \ln y}{1-y}\,. \tag{1.90}$$

A single mutation appearing in a diploid population corresponds to $y = 1/(2N)$, for which the mean number of generations to loss is $\langle \tau(y) \rangle_0 \approx 2\ln(2N)$. The mean time to fixation is obtained simply by replacing $y$ with $(1 - y)$ in Eq. (1.90) as

$$\langle \tau(y) \rangle_1 = -4N\frac{(1-y)\ln(1-y)}{y}\,. \tag{1.91}$$

The mean time for fixation of a newly appearing mutation ($y = 1/(2N)$) is thus $\langle \tau(y) \rangle_1 \approx (4N)$.

We can also examine the amount of time that the mutation survives in the population. The net probability that the mutation is still present at time $t$ is

$$S(t|y) = \int_{0^+}^{1^-} dx p(x, t|y)\,, \tag{1.92}$$

where the integrations exclude the absorbing points at 0 and 1. Conversely, the PDF that the mutation disappears (by loss *or* fixation) at time $t$ is

$$p_\times(t|y) = -\frac{dS(t|y)}{dt} = -\int_{0^+}^{1^-} dx \frac{dp(x, t|y)}{dt}\,. \tag{1.93}$$

(Note that the above PDF is properly normalized as $S(\infty) = 0$, while $S(0) = 1$.) The mean survival time is thus given by

$$\langle \tau(y) \rangle_\times = -\int_0^\infty dt\, t \int_{0^+}^{1^-} dx \frac{dp(x,t|y)}{dt} = \int_{0^+}^{1^-} dx \int_0^\infty dt\, p(x,t|y)\,, \tag{1.94}$$

where we have performed integration by parts and noted that the boundary terms are zero. Applying the backward Kolmogorov operator to both sides of the above equation gives

$$\begin{aligned} \mathcal{B}_y \langle \tau(y) \rangle_\times &= \int_{0^+}^{1^-} dx \int_0^\infty dt\, \mathcal{B}_y p(x,t|y) \\ &= \int_{0^+}^{1^-} dx \int_0^\infty dt \frac{dp(x,t|y)}{dt} \\ &= S(\infty|y) - S(0|y) = -1\,. \end{aligned} \tag{1.95}$$

In the absence of selection, we obtain

$$\frac{y(1-y)}{4N} \frac{\partial^2 \langle \tau(y) \rangle_\times}{\partial y^2} = -1 \;\Rightarrow\quad \frac{\partial^2 \langle \tau(y) \rangle_\times}{\partial y^2} = -4N \left( \frac{1}{y} + \frac{1}{1-y} \right)\,. \tag{1.96}$$

After two integrations we obtains

$$\langle \tau(y) \rangle_\times = -4N \left[ y \ln y + (1-y) \ln(1-y) \right]\,, \tag{1.97}$$

where the constants of integration are set by the boundary conditions $\langle \tau(0) \rangle_\times = \langle \tau(1) \rangle_\times = 0$. Note the interesting relation

$$\langle \tau(y) \rangle_\times = \Pi_0(y) \langle \tau(y) \rangle_0 + \Pi_1(y) \langle \tau(y) \rangle_1\,, \tag{1.98}$$

which is easily generalized to any number of absorbing states. By adding and subtracting the contribution of absorbing sites to the positional integral in Eq. (1.94), we obtain

$$\langle \tau(y) \rangle_\times = -\int_0^\infty dt\, t \left[ \int dx \frac{dp(x,t|y)}{dt} - \sum_a \frac{dp(x_a,t|y)}{dt} \right]\,. \tag{1.99}$$

By taking the time derivative over $t$ outside the integration over $x$, we get

$$\langle \tau(y) \rangle_\times = -\int_0^\infty dt\, t \frac{\partial}{\partial t} \left( \int dx\; p(x,t|y) \right) + \sum_a \frac{dp(x_a,t|y)}{dt}\,. \tag{1.100}$$

The first term is zero since the integral over $x$ is always unity, and from Eq. (1.82) we obtain

$$\langle \tau(y) \rangle_\times = \sum_a \Pi^*(x_a, y) \langle \tau(y) \rangle_a\,. \tag{1.101}$$

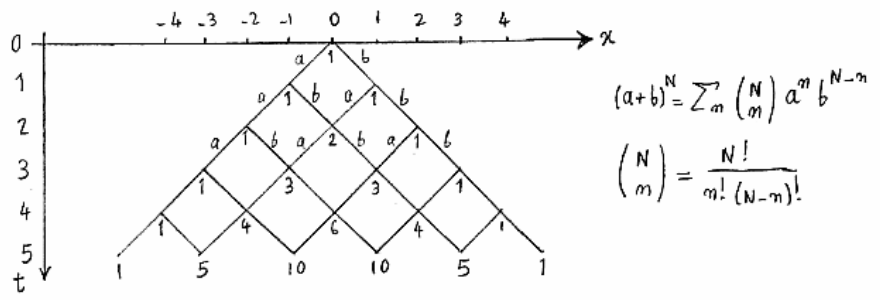
## 1.6 Sequence alignment

The dramatic increase in the number of sequenced genomes and proteomes has lead to development of quite a few *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 surmising the function of a sequenced protein.

The *explicit inputs* are two (or more) sequences (of 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 that 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 (or several such matches), and a corresponding score $S$. An important question is whether this output is due to a meaningful relation (homology) 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 the 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.

### 1.6.1 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

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

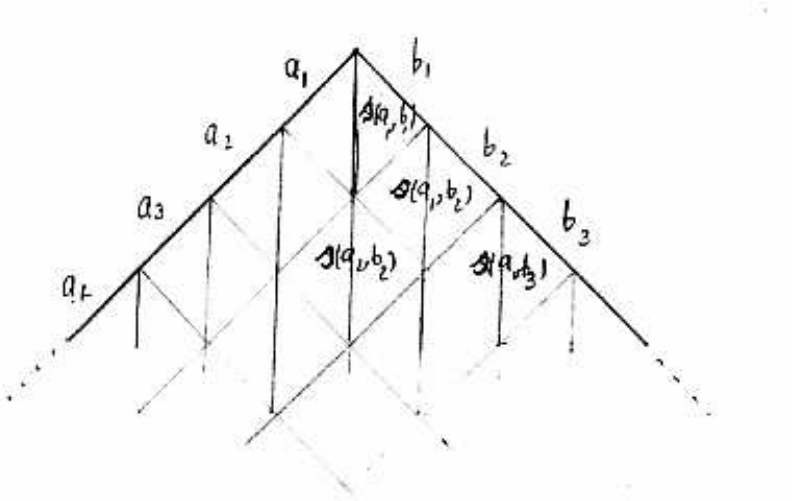
of lengths $m$ an $n$ respectively. We can define a matrix of alignment scores

$$S_{ij} \equiv \text{Score of best alignment terminating on } a_i, b_j. \tag{1.102}$$

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

$$S_{ij} = S_{i-1,j-1} + s(a_i, b_j), \tag{1.103}$$

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 $\pm 45°$. 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 build an analogy with a dynamical system, we now introduce new coordinates $(x, t)$ by

$$x = j - i, \;\; t = i + j. \tag{1.104}$$

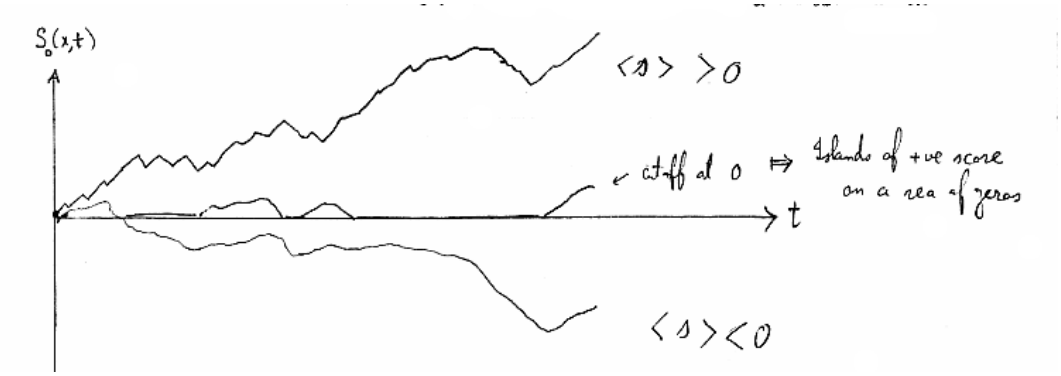
The recursion relation in Eq. (1.103) is now recast as

$$S(x, t) = S(x, t - 2) + s(x, t)\,, \tag{1.105}$$

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 include also jumps between columns. For a global (*Needleman–Wunsch*) alignment (including gaps only at beginning or end), 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 distributed, 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. (1.105). 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.



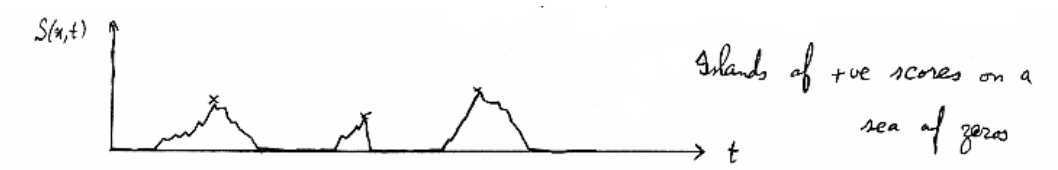

To identify matching segments shorter than a complete sequence, we can employ the local (*Smith-Waterman*) alignment scheme in which negative scores are cut off according to

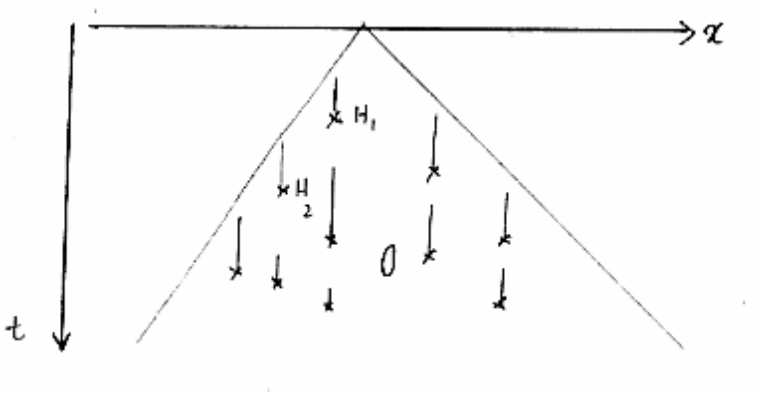
$$S_{ij} = \max\{S_{i-1,j-1} + s(a_i, b_j), 0\}, \tag{1.106}$$

or in the notation of $x$ and $t$,

$$S(x,t) = \max\{S(x,t-2) + s(x,t-1), 0\}. \tag{1.107}$$

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 $H_\alpha$ the peak value of the score on island $\alpha$. 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

$$S = \max_{\alpha=1,\cdots,K}\{H_\alpha\}. \tag{1.108}$$

To assign significance to a match, we need to know how likely it is that a score $S$ 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.

$$K \propto mn. \tag{1.109}$$

Equation (1.108) 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

$$X = \max\{H_1, H_2, \ldots, H_K\}\,. \tag{1.110}$$

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

$$\begin{aligned} P_K(X \leq S) &= \text{Prob.}(H_1 \leq S) \times \text{Prob.}(H_2 \leq S) \times \cdots \times \text{Prob.}(H_K \leq S) \\ &= \left[\int_{-\infty}^{S} dH p(H)\right]^K \\ &= \left[1 - \int_{S}^{\infty} dH p(H)\right]^K . \end{aligned} \tag{1.111}$$

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

$$P_K(S) \approx \exp\left[-K \int_{S}^{\infty} dH p(H)\right]. \tag{1.112}$$

Assume, as we shall demonstrate shortly, that $p(H)$ falls exponentially in its tail, as $ae^{-\lambda H}$. We can then write

$$\int_S^\infty dHp(H) = \int_S^\infty dHae^{-\lambda H} = \frac{a}{\lambda}e^{-\lambda H}. \tag{1.113}$$

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

$$P_K(S) = \exp\left[-\frac{Ka}{\lambda}e^{-\lambda S}\right], \tag{1.114}$$

with a corresponding PDF of

$$p_K(S) = \frac{dP_K(S)}{dS} = Ka\exp\left(-\lambda S - \frac{Ka}{\lambda}e^{-\lambda S}\right). \tag{1.115}$$

The exponent

$$\phi(S) \equiv -\lambda S - \frac{Ka}{\lambda}e^{-\lambda S}, \tag{1.116}$$

has an extremum when

$$\frac{d\phi}{dS} = -\lambda + Kae^{-\lambda S^*} = 0, \tag{1.117}$$

corresponding to a score of

$$S^* = \frac{1}{\lambda}\log\left(\frac{Ka}{\lambda}\right). \tag{1.118}$$

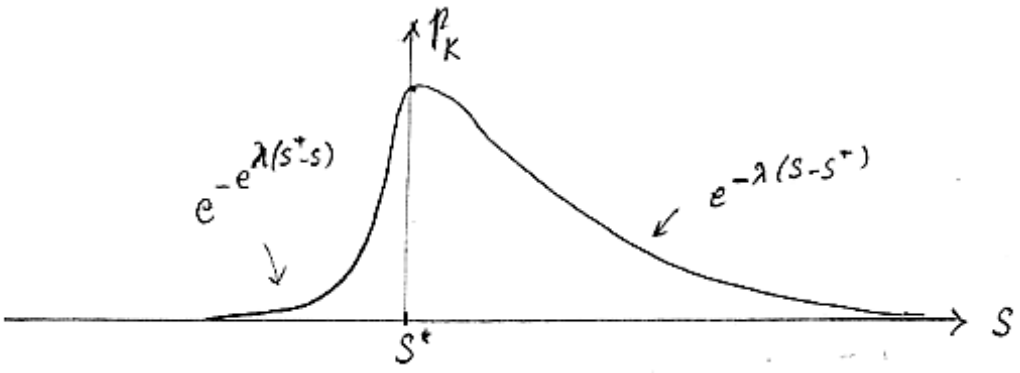
Equation (1.118) gives the most probable value for the score, in terms of which we can re-express the PDF in Eq. (1.115) as

$$p_K(S) = \lambda\exp\left[-\lambda(S-S^*) - e^{-\lambda(S-S^*)}\right]. \tag{1.119}$$

Evidently, once we determine the most likely score $S^*$, the rest of the probability distribution is determined by the single parameter $\lambda$. The PDF in Eq. (1.119) 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 $\lambda$, and indeed to confirm the assumption in Eq. (1.113) that $p(H)$ decays exponentially at large $H$. The height profile of an island (by definition positive) evolves according to Eq. (1.105). 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 of jumps that render $S$ negative). Thus the probability for a height (island score) $h$ evolves according to

$$\begin{aligned} p(h,t) &= \sum_s p_s p(h-s, t-2) \\ &= \sum_{a,b} p_a p_b p[h - s(a,b), t-2], \end{aligned} \tag{1.120}$$




where the second equality is obtained by assuming that the characters are chosen randomly with 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

$$p^*(h) \propto e^{-\lambda h}. \tag{1.121}$$

Substituting this ansatz into Eq. (1.120), we obtain

$$e^{-\lambda h} = \sum_{a,b} p_a p_b e^{-\lambda(h - s(a,b))}. \tag{1.122}$$

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 $\lambda$

$$\boxed{\sum_{a,b} p_a p_b e^{\lambda s(a,b)} = 1.} \tag{1.123}$$

Equation (1.121) 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. (1.119) the maximization will not change the tail of the distribution. Hence the value of $\lambda$ in Eq. (1.123), along with $S^*$ completely characterizes the Gumbel distribution for statistics of random local gapless alignments. (In addition to the trivial solution $\lambda = 0$, Eq. (1.123) has a unique positive solution.) This expression was first derived in by Karlin and Altschul in 1990.[4]

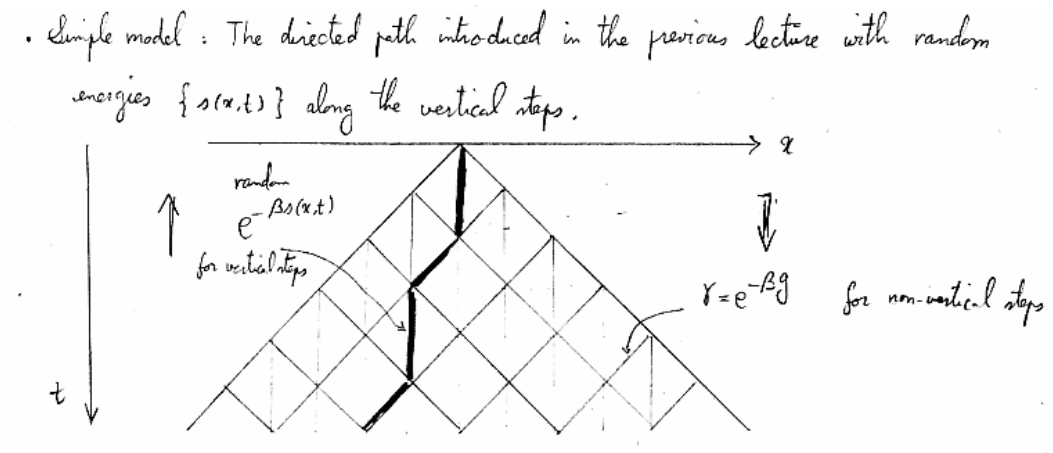
### 1.6.2 Gapped alignments

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

[4] *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes*, S. Karlin AND S.F. Altschul, Proc. Natl. Acad. Sci. USA **87**, 2264-2268 (1990).

gaps and insertions (more so for further 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 can again be justified by noting that local alignments rely on selecting the best score amongst a large number. The shape of the 'islands' is now somewhat different, and their statistics (alluded to below) is considerably harder to obtain.

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.




In the spirit of statistical physics, we may even introduce a fuzzier version of alignment corresponding to a finite temperature $\beta^{-1}$. 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

$$W(x,t) = \text{sum of all paths' Boltzmann weights from } (0,0) \text{ to } (x,t). \tag{1.124}$$

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

$$W(x,t) = e^{\beta s(x,y)} W(x,t-2) + e^{-\beta g} \left[ W(x+1,t-1) + W(x-1,t-1) \right]. \tag{1.125}$$

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. (1.125) is dominated by the largest term, leading to $(W(x,t) = \exp[\beta S(x,t)])$, with

$$S(x,t) = \max\{S(x,t-2) + s(x,t), S(x+1,t-1) - g, S(x-1,t-1) - g\}. \qquad (1.126)$$

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