transition phenomena for optimal gapped alignment scores. Section 7.3.2 introduces two methods for estimating the key parameters of the distribution. Section 7.3.3 lists the empirical values of these parameters for BLOSUM and PAM matrices.

In Section 7.4, we describe how P-value and E-value (also called Expect value) are calculated for BLAST database search.

Finally, we conclude the chapter with the bibliographic notes in Section 7.5.

7.1 Introduction

In sequence similarity search, homology relationship is inferred based on P-values or its equivalent E-values. If a local alignment has score s, the P-value gives the probability that a local alignment having score s or greater is found by chance. A P-value of 10^{-5} is often used as a cutoff for BLAST database search. It means that with a collection of random query sequences, only once in a hundred thousand of instances would an alignment with that score or greater occur by chance. The smaller the P-value, the greater the belief that the aligned sequences are homologous. Accordingly, two sequences are reported to be homologous if they are aligned extremely well.

Extremes are rare events that do not happen very often. In the 1950s, Emil Julius Gumbel, a German mathematician, proposed new extreme value distributions. These distributions had quickly grown into the extreme value theory, a branch of statistics, which finds numerous applications in industry. One original distribution proposed by Gumbel is the extreme value type-I distribution, whose distribution function is

$$\Pr[S \geq s] = 1 - \exp(-e^{-\lambda(s-u)}), \tag{7.1}$$

where u and λ are called the location and scale parameters of this distribution, respectively. The distribution defined in (7.1) has probability function

$$f(x) = \lambda \exp(-\lambda(x-u) - e^{-\lambda(x-u)}).$$

Using variable substitution

$$z = e^{-\lambda(x-u)},$$

we obtain its mean and variance as

$$\begin{aligned} \mu &= \lambda \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_0^{\infty} (u - \ln(z)/\lambda) e^{-z} dz \\ &= u \int_0^{\infty} e^{-z} dz - (1/\lambda) \int_0^{\infty} \ln(z) e^{-z} dz \\ &= u + \gamma/\lambda, \end{aligned} \tag{7.2}$$

and

$$\begin{aligned} V &= \lambda \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 \\ &= \int_0^{\infty} (u - \ln(z)/\lambda)^2 e^{-z} dz - (u + \gamma/\lambda)^2 \\ &= \pi^2 \lambda^2 / 6, \end{aligned} \tag{7.3}$$

where γ is Euler's constant $0.57722\ldots$.

Both theoretical and empirical studies suggest that the distributions of optimal local alignment scores S with or without gaps are accurately described by an extreme value type-I distribution. To given this an intuitive account, we consider a simple alignment problem where the score is 1 for matches and $-\infty$ for mismatches. In this case, the optimal local ungapped alignment occurs between the longest exact matching segments of the sequences, after a mismatch. Assume matches occur with probability p. Then the event of a mismatch followed by k matches has probability $(1-p)p^k$. If two sequences of lengths n and m are aligned, this event occurs at nm possible sites. Hence, the expected number of local alignment with score k or more is

$$a = mn(1-p)p^k.$$

When k is large enough, this event is a rare event. We then model this event by the Poisson distribution with parameter a. Therefore, the probability that there is a local alignment with score k or more is approximately

$$1 - e^{-a} = 1 - \exp\left(-mn(1-p)p^k\right).$$

Hence, the best local alignment score in this simple case has the extreme value type-1 distribution (7.1) with

$$u = \ln\left(mn(1-p)\right) / \ln(1/p)$$

and

$$\lambda = \ln(1/p).$$

In general, to study the distribution of optimal local ungapped alignment scores, we need a model of random sequences. Through this chapter, we assume that the two aligned sequence are made up of residues that are drawn independently, with respective probabilities p_i for different residues i. These probabilities (p_i) define the *background frequency distribution* of the aligned sequences. The score for aligning residues i and j is written s_{ij}. Under the condition that the expected score for aligning two randomly chosen residues is negative, i.e.,

$$E(s_{ij}) = \sum_{i,j} p_i p_j s_{ij} < 0, \tag{7.4}$$

the optimal local ungapped alignment scores are proved to approach an extreme value distribution when the aligned sequences are sufficiently long. Moreover, simple formulas are available for the corresponding parameters λ and u.

Fig. 7.1 The accumulative score of the ungapped alignment in (7.7). The circles denote the ladder positions where the accumulative score is lower than any previously reached ones.

The scale parameter λ is the unique positive number satisfying the following equation (see Theorem B.1 for its existence):

$$\sum_{i,j} p_i p_j e^{\lambda s_{ij}} = 1. \tag{7.5}$$

By (7.5), λ depends on the scoring matrix (s_{ij}) and the background frequencies (p_i). It converts pairwise match scores to a probabilistic distribution $\left(p_i p_j e^{\lambda s_{ij}}\right)$.

The location parameter u is given by

$$u = \ln(Kmn)/\lambda, \tag{7.6}$$

where m and n are the lengths of aligned sequences and $K < 1$. K is considered as a space correcting factor because optimal local alignments cannot locate in all mn possible sites. It is analytically given by a geometrically convergent series, depending only on the (p_i) and (s_{ij}) (see, for example, Karlin and Altschul, 1990, [100]).

7.2 Ungapped Local Alignment Scores

Consider a fixed ungapped alignment between two sequences given in (7.7):

$$\begin{array}{c} a\ g\ c\ g\ c\ c\ g\ g\ c\ t\ \ t\ a\ t\ t\ c\ t\ \ t\ g\ c\ g\ c\ t\ g\ c\ a\ c\ c\ g \\ |\ |\ \ \ |\ |\ \ \ |\ |\ |\ \ \ \ \ \ \ |\ \ \ \ \ \ \ |\ \ \ \ \ \ \ |\ |\ \ \ |\ |\ |\ |\ |\ | \\ a\ g\ t\ g\ c\ g\ g\ g\ c\ g\ a\ t\ t\ c\ t\ g\ c\ g\ t\ c\ c\ t\ c\ c\ a\ c\ c\ g \end{array} \tag{7.7}$$

We use s_j to denote the score of the aligned pair of residues at position j and consider the accumulative score

$$S_k = s_1 + s_2 + \cdots + s_k, \ \ k = 1, 2, \ldots.$$

Starting from the left, the accumulative score S_k is graphically represented in Figure 7.1.