Calculating the Secrets of Life

Applications of the Mathematical Sciences in Molecular Biology

Eric S. Lander and
Michael S. Waterman, Editors

Committee on the Mathematical Sciences in
Genome and Protein Structure Research

Board on Mathematical Sciences

Commission on Physical Sciences, Mathematics, and Applications

National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1995

Chapter 4
Hearing Distant Echoes: Using Extremal Statistics to Probe Evolutionary Origins

Michael S. Waterman
University of Southern California

The comparison of DNA and protein sequences provides a powerful tool for discerning the function, structure, and evolutionary origin of important macromolecules. Sequence comparison sometimes reveals striking matches between molecules that were hitherto not known to be related—immediately suggesting hypotheses that can be tested in the laboratory. In other cases, sequence comparison reveals only weak similarities. In such instances, statistical theory is essential for interpreting the significance of such matches. The author discusses large deviation theory for sequence matching and applies it to evaluate a tantalizing report concerning distant echoes from the earliest period in the origin of life.

As soon as new deoxyribonucleic acid (DNA) or protein sequences are determined, molecular biologists immediately examine them for clues about their biological significance. A number of important questions about the function of a newly determined protein are often asked, including the following. What can be inferred about the function of a new protein on the basis of its amino acid sequence? Can one discern the reactions it catalyzes or the molecules it binds? What three-dimensional shape will the linear amino acid sequence of a protein assume when it folds up according to the laws of thermodynamics? Another class of questions concerns the evolutionary relationships between known sequences. For example, some questions concerning

hemoglobin are the following. What is the evolutionary relationship between three related α, β, and γ hemoglobin genes? What is the evolutionary relationship between the hemoglobin molecules from various organisms? What do these sequences tell us about the evolutionary history of humans, chimpanzees, and gorillas? Each of these questions can be approached, if not always entirely solved, by sequence comparison.

Sequence comparison is of tremendous interest to molecular biologists because it is becoming easy to determine DNA and protein sequences, whereas it remains difficult to determine molecular structure or function by experimental means. Thus, functional and structural clues from sequence analysis can save years of work at the laboratory bench. An important early example illustrates the point. Some years ago, molecular biologists compared the protein sequence encoded by a cancer-causing gene (or oncogene) called *v-sis* to the available database of protein sequences. Remarkably, a computer search revealed that the sequence showed more than 90 percent identity to the sequence of a previously discovered gene encoding a growth-stimulating molecule, called platelet-derived growth factor (PDGF). Instantly, cancer researchers had a precise hypothesis about how this oncogene causes unregulated cell growth. Subsequent experiments confirmed the guess.

Nowadays, molecular biologists routinely carry out such computer searches against the current databases (which now contain both protein and DNA sequences) and are rewarded with striking and suggestive matches at a high frequency (perhaps 20 to 30 percent for a new gene). In some cases, the matches extend across the entire length of the protein. In other cases, there is a strong match across a restricted domain—examples include particular sequences at the catalytic site of enzymes that hydrolyze adenosinetriphosphate (ATP) or at the DNA-binding site of proteins that regulate the activity of genes. The frequency with which such strong matches are found is a tribute to the tremendously conservative nature of evolution: many of the basic building blocks of proteins and DNA have been reused in hundreds of different ways.

For the majority of new sequences, however, there is no striking match in the database. Although this may change with time (some molecular biologists believe that there are only a few thousand or a few tens of thousands of basic architectural motifs for proteins and that it is

just a matter of time before we collect them all), computer searches will turn up only weak similarities. Before attempting to read biological significance into such weak similarities, one must evaluate their statistical significance. Not surprisingly, this is an area in which mathematics has much to offer molecular biology. To motivate the study of the statistical significance of sequence similarities, we consider a single data set that provoked a great deal of excitement a few years ago when a team of researchers thought that they saw extraordinary clues about early evolution in the sequences of genes encoding certain ribonucleic acid (RNA) molecules.

The origin of the universe and the origin of life are topics of wide interest to both biologists and nonbiologists. One approach to studying the origin of the universe is to listen to faint echoes from the Big Bang. Similar approaches are used in studying the origin of life. Are there any molecular echoes remaining from the origin of life? Each of the three key molecules in molecular biology—DNA, RNA, and protein—has been championed by some theorists as the earliest self-replicating molecule. Proteins have seemed attractive to some because of their ability to catalyze chemical reactions. DNA has seemed attractive to others because it is a stable store of information. Lately, however, RNA has taken the lead based on the well-known ability of RNA to encode information in the same manner as DNA and the recently discovered ability of RNA to act as nonprotein enzymes that are able to catalyze some chemical reactions. These properties suggest that some RNA sequence might have been able to achieve the key feat of self-replication—serving as both self-template and replication enzyme. Thus, life may have started out as an RNA world.

As indicated in Chapter 1, modern RNAs come in three varieties: messenger RNAs (mRNAs), ribosomal RNAs (rRNAs), and transfer RNAs (tRNAs). mRNAs are the messages copied from genes. rRNAs are components of the macromolecular structure, called the ribosome, used for translating RNA sequences into protein sequences. tRNAs are the "adapter molecules" that read the genetic code, with an anticodon loop recognizing a particular codon at one end and an attachment site for the amino acid corresponding to this codon at the other. rRNAs and tRNAs are clearly ancient inventions, necessary for the progression from life based only on RNA to organisms employing proteins for efficient catalysis of biochemical reactions.

In the early 1980s David Bloch and colleagues reported that they had found that these two types of RNA—tRNA and rRNA—had significant sequence similarities implying a common evolutionary ancestry (Bloch et al., 1983). In his paper, Bloch reported:

> Many tRNAs of *E. coli* and yeast contain stretches whose base sequences are similar to those found in their respective rRNAs. The matches are too frequent and extensive to be attributed to coincidence. They are distributed without discernible pattern along and among the RNAs and between the two species. They occur in loops as well as in stems, among both conserved and non-conserved regions. Their distributions suggest that they reflect common ancestral origins rather than common functions, and that they represent true homologies.

Such tantalizing arguments should be grounded in statistics—since we cannot test the origin of life by direct experiment (as we could test a proposed function for a protein based on sequence similarity). In this chapter, we develop some tools for evaluating statistical significance and apply them to Bloch's data.

The biological hypothesis that relationships between the RNAs are true homologies is necessarily imprecise. The evidence given is frequent and extensive matchings of stretches of sequences between the molecules, just the sort of matchings that the local algorithm presented below is designed to find. To "test" the biological hypothesis, we form a statistical hypothesis that the sequences are generated with independent and identically distributed letters. Then we test this hypothesis by computing scores using the local algorithm. Since letters in real sequences are not independent, it is possible to change the hypothesis to a Markov hypothesis, for example. This does not change the score distribution very much for the distributions obtained from real sequences. If the score distribution is consistent with that from comparison of random sequences, we would fail to reject the statistical hypothesis and thus have evidence against the biological hypothesis. If, on the other hand, the scores are frequently too large, showing strongly matching stretches or intervals of sequence, we have evidence for the biological hypothesis and against the statistical hypothesis.

Statistical questions are increasingly important in molecular biology. While statistical significance is not directly related to biological

importance, it is a good indicator and can lead to the formulation of important biological hypotheses, as noted above. Conversely, lack of statistical significance is an important clue in considering whether to reject a relationship that may seem interesting to the human eye. With over 70,000 sequences in modern databases, molecular biologists require an automatic way to reject all but the most interesting results from a database search. Comparing one sequence to the database involves 70,000 comparisons. Comparing all pairs of sequences involves $\binom{70{,}000}{2}$, or about 2.4×10^9, comparisons. As we will see with the tRNA and rRNA comparison, even a small number of comparisons can raise subtle questions.

GLOBAL SEQUENCE COMPARISONS

We will now discuss a number of situations for sequence comparisons and some probability and statistics that can be applied to these problems. Some powerful and elegant mathematics has been developed to treat this class of problems. Our discussion will naturally break into two parts, global comparisons and local comparisons.

Sequence Alignment

In this section we study the comparison of two sequences. For simplicity the two sequences $A_1A_2\ldots A_n$ and $B_1B_2\ldots B_m$ will consist of letters drawn independently with identical distribution from a common alphabet.

Sequences evolve at the molecular level by several mechanisms. One letter, A for example, can be substituted for another, G for example. These events are called substitutions. Letters can be removed from or added to a sequence, and these events are called deletions or insertions. Given two sequences such as ATTGCC and ACGGC, it is usually not clear how they should be related. The possible relationships are often written as alignments such as:


```
ATTGCC
ACGG-C
```

or

```
ATTGCC
-ACGGC
```

where in the first case there are three identities, two substitutions, and one insertion/deletion (indel) and in the second case there are two identities, three substitutions, and one indel.

An alignment can be obtained by inserting gaps ("–") into the sequences so that

$$A_1A_2\ldots A_n \rightarrow A_1^*A_2^*\ldots A_L^*$$

and

$$B_1B_2\ldots B_m \rightarrow B_1^*B_2^*\ldots B_L^*.$$

Here the subsequence of all $A_i^* \neq$ "–" is identical to $A_1A_2\ldots A_n$. Then, since the *-sequences have equal length, A_i^* is aligned with B_i^*. In Chapter 3, algorithms to achieve optimal alignments are discussed. Here we are interested in the statistical distribution of these scores, not in how they are obtained. Global alignments refer to the situation where all the letters of each sequence must be accounted for in the alignments. There are two types of global alignments, alignments where the pairing is given and alignments where the pairing is not given.

Alignment Given

In this section, we assume the alignment is given with the sequences:

$$\begin{matrix} A_1A_2\ldots A_n \\ B_1B_2\ldots B_n. \end{matrix}$$

(Gaps "–" have been added so that these sequences both have the same length—L in the previous section, n here—and the stars have been omitted to simplify the notation.) In this case the alignment is given and

therefore cannot be optimized. We give the statistical distribution of the alignment score for completeness, however. Let $s(A,B)$ be a real valued random variable. Define the score S by

$$S = \sum_{i=1}^{n} s(A_i, B_i),$$

and let $\mathbf{E}(S)$ denote the expectation of S and Var (S) denote the variance. Clearly, $\mathbf{E}(S) = n\mathbf{E}(s(A,B))$ and

$$\mathrm{Var}(S) = n\,\mathrm{Var}(s(A,B)).$$

Since S is the sum of independent, identically distributed random variables $s(A,B)$, the central limit theorem implies that for large n

$$S \approx \mathrm{Normal}(n\,\mathbf{E}(s(A,B)), n\,\mathrm{Var}(s(A,B))).$$

If $s(A,B) \in \{0,1\}$, then S is binomial (n,p), where $p = \mathbf{P}\{s(A,B) = 1\}$. Even when the letters are not identically distributed, the limiting distribution is normal under mild assumptions (Chung, 1974).

Alignment Unknown

The assumptions of the last section are carried over: $A_1A_2\ldots A_n$ and $B_1B_2\ldots B_m$ are composed of independent and identically distributed letters and $s(A,B)$ is a real valued random variable on pairs of letters. We extend $s(\cdot,\cdot)$ to $s(A,-)$ and $s(-,B)$ so that deletions are included. We assume that the value of s for all deletion scores is smaller than $\max s(A,B)$. An alignment score S is the maximum over all possible alignments

$$S = \max_{\text{alignments}} \sum_{i=1}^{L} s(A_i^*, B_i^*).$$

The optimization destroys the classical normal distribution of alignment score, but an easy application of a beautiful theorem known as Kingman's subadditive ergodic theorem gives an interesting result:

Theorem 4.1 (Kingman, 1973) *For s,t nonnegative integers with $0 \le s \le t$, let $X_{s,t}$ be a collection of random variables that satisfy*

(i) Whenever $s < t < u$, $X_{s,u} \le X_{s,t} + X_{t,u}$,

(ii) The joint distribution of $\{X_{s,t}\}$ is the same as that of $\{X_{s+1,t+1}\}$,

(iii) The expectation $g_t = \mathbf{E}(X_{0,t})$ exists and satisfies $g_t \ge -Kt$ for some constant K and all $t > 1$.

Then the finite $\lim_{t\to\infty} X_{0,t}/t = \lambda$ exists with probability 1 and in the mean.

The essential assumption is (ii), the subadditivity condition. To motivate and illustrate this theorem, recall the strong law of large numbers (SLLN), which treats independent, identically distributed (iid) random variables $W_1, W_2, \ldots$ with $\mu = \mathbf{E}(W_i)$. The SLLN asserts that

$$\frac{W_1 + W_2 + \ldots + W_n}{n} \to \mu$$

with probability 1.

It is easy to see that additivity holds. Set

$$U_{s,t} = \sum_{s+1 \le i \le t} W_i.$$

Of course (i) is satisfied:

$$U_{s,u} = \sum_{s+1 \le i \le t} W_i + \sum_{t+1 \le i \le u} W_i$$

$$= U_{s,t} + U_{t,u}.$$

Since the W_i are iid, (ii) is evidently true. Finally, $g(t) = \mathbf{E}(U_{0,t}) = t\mu$, so that (iii) holds with $\mu = -K$. Therefore the limit

$$\lim_{t\to\infty} \sum_{1\le i\le t} W_i / t$$

exists and is constant with probability 1. Notice that this setup does not allow us to conclude that the limit is μ. This is a price of relaxing the assumption of additivity.

Returning to the statistical distribution of alignment score, recall that an alignment score S is the maximum over all possible alignments

$$S = \max_{\text{alignments}} \sum_{i=1}^{L} s(A_i^*, B_i^*) .$$

Define $X_{s,t}$ by

$$-X_{s,t} = \text{ score of } A_{s+1}A_{s+2}\dots A_t \text{ vs } B_{s+1}B_{s+2}\dots B_t .$$

Then evidently,

$$-X_{s,u} \ge (-X_{s,t}) + (-X_{t,u})$$

and

$$X_{s,u} \le X_{s,t} + X_{t,u} .$$

We have that $g_t = \mathbf{E}(X_{0,t})$ exists since the expectation of a single alignment exists and $-X_{0,t}$ is the maximum of a *finite* number of alignment scores. The final hypothesis to check is $g_t \ge -Kt$ for some constant K and all $t > 1$. Clearly,

$$\mathbf{E}(-X_{0,t}) \le t \max s(A,B)$$

so that

$$g_t \ge -(\max s(A,B))t = -Kt .$$

Our conclusion is that

$$\lim_{t\to\infty} X_{0,t} / t = \lambda$$

exists with probability 1 and in the mean. Therefore optimal alignment score grows linearly with sequence length. Obviously, $\lambda \ge \mathbf{E}(s(A,B))$.

In the simplest case of interest, the alphabet has two uniformly distributed letters and $s(A,B) = 0$ if $A \ne B$ and $s(A,A) = s(B,B) = 1$. The alignment score is known as the *longest common subsequence*, and Chvátal and Sankoff (1975) wrote a seminal paper on this problem in the 1970s. In spite of much effort since then, λ remains undetermined. Deken (1979) gives bounds for λ: $0.7615 \le \lambda \le 0.8602$. Without alignment the fraction of matching letters is 0.5 by the strong law of large numbers. Not too much is known about the variance either, although Steele (1986) proves it is $O(n)$.

LOCAL SEQUENCE COMPARISONS

Alignment Given

Consider many independent throws of a coin with probability p of heads, where $0 < p < 1$. For any p, there will be stretches where the coin comes up heads every time. What is the distribution of the length of the longest of these head runs? This maximum length is known as a "local" score; while it is a global maximum, it is a function only of the nearby tosses. A related problem is to consider sequence $A_1A_2\dots A_n$ of letters chosen independently and from a common alphabet, {A, C, G, T} for DNA for example. The letters A and G are known as purines (R), and C and T are known as pyrimidines (Y). A two-letter alphabet is natural when grouping nucleotides by chemical similarity. In fact, there is a hypothesis that the first nucleic acid sequences were made up of just two elements, R and Y. It is natural to ask how random the distribution of R and Y is for a given sequence. We will study how large is R_n, the length of the longest run of purines R in a sequence of length n. Here an occurrence of R is a "head" for the coin tossing analogy.

The coin tossing question was considered by Erdös and Rényi (1970), who found the strong law

$$\lim_{n\to\infty} \frac{R_n}{\log_{1/p} n} = 1 \text{ with probability 1.} \qquad (4.1)$$

Of course, one may desire more detailed information about R_n. For an example, we look at "16S" rRNA from *E. coli,* which is 1,541 letters in length and is known by its sedimentation rate S of 16. (The sedimentation rate is an indication of mass: the greater the mass, the higher the rate of sedimentation and the larger the value of S.) Equation (4.1) tells us that we would typically see about $\log_{1/p} 1{,}541 = 10.6$ R's in a row where $p = \frac{1}{2}$. What if we have a head run of length 14 in our 16S sequence? Is this score extreme enough to be of note? For the statistical question of significance, we need to have a way to compute such probabilities. This is supplied by Poisson approximation.

For an appropriately chosen test length t, we see an R run of length t begins at a given position with a small probability. Since the number of positions where such a run could occur is large, the number of long head runs should be approximately Poisson. Our discussion about the mathematics behind this intuition follows Goldstein (1990).

This intuition is almost correct. One must first, however, adjust for the fact that runs of heads occur in "clumps"; that is, if there is a run of heads of length t beginning at position α, then with probability p there will also be a run of heads of length t beginning at position $\alpha+1$, with probability p^2 a run of heads of length t beginning at position $\alpha+2$, and so forth. Hence, the total number of runs of length t or more is seen to have a compound Poisson distribution. By counting only the first such run in every clump, the occurrences now counted are no longer clumped and their number is approximately Poisson. This is an example, with average clump size $1+p+p^2\ldots = 1/(1-p)$, of the "Poisson clumping heuristic," as described by Aldous (1989). Using the fact that having no runs of length t is equivalent to having the longest head run shorter than t, we can approximate the distribution of the length of the longest run of heads. In the remainder of this section we explore the approximation of this distribution by the Poisson distribution.

Let I be an index set, and for each $\alpha \in I$, let X_α be an indicator random variable, that is, $X_\alpha = 1$ if an event occurs and $X_\alpha = 0$ if the event does not occur. The total number of occurrences of events can be expressed as

$$W = \sum_{\alpha\in I} X_\alpha .$$

It seems intuitive that if $p_\alpha = \mathbf{P}(X_\alpha = 1)$ is small, and $|I|$, the size of the index set, is large, then W should be approximately Poisson distributed. Certainly this is true when all the $X_\alpha, \alpha \in I$, are independent. In the case of dependence, it seems plausible that the same approximation should hold when dependence is somewhat confined. For each α, we let B_α be the set of dependence for α; that is, for each $\alpha \in I$, assume we are given a set $B_\alpha \subset I$ such that

$$X_\alpha \text{ is independent of } X_\beta, \beta \notin B_\alpha . \qquad (4.2)$$

Define

$$b_1 \equiv \sum_{\alpha\in I}\sum_{\beta\in B_\alpha} p_\alpha p_\beta \quad \text{and}$$

$$b_2 \equiv \sum_{\alpha\in I}\sum_{\alpha\neq\beta\in B_\alpha} p_{\alpha\beta}, \quad \text{where } p_{\alpha\beta} \equiv \mathbf{E}\left(X_\alpha X_\beta\right).$$

Let Z denote a Poisson random variable with mean λ, so that for $k = 0,1,2,\ldots,$

$$\mathbf{P}(Z=k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Classically, the Poisson distribution is the probability law of rare events. It is remarkable that a few probability distributions arise with great frequency. The three principal distributions are the binomial, the normal, and the Poisson. (See Feller (1968) for an extensive discussion of these matters.)

Let $h: Z^+ \to R$, where $Z^+ = \{0,1,2,\ldots\}$, and $\|h\| \equiv \sup_{k\geq 0}|h(k)|$. We denote the total variation distance between the distributions of W and Z by

$$\|W - Z\| \equiv \sup_{\|h\|=1} \left|\mathbf{E}(h(W)) - \mathbf{E}(h(Z))\right| = 2 \sup_{A\subset Z^+} \left|\mathbf{P}(W \in A) - \mathbf{P}(Z \in A)\right|.$$

More general versions of the following theorem appear in Arratia et al. (1989, 1990). They refer to this approach as the Chen-Stein method.

Theorem 4.2 *Let W be the number of occurrences of dependent events, and let Z be a Poisson random variable with $\mathbf{E}(Z) = \mathbf{E}(W) = \lambda$. Then*

$$\|W - Z\| \leq 2(b_1 + b_2)\frac{1 - e^{-\lambda}}{\lambda} \leq 2(b_1 + b_2),$$

and in particular

$$\left|\mathbf{P}(W = 0) - e^{-\lambda}\right| \leq (b_1 + b_2)(1 - e^{-\lambda})/\lambda.$$

We first apply Poisson approximation to the distribution of the length of long success runs in Bernoulli trials. This has application to molecular sequences and provides a good illustration of the methods needed for sequence comparisons in the case when the alignment is unknown. Let $C_1, C_2, \ldots$ be independent Bernoulli random variables with success probability p, and let R_n be the length of the longest run of heads contained in the first n tosses. Fix a test level t and let the index set be $I = \{1,2,\ldots,n-t+1\}$; the elements of the index set will denote locations where head runs of length t or greater may begin. A head run of length t or more begins at position 1 if and only if the indicator random variable

$$X_1 = C_1 C_2 \ldots C_t$$

takes on the value 1. Now, to unclump the remaining runs define

$$X_\alpha = (1 - C_{\alpha-1})C_\alpha C_{\alpha+1}\ldots C_{\alpha+t-1} \text{ for } \alpha = 2,3,\ldots,n-t+1.$$

For $\alpha = 2,3,\ldots,n-t+1$, X_α will be 1 if and only if a run of t or more heads begins at position α, preceded by a tail. Below we calculate b_1, show that $b_2 = 0$, and find a bound for the approximation.

Write now the total number of clumps of runs of length t or more as the sum of dependent indicator random variables

$$W = \sum_{\alpha\in I} X_\alpha.$$

The Poisson approximation heuristic says we should be able to approximate the distribution of W by a Poisson random variable with mean

$$\lambda_n(t) = \mathbf{E}(W) = p^t\left((n-t)(1-p)+1\right). \tag{4.3}$$

In particular then, since we have as events

$$\{R_n < t\} = \{W = 0\},$$

the distribution function of R_n can be approximated as

$$\mathbf{P}(R_n < t) = \mathbf{P}(W = 0) \cong e^{-\lambda_n(t)}.$$

The test length t is dictated by requiring λ to be bounded away from 0 and ∞; this is equivalent to the condition that $t - \log_{1/p} n$ is bounded. In fact, for integer t, with c defined by

$$t = \log_{1/p}((n-t)(1-p)+1) + c,$$

the above approximation predicts that

$$\mathbf{P}(R_n < t) \cong e^{-\lambda_n(t)} = \exp(-p^c),$$

that is, that $R_n - \log_{1/p}((n-t)(1-p)+1)$ has an asymptotic extreme value distribution. This is almost so; the limiting distribution is complicated by the fact that R_n can assume only integer values. However, this fact does not complicate the approximation itself.

For an example we return to our problem with the 16S rRNA sequence. We model an R run of length 14 by n = 1,541 independent tosses of a fair coin. Using formula (4.3), we calculate that $\lambda_n = 0.0700$. Using the Poisson distribution, $\mathbf{P}(R_{1541} \geq 14)$ is approximately $1 - \exp(-\lambda_n) = 0.0676$.

Even so, without a bound on the error we have no way of knowing if the event is likely or not. To assess the accuracy of the above approximation, we apply Theorem 4.2. We define $B_\alpha = \{\beta \in I: |\alpha - \beta| \leq t\}$ for all α. Since X_α is independent of $\{X_\beta : \beta \notin B_\alpha\}$, condition 1 is satisfied. Furthermore, if $1 \leq |\alpha - \beta| \leq t$, we cannot have both X_α and X_β equal to 1 since we require that a run begin with a tail. Therefore $p_{\alpha\beta} = 0$ for $\beta \in B_\alpha, \beta \neq \alpha$, and hence $b_2 = 0$.

In order to calculate $b_1 = \sum_\alpha \sum_{\beta \in B_\alpha} p_\alpha p_\beta$, we break up the sum over $\beta \in B_\alpha$ into two parts, depending on whether or not p_1 appears. This yields the bound

$$b_1 < \lambda^2(2t+1)/(n-t+1) + 2\lambda p^t. \tag{4.4}$$

Theorem 4.2 now shows us that the Poisson approximation is quite accurate for the example considered above; the probability computed is correct to within $b_1 < 10^{-4}$, that is,

$$0.0699 \leq \mathbf{P}(R_n \geq 13) \leq 0.0701.$$

Alignment Unknown

The situation for matching between two sequences is closely related, although the dependence structure becomes more complex. Suppose that the two sequences $A_1A_2\ldots A_n$ and $B_1B_2\ldots B_m$ are made up of letters independently and uniformly chosen from a d-letter alphabet. It must be emphasized that whenever the letters are not uniformly chosen, Theorem 4.2 holds but is not straightforward to apply. In matching DNA, $d = 4$; for protein sequences, $d = 20$. Let

$$I = \{(i,j): 1 \leq i \leq n-t+1, 1 \leq j \leq m-t+1\}.$$

Define indicator random variables

$$E_{i,j} = 1 \text{ if } A_i = B_j.$$

Let $p = \mathbf{P}(E_\alpha = 1) = 1/d$.

As in the case of head runs, we need to unclump matches and consider "boundary effects." Let

$$X_{i,j} = E_{i,j}E_{i+1,j+1}\cdots E_{i+t-1,j+t-1} \text{ if } i = 1 \text{ or } j = 1$$

and otherwise

$$X_{i,j} = (1 - E_{i-1,j-1})E_{i,j}E_{i+1,j+1}\cdots E_{i+t-1,j+t-1}.$$

With $W = \sum_{\alpha \in I} X_\alpha$, calculating $\lambda = \mathbf{E}(W)$ yields

$$\lambda = p^t[(n+m-2t+1) + (n-t)(m-t)(1-p)]. \tag{4.5}$$

In matching two tRNA sequences, one of length 76, the other of length 77, would a match of length 9 be unusual? For the given parameters, $\lambda = 0.0136$ and under the model above, the event has a probability of approximately

$$1-\exp(-\lambda) = 0.135.$$

A bound on the error may be calculated in a way similar to that for coin tossing. We note that again $b_2 = 0$, and by breaking the sum for b_1 into two sums, one of which is made up of all terms that involve the boundary, we find

$$b_1 < \lambda^2(2t+1)/((n-t+1)(m-t+1)) + 2\lambda p^t.$$

Hence, the probability above is correct to within 8.5×10^{-7}.

APPLICATION TO RNA EVOLUTION

Now we bring these ideas to bear on our RNA evolution problem. We have a set of 33 tRNA molecules and one 16S rRNA molecule from *E. coli*. In Bloch et al. (1983), matchings between 16S and each of the tRNAs were intensely studied. tRNA evolution is a complex topic and tRNA/tRNA comparisons were not made in this study. Table 4.1 shows the length of the longest exact match H_n between these sequences, along with estimates of significance or *p*-values $(1-e^{-\lambda n})$ from our Chen-Stein method. There are no exceptionally good matchings in this list, and so this analysis discounts any deep relationship between the sequences. In fact the *p*-values seem unusually large. In the 33 comparisons the minimum *p*-value is 0.26. Still we should not give up the search. One estimate puts the origin of these sequences at 3 billion years ago. We should not expect large segments of sequence to be preserved in every position over such vast amounts of time. Instead, mutations such as substitutions, insertions, and deletions will accumulate, greatly complicating our task. It is possible that the hy-pothesis of common origin is correct and that so much evolutionary change has taken place that no significant similarity remains. The next section, "Two Behaviors Suffice," examines the results of this search for unusual similarity using more subtle sequence comparison algorithms.

Table 4.1 Exact Match P-Values

tRNA	GenBank Locus	Length (n)	H_n	$1 - e^{-\lambda n}$	b_1
ala-1a	ECOTRA1A	76	9	0.26	1.87×10^{-5}
ala-1b	ECOTRA1B	76	9	0.26	1.87×10^{-5}
cys	ECOTRC	74	8	0.69	2.67×10^{-4}
asp-1	ECOTRD1	77	8	0.71	2.79×10^{-4}
glu-1	ECOTRE1	76	10	0.71	1.25×10^{-6}
glu-2	ECOTRE2	76	10	0.71	1.25×10^{-6}
phe	ECOTRF	76	9	0.26	1.87×10^{-5}
gly-1	ECOTRG1	74	7	0.99	3.90×10^{-3}
gly-2	ECOTRG2	75	6	1.00	5.70×10^{-2}
gly-3	ECOTRG3	76	9	0.26	1.87×10^{-5}
his-1	ECOTRH1	77	9	0.26	1.89×10^{-5}
ile-1	ECOTRI1	77	9	0.26	1.89×10^{-5}
ile-2	ECOTRI2	76	10	0.71	1.25×10^{-6}
lys	ECOTRK	76	6	1.00	5.78×10^{-2}
leu-1	ECOTRL1	87	8	0.76	3.19×10^{-4}
leu-2	ECOTRL2	87	8	0.76	3.19×10^{-4}
leu-5	ECOTRL5	87	9	0.29	2.16×10^{-5}
met-f	ECOTRMF	77	9	0.26	1.89×10^{-5}
met-m	ECOTRMM	77	8	0.71	2.79×10^{-4}
asn	ECOTRN	76	7	0.99	4.01×10^{-3}
gln-1	ECOTRQ1	75	8	0.70	2.71×10^{-4}
gln-2	ECOTRQ2	75	8	0.70	2.71×10^{-4}
arg-1	ECOTRR1	76	7	0.99	4.01×10^{-3}
arg-2	ECOTRR2	77	7	0.99	4.07×10^{-3}
ser-1	ECOTRS1	88	8	0.76	3.23×10^{-4}
ser-3	ECOTRS3	93	9	0.31	2.33×10^{-5}
thr-ggt	ECOTRTACU	76	7	0.99	4.01×10^{-3}
val-1	ECOTRV1	76	8	0.70	2.75×10^{-4}
val-2a	ECOTRV2A	77	8	0.71	2.79×10^{-4}
val-2b	ECOTRV2B	77	9	0.26	1.89×10^{-5}
trp	ECOTRW	76	7	0.99	4.01×10^{-3}
tyr-1	ECOTRY1	85	8	0.75	3.11×10^{-4}
tyr-2	ECOTRY2	85	8	0.75	3.11×10^{-4}

H_n, the length of the longest exact match between the listed tRNA molecule and a 16S rRNA molecule; $1-e^{-\lambda n}$, the p-value (estimate of significance) for n^{th} tRNA molecule; b_1, column entry is the calculated bound on b_1.

TWO BEHAVIORS SUFFICE

In this section we describe a statistic that provides a link between the sections "Global Sequence Comparison" and "Local Sequence Comparison" of this chapter. This statistic is the score of the best matching intervals between two sequences, where nonidentities in the alignments receive penalties. In the section "Global Sequence Comparisons," we showed that the growth of score of global alignments of random sequences is linear with sequence length. In the section "Local Sequence Comparisons," we showed that the number of long runs of exact matches between random sequences has an approximate Poisson distribution. Below we show that the Poisson distribution implies that, for exact matching, the growth of longest run length is proportional to the logarithm of the product of sequence length. Then we state the result that all optimal alignments of a broad class have a score that has either logarithmic or linear growth, depending on the penalties for nonidentities. We will consider two sequences $\mathbf{A} = A_1 A_2 \ldots A_n$ and $\mathbf{B} = B_1 B_2 \ldots B_n$ of equal length n.

Recall that $p := \mathbf{P}(\text{two random letters are identical}) = \mathbf{P}(C_\alpha = 1)$. In the case of unknown alignments, $\lambda = \mathbf{E}(W)$ is given from equation (4.5) by

$$\lambda = p^t((n+n-2t+1)+(n-t)(n-t)(1-p)).$$

For $\lambda \approx 1$, we expect one run of length t. Then

$$\begin{aligned} 1 &= p^t((n+n-2t+1)+(n-t)(n-t)(1-p)) \\ &\approx p^t(nn(1-p)). \end{aligned}$$

Solving for t yields

$$t = \log_{1/p}(nn(1-p)).$$

Therefore the length of the longest run of identities grows like $\log_{1/p}(n^2) = 2\log_{1/p}(n)$.

To relax our stringent requirement of identities, we recall scoring for the alignments as introduced in the section "Global Sequence Comparisons." Extend the sequence $A_1 A_2 \ldots A_n$ to $A_1^* A_2^* \ldots A_L^*$ by inserting gaps "–" and similarly extend $B_1 B_2 \ldots B_n$ to $B_1^* B_2^* \ldots B_L^*$.

Define

$$S(\mathbf{A},\mathbf{B}) = \max \sum_{i=1}^{L} s(A_i^*, B_i^*),$$

where

$$s(A,B) = \begin{cases} +1 \text{ if } A = B \\ -\mu \text{ if } A \neq B \end{cases},$$

$$s(-,B) = s(A,-) = -\delta,$$

and $\mu \geq 0, \delta \geq 0$. The maximum is extended over all ways of inserting gaps and all L.

In Smith and Waterman (1981) and Waterman and Eggert (1987), dynamic programming algorithms are presented to compute

$$H(\mathbf{A},\mathbf{B}) = H(\mathbf{A},\mathbf{B};\mu,\delta) = \max\{S(I,J): I \subset \mathbf{A}, J \subset \mathbf{B}\}$$

in time $O(n^2)$. By $I \subset \mathbf{A}$, for example, we mean all $I = A_i A_{i+1} \ldots A_j$, where $1 \leq i \leq j \leq n$. This algorithm was designed to study situations like our 16S rRNA/tRNA relationships. We are searching for segmental alignments that are not necessarily perfect matchings but are unusually good matches. After some discussion of the statistical properties of $H(\mathbf{A},\mathbf{B};\mu,\delta)$, we will apply the algorithm to our data.

The statistic $H(\mathbf{A},\mathbf{B};\mu,\delta)$ is for one of the so-called local alignment algorithms. However, when the penalties μ and δ are set to 0, the algorithm computes a global alignment. The results in the section "Global Sequence Comparisons" imply that

$$H(\mathbf{A},\mathbf{B};0,0) \sim a \cdot n;$$

but when the parameters are set to ∞, the results in the section "Local Sequence Comparisons" imply that

$$H(\mathbf{A},\mathbf{B};\infty,\infty) \sim b \log n .$$

It is natural to ask if there are other growth rates. The answer is presented in Waterman et al. (1987) and Arratia and Waterman (1994), where the following result is proved: Assume both sequences have equal lengths n. There is a continuous curve in the nonnegative (μ,δ) plane such that when (μ,δ) belongs to F_0, the same component as $(0,0)$, the growth of H is linear with sequence length. When (μ,δ) belongs to F_∞, the same component as (∞,∞), the growth is logarithmic with sequence length. In any curve crossing from F_0 to F_∞ there is a phase transition in growth of the score $H(\mu,\delta)$. This behavior is quite general, and in Arratia and Waterman (1994) it is shown to hold with very general penalties for scoring matches, mismatches, and indels. The behavior of $H(\mathbf{A},\mathbf{B};\mu,\delta)$ when (μ,δ) lies on the line between F_0 and F_∞ remains an open question.

RNA EVOLUTION REVISITED

How do the results in the previous section apply to our comparisons of 16S rRNA with tRNAs? As we have seen, the matchings of Bloch et al. (1983) were the result of applying a local algorithm, and so we will apply the local algorithm H to the data and study the distribution of scores. The first task is to compare the sequences using the statistic $H(\mathbf{A},\mathbf{B};\mu,\delta)$ with $\mu = 0.9$ and $\delta = 2.1$. These values have been used in several database searches, and the growth of scores from aligning random sequences lies in the logarithmic region. The results of the algorithm applied to our data can be found in Table 4.2. No closed-form Chen-Stein method has been arrived at for alignments with indels, so the results are presented in number of standard deviations ($\#\sigma$) above the mean value for comparing two random sequences of similar lengths. (See Waterman and Vingron (1994) for recent work on estimating statistical significance.) The estimated mean as a function of the tRNA length is

$$H(\mathbf{A},\mathbf{B};\mu = 0.9,\delta = 2.1) \;=\; 5.04 \log n - 30.95 ,$$

Table 4.2 Scores and Alignment Statistics

tRNA	Score	# σ	Matches	mms.	Indels
ala-1a	12.2	-.02	14	2	0
ala-1b	12.2	-0.1	14	2	0
cys	21.0	6.2	40	10	5
asp-1	10.8	-1.1	22	8	2
glu-1	10.9	-0.8	21	9	1
glu-2	12.8	0.6	22	8	1
phe	13.0	0.6	32	10	5
gly-1	9.4	-1.4	15	4	1
gly-2	9.5	-1.2	35	15	6
gly-3	14.4	1.5	41	14	7
his-1	13.2	1.1	28	12	2
ile-1	13.6	0.9	41	26	2
ile-2	14.0	1.3	35	10	6
lys	10.7	-0.5	23	7	3
leu-1	13.8	0.7	49	28	5
leu-2	11.7	-0.7	33	17	3
leu-5	13.4	0.4	36	14	5
met-f	12.0	-.03	44	20	7
met-m	11.4	-0.2	21	4	3
asn	15.3	2.4	33	13	3
gln-1	11.8	0.1	23	8	2
gln-2	12.1	0.2	26	11	2
arg-1	13.3	0.7	48	23	7
arg-2	12.8	0.3	26	8	3
ser-1	11.1	-1.3	29	11	4
ser-3	13.8	0.3	42	18	6
thr-ggt	10.1	-1.3	15	1	2
val-1	11.9	-0.2	22	9	1
val-2a	11.3	-0.7	14	3	0
val-2b	11.3	-0.4	14	3	0
trp	11.0	-0.7	22	10	1
tyr-1	11.7	-0.4	31	17	2
tyr-2	10.9	-0.9	42	19	7

σ, the number of standard deviations above the mean value (for comparing the two random sequences of similar lengths); mms., mismatches; indels, insertions/deletions.

while the standard deviation is estimated to be $\hat{\sigma} = 1.49$. In contrast to Table 4.1, there is one tRNA, that for cystine, that scores exceptionally high. The tails of the extremal distributions in the logarithmic region probably behave like $\exp(-\lambda t)$, where t is the test value as in the section "Local Sequence Comparisons" and λ is a constant, but this has not yet been proven rigorously. The usual intuition informed by the tail of a normal distribution has the probabilities behaving like $\exp(-t^2)/2$, which converges to 0 much more rapidly than the Poisson or exponential tails. Thus except for the cystine score, the remaining scores look very much like scores from random sequences. Simulations were performed, and the score 21.0 has an approximate p-value of 10^{-3}, so that it is not possible to dismiss this matching for statistical reasons alone. As far as we know, no one has offered a biological explanation of this interesting match. As to the hypothesis of Bloch et al. (1983), while their work concluded that "matches are too frequent and extensive to be attributed to coincidence," it is not supported by the data but is instead the result of incorrect estimation of p-values. This data set received their most extensive analysis, and they concluded that over 30 percent of the matchings between *E. coli* 16S rRNA and tRNAs were significant at the level $\alpha = 0.10$. Correct estimates show about 10 percent of the matching at the level $\alpha = 0.10$. While the origin of life may be hiding in tRNA and 16S rRNA, it remains elusive.

REFERENCES

Aldous, D.J., 1989, *Probability Approximations via the Poisson Clumping Heuristic*, New York: Springer-Verlag.

Arratia, R.A., L. Goldstein, and L. Gordon, 1989, "Two moments suffice for Poisson approximation: The Chen-Stein method," *Annals of Probability* **17**, 9-25.

Arratia, R.A., L. Goldstein, and L. Gordon, 1990, "Poisson approximation and the Chen-Stein method," *Statistical Science* **5**, 403-434.

Arratia, R.A., and M.S. Waterman, 1994, "A phase transition for the score in matching random sequences allowing deletions," *Annals of Applied Probability* **4**, 200-225.

Bloch, D.P., B. McArthur, R. Widdowson, D. Spector, R.C. Guimaraes, and J. Smith, 1983, "tRNA-rRNA sequence homologies: Evidence for a common evolutionary origin?," *Journal of Molecular Evolution* **19**, 420-428.

Chung, K.L., 1974, *A Course in Probability Theory*, 2nd ed, San Diego, CA: Academic Press.

Chvátal, V, and D. Sankoff, 1975, "Longest common subsequences of two random sequences," *Journal of Applied Probability* **12**, 306-315.

Deken, J., 1979, "Some limit results for longest common subsequences," *Discrete Mathematics* **26**, 17-31.

Erdös, P., and A. Rényi, 1970, "On a new law of large numbers," *Journal d' Analyse Mathematique* **22**, 103-111. Reprinted in 1976 in *Selected Papers of Alfred Rényi*, Vol. 3, 1962-1970, Budapest: Akadémiai Kiadó.

Feller, W., 1968, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., New York: Wiley and Sons.

Goldstein, L., 1990, "Poisson approximation and DNA sequence matching," *Communications in Statistics. Part A: Theory and Methods* **19**, 4167-4179.

Kingman, J.F.C., 1973, "Subadditive ergodic theory," *Annals of Probability* **1**, 883-909.

Smith, T.F., and M.S. Waterman, 1981, "Identification of common molecular subsequences," *Journal of Molecular Biology* **147**, 195-197.

Steele, J.M., 1986, "An Efron-Stein inequality for nonsymmetric statistics," *Annals of Statistics* **14**, 753-758.

Waterman, M.S., and M. Eggert, 1987, "A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons," *Journal of Molecular Biology* **197**, 723-728.

Waterman, M.S., L. Gordon, and R. Arratia, 1987, "Phase transitions in sequence matches and nucleic acid structure," *Proceedings of the National Academy of Sciences USA* **84**, 1239-1243.

Waterman, M.S., and M. Vingron, 1994, "Rapid and accurate estimates of statistical significance for data base searches," *Proceedings of the National Academy of Sciences USA* **91**, 4625-4628.