Comment

Michael S. Waterman

The authors of this article mention that their interest in Poisson approximation was motivated by questions in sequence matching. Sequence matching refers to the comparison of two or more sequences to locate regions that are exceptionally similar. While these questions are of interest in computer science, my own motivation to study sequence matching has been molecular biology. Section 5 of the paper is devoted to a biological example. I will take this opportunity to expand upon some statistical questions of interest to molecular biology.

Biology is embarked on one of the most exciting scientific endeavors of the century. The widely publicized Human Genome Initiative (*Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years, FY 1991–1995*, April 1990; DOE/ER-045 2P) has as its goal the analysis of the structure of human DNA and the determination of the location of the estimated 100,000 genes. Other model organisms are included in the Initiative to provide the comparative information necessary for understanding the human and other genomes. Medical doctors and legislators may choose to focus on the understanding and possible consequent treatment of more than 4000 human genetic diseases. Some may well view the project as providing the initial data for a fundamental understanding of the processes of life. In any case, the rate at which information is being gathered is astonishing. International DNA databases began to be formed in 1982. The databases are DDBJ (Japan), EMBL (Europe) and GenBank (US). By 1986, Release #42.0 of GenBank had 6.7×10^6 nucleotides (bases) of sequence data. Release #62.0 cited by the authors had 37.2×10^6 nucleotides, while the most recent release #65.0 in 1990 has 49.2×10^6 nucleotides. New technology promises to accelerate the rate of sequence determination. Molecular biology has been an experimental and empirical science. The flow of sequence information is changing the character of the subject.

Our interest in Poisson approximation began with an early analysis of the DNA database. DNA sequences average 1000 nucleotides in length and have a four letter alphabet adenine (A), guanine (G), cytosine (C) and thymine (T). In 1981, Temple Smith and I devised a method or algorithm for finding similar regions of sequences. Briefly, this method optimized a score for all segments I of sequence $\mathbf{x} = x_1 x_2 \cdots x_n$ and all segments J of sequence $\mathbf{y} = y_1 y_2 \cdots y_n$. The score, in its simplest form, counts $+1$ for a match or identical letter from I and J, counts $-\mu$ for a mismatch or nonidentity and counts $-\delta$ for a letter inserted or deleted from a sequence (an indel). For example AAGTC and AGCC can be arranged or aligned as

```
AAGTC
A-GCC
```

to receive score $S = 3 - \mu - \delta$. They can also be aligned as

```
AAGTC
AGCC-
```

to receive score $S = 1 - 3\mu - \delta$. The algorithm, based on dynamic programming, provides a straightforward

Michael S. Waterman is Professor of Mathematics and of Molecular Biology, University of Southern California, Los Angeles, California 90089-1113.

way to compute

$$M_{n,m} = M(\mathbf{x}, \mathbf{y})$$
$$= \max_{\substack{I \subset \mathbf{x} \\ J \subset \mathbf{y}}} \{\#\text{matches} - \mu\,\#\text{mismatches} - \delta\,\#\text{indels}\}.$$

$M(\mathbf{x}, \mathbf{y})$ can be found in time $O(nm)$ (Smith and Waterman, 1981; Waterman and Eggert, 1987).

In Smith, Burks and Waterman (1985), all pairs of vertebrate sequences (and their reverse complements) were compared using the above algorithm with $\mu = .9$ and $\delta = 2.1$. The goal was to understand the distribution of M when there was no biologically significant relationship between the sequences. Regression techniques that trimmed the outliers were used and an excellent fit to the data was obtained:

$$\hat{M}_{n,m} = 2.55 \frac{\log(nm)}{\log(1/p)} - 8.99,$$

where $p = \sum_{a \in \mathscr{A}} P(X = Y = a)$, and X and Y are idd with the frequencies of nucleotides in the sequences.

The estimated standard deviation was $\hat{\sigma} = 1.78$. A simulation of same set of sequences with independent letters was performed and almost the identical fit was obtained. This was not obvious, since it is known that biological sequences have dependencies (Tavaré and Giddings, 1989). Today these calculations occupy a SUN workstation for 2 or 3 hours; it should be pointed out that in 1982 the same calculations took a similar amount of time on a CRAY, then the world's fastest computing machine.

Not long after the data analysis in 1982, we learned of the Erdös–Renyí law (1970). The longest run of heads with no tails is described beautifully in Arratia, Goldstein and Gordon. It corresponds exactly to the case when we know the alignment of the independent sequences:

$$\begin{matrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{matrix}.$$

Whenever $x_i = y_i$ replace the aligned pair by "H", otherwise $x_i \neq y_i$ and we replace the aligned pair by "T." The random variable R_n corresponds to M where only one alignment is considered and where $\mu = \delta = \infty$. Several generalizations are suggested by the biology. The sequences need to be shifted against one another to produce different alignments, and alignments with "errors"—mismatches, insertions or deletions—need to be included.

It is straightforward to handle the complications that arise from shifts of the sequences when the sequence letters are independently and uniformly distributed. For a test length t, the index set I is the possible starting locations for matchings of length t:

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

Define $C_{ij} = 1\{X_i - Y_j\}$ and $p = P(C_{ij} = 1)$. To declump, define

$$Z_{i,j} = C_{i,j} C_{i+1,j+1} \cdots C_{i+t-1,j+t-1} \quad \text{if } i = 1 \text{ or } j = 1,$$

and otherwise

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

With $W = \sum_{\alpha \in I} Z_\alpha$, calculating $\lambda = E[W]$ yields

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

This is used to assign statistical significance by the formula

$$P(M_{n,m} < t) = P(W = 0) \cong e^{-\lambda}.$$

It is easy to show that $b_2 = b_3 = 0$ for the natural choice of neighborhoods B_α, and

$$|P(W = 0) - e^{-\lambda}| \le b_1(1 - e^{-\lambda})/\lambda.$$

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

It is possible to introduce inexact matching into these formulas. The next step was the longest match with k mismatches, but with the Chen–Stein method it is possible to go beyond that problem. Equation (31) of Arratia, Goldstein and Gordon was obtained by using the ballot theorem with the Chen–Stein method. Still these results only hold for $\delta = \infty$ and $\mu > p/(1 - p)$. In Waterman, Gordon and Arratia (1987), a strong law is described where

$$M_{n,n} \sim a \cdot n \quad \text{if } (\mu, \delta) \in S_1,$$

while

$$M_{n,n} \sim b \log(n) \quad \text{if } (\mu, \delta) \in S_2.$$

S_1 and S_2 are separated by a single, continuous curve. Essentially, if S_1 is a set of "large" (μ, δ) while S_2 is a set of "small" (μ, δ). The contrast between behavior on these (adjacent) regions is sharp and constitutes a phase transition. The case $\mu = \delta = 0$ is famous and is known as the longest common subsequence problem (Chvátal and Sankoff, 1975). Kingman's subadditive ergodic theorem can be used to show that there exists $0 < \alpha < 1$ where

$$P\left(\lim \frac{M_{n,n}}{n} = \alpha\right) = 1$$

but α remains unknown (Deken, 1979). It is interesting that the "symmetric" case, $\mu = \delta = \infty$, indexes a manageable problem.

Generalization to more general scoring schemes was begun by Arratia, Morris and Waterman (1988). In that paper, matching letters are given nonnegative scores. The maximum scoring matching contiguous subsequence between sequences of length n and m has behavior $M_{n,m}/k \log(nm) \to 1$ a.s. The constant k is

the largest solution of a certain equation. In addition, the proportion of letter i in the maximum scoring subsequence was given explicitly. These results were extended by Karlin and Altschul (1990), who consider more general scoring schemes. They require the expected score of two letters to be negative and do not allow insertions or deletions. They describe generalizations of Arratia, Morris and Waterman (1988) and give a Chen–Stein style formula to assess statistical significance:

$$P\left(M_{n,m} > \frac{\ln(nm)}{\lambda^*} + x\right) \leq Ce^{-\lambda^* x}.$$

It is straightforward to simulate random sequences and to obtain estimates of the 95th percentile of the score distribution. Why then is the Chen–Stein analysis valuable to molecular biology? An important part of the answer lies in the number of sequence comparisons that are made. When a newly determined sequence is compared to the full GenBank database, the sequence is compared to about 49,000 sequences of average length 1,000 nucleotides. These sequences are of differing compositions and could each require a simulation. Since comparison of sequences is the rate limiting step in database searches, we need to have a rapid, accurate way to assess statistical significance. The Chen–Stein method provides that. In addition, very small p-values are almost impossible to determine by simulation, and the Poisson approximation is of much use there.