Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 4625–4628, May 1994
Biochemistry

Rapid and accurate estimates of statistical significance for sequence data base searches

MICHAEL S. WATERMAN† AND MARTIN VINGRON‡

Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, CA 90089-1113

Communicated by Gian-Carlo Rota, January 24, 1994

ABSTRACT **A central question in sequence comparison is the statistical significance of an observed similarity. For local alignment containing gaps to optimize sequence similarity this problem has so far not been solved mathematically. Using as a basis the Chen–Stein theory of Poisson approximation, we present a practical method to approximate the probability that a local alignment score is a result of chance alone. For a set of similarity scores and gap penalties only one simulation of random alignments needs to be calculated to derive the key information allowing us to estimate the significance of any alignment calculated under this setting. We present applications to data base searching and the analysis of pairwise and self-comparisons of proteins.**

Every new DNA or protein sequence is compared with one or more sequence data bases to find similar or homologous sequences that have already been studied. There are numerous examples of important discoveries resulting from these data base searches. One of the most famous is the similarity between platelet-derived growth factor and the v-*sis* oncogene product (1). Another example is the similarity between bovine cAMP-dependent protein kinase and the Rous avian sarcoma virus Src proteins (2), which supports the origination of the *src* genes in host genomes. When the cystic fibrosis gene was cloned and sequenced, a data base search revealed that the gene product had similarity to a family of related ATP-binding proteins involved in active transport of small hydrophilic molecules across the cytoplasmic membrane (3). A great deal of biology is learned from these searches, which are routinely used to create or test hypotheses about the function of a protein or DNA sequence or the membership of a sequence in a family.

There are two mathematical aspects to data base searches: the algorithm used to find sequence similarities and the method used to determine which similarities are interesting. Clearly, the algorithm for searching is important. Both the objective function for scoring alignments and the speed of the algorithm are relevant, since scoring determines the alignments themselves and speed determines how easy and practical searching is. Less well appreciated is the importance of the criteria used to determine whether similarities are of interest. When searching data bases containing tens of thousands of individual sequences, automatic criteria are required. Since detailed sequence-by-sequence biological reasoning cannot be included, statistical significance is usually defined by comparison with the similarity of random sequences.

Two algorithms have become famous for rapid searches of data bases, FASTA and BLAST. They are local rather than global algorithms that look for intervals or segments of good matching between sequences. The FASTA family of programs (4, 5) achieves its speed by searching for diagonals in the comparison matrix where there is dense matching of k-tuples. In the last stage of the analysis, restricted Smith–Waterman dynamic programming is performed, using Dayhoff scoring to weight amino acid substitutions. BLAST (6), which is faster, precomputes all segments or patterns that could score above some test value against a segment of the sequence. Then all instances of this collection of patterns in the data base are found and statistical significance is estimated by Poisson approximation. Both FASTA and BLAST look for best local matchings with scores, although BLAST does not include insertions/deletions (indels). The most rigorous method to search a data base uses the Smith–Waterman local alignment algorithm (7–9). Its associated statistical questions are so complex that until now only a few efforts to calculate significance of alignments with gaps have been undertaken (10, 11).

The traditional method for assigning statistical significance is by simulation. Random sequences with the same distribution of letters as the real sequences are aligned, thus creating a random sample of scores. Then the mean and standard deviation of this sample is computed and the real sequence alignment score is reported in "number of standard deviations above the mean." There are problems with this approach. It is simply too expensive in time to perform such a simulation for all sequences in a data base. Moreover "number of standard deviations" carries with it an implicit and incorrect assignment of significance by the normal distribution. Theory is needed because simulations rarely cover the extreme tails of a distribution. In addition, taking such an approach requires redoing the simulation not only when one uses a new scoring system but for different-length sequences. Both shortcomings are remedied by the conceptually simple and practical method we propose in this paper to estimate the significance of a wide class of local alignments.

First we describe the Smith–Waterman algorithm (9) along with an important extension. The function $s_{x,y}$ is the weight given to a substitution of letter y for letter x, often from the Dayhoff matrix. The function $-g_k$ is the weight given to a gap or indel of length k. $g_k = \alpha + \delta k$ provides an efficient algorithm (12). For simplicity we present $g_k = \delta k$. The sequences $\mathbf{x} = x_1 x_2 \ldots x_n$ and $\mathbf{y} = y_1 y_2 \ldots y_m$ are to be compared. Let $H_{i,j}$ be the best score of any alignment ending at x_i and y_j or 0, whichever is larger. Define $H_{i,j} = 0$ if i or j is 0. Then the recursion is

$$H_{i,j} = \max\{H_{i-1,j-1} + s_{x_i,y_j},\ H_{i-1,j} - \delta,\ H_{i,j-1} - \delta,\ 0\}. \quad [1]$$

The best local similarity score between $\mathbf{x}$ and $\mathbf{y}$ is then $H(\mathbf{x},\mathbf{y}) = \max\{H_{i,j}\colon 1 \le i \le n,\ 1 \le j \le m\}$, and an alignment is determined by traceback from that location, (i, j) say. The computing time is proportional to $n \cdot m$. Frequently the analyst is interested in finding all alignments that score well. This creates an immediate problem, as there are many alignments

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†To whom reprint requests should be addressed.
‡Present address: Gesellschaft für Mathematik und Datenverarbeitung Institute I1, Schoss Birlinghoven, D-53757 St. Augustin, Germany.

that score well but differ in only minor ways from an optimal alignment ending at (i, j). To cope with this difficulty first define an *alignment clump* as the group of alignments sharing at least one pair of aligned letters with a given alignment. The procedure is to select one alignment from a clump and then to *declump* or remove the effects of all alignments in the clump. The method of declumping was introduced to find nonintersecting suboptimal local alignments (13). Let $H_{(1)}$ be the score of the best alignment score. In a traceback from that score, an optimal alignment is produced. Now $H_{(2)}$ is the maximum score where no pair of alignment letters from the first alignment is used. The second score can be produced very quickly from the first by declumping and recomputing in the vicinity of the first alignment. In this way $H_{(1)}, H_{(2)}, \ldots, H_{(N)}$ can be rapidly found. When there is more than one local alignment of interest, this is an important tool as, e.g., the analysis of repeats in a sequence. In addition, the ability to quickly produce the scores of clumps is key to our statistical analysis below.

Next, we turn to the statistical distribution of local alignment scores $H(\mathbf{x}, \mathbf{y})$. The setup is simple. The sequences $\mathbf{x}$ and $\mathbf{y}$ are assumed to be random with letters statistically independent or given by a Markov chain, and the scoring weights $s_{x,y}$ and g_k are fixed. The local alignment score of biologically unrelated sequences then has a statistical distribution that depends on the sequence lengths, letter distribution, and scoring weights. There are two extreme cases. First, let $s_{x,y} = +1$ if $x = y$ and 0 if $x \neq y$ and $g_k = 0$, so that optimal alignment scores are the length of the longest common subsequence common to $\mathbf{x}$ and $\mathbf{y}$. In this case, $H(\mathbf{x}, \mathbf{y}) = H_n$ is proportional to sequence length $n = m$. With probability 1, the limit of H_n/n converges to c, and while the existence of c was established in 1975 (14), its precise value is still unknown. The idea is that there is no penalty for errors, so waiting for another identity cannot decrease the score. In contrast, if $s_{x,y} = 1$ if $x = y$ and $-\infty$ if $x \neq y$ and $g_k = \infty$, then $H(\mathbf{x}, \mathbf{y}) = H_n$ is the length of the longest perfect match between $\mathbf{x}$ and $\mathbf{y}$. The growth of H_n is now proportional to $\log(n)$, and limit $H_n/\log(n) \to 2$ with probability 1. These are the only two growths possible (15). That is for a general scoring $(s_{x,y})$ and $g_k = \alpha + \delta k$, there are two exhaustive sets, S_{lin} and $S_{\log}$, where $((s_{x,y}), \alpha, \delta) \in S_{\text{lin}}$ implies $H_n \sim cn$ and $((s_{x,y}), \alpha, \delta) \in S_{\log}$ implies $H_n \sim d\log(n)$.

Let us look more carefully at the linear/logarithmic growths. The linear region of the parameter space is defined as those parameters where the average score per letter of a global alignment is positive. Hence, positive score is the average event. The logarithmic region is where the average score per letter of a global alignment is negative and the local alignment algorithm "shrinks" the alignment to a truly local one (16). Here, positive-scoring local alignments are rare events in the exponential number of possible local alignments. By using a beautiful development of the Chen–Stein method (17), Poisson approximations can be established in some cases. The subtlety comes in the dependence between clumps of intersecting alignments and alignments that share sequence positions but not matched pairs of positions. The first type of dependence can be removed by declumping, and the second can be controlled by the Chen–Stein method.

We introduce the most relevant known result from sequence matching (18, 19). Let $g_k = \infty$ (no gaps) and $s_{x,y}$ be the scoring parameters. Given $p \in (0, 1)$, the largest root of $f(\lambda) = 1 - \mathbb{E}(\lambda^{-s_{x,y}}) = 0$, where x and y are random letters, then $\lim H_n/\log_{1/p}(n^2) \to 1$. For sequences of length n and m, the "center" of the distribution of optimal local alignment scores is $\log_{1/p}(nm)$. Notice that p is determined by the letter distribution and $s_{x,y}$. For random sequences $\mathbf{X}$, $\mathbf{Y}$ there is a constant γ that can be numerically determined by solving an equation, such that

$$\mathbb{P}\{H(\mathbf{X}, \mathbf{Y}) > t = \log_{1/p}(mn) + c\} \approx 1 - e^{-\gamma mnp^t}. \quad [2]$$

This method of assigning statistical significance is used in BLAST (6). The Poisson distribution is needed to approximate the number of clumps exceeding the center by c—i.e., those for which $H > \log_{1/p}(nm) + c$. This Poisson distribution has mean γmnp^t. This implies there are no scores that large with probability $e^{-\gamma mnp^t}$. The first result obtaining the center $\log_{1/p}(nm)$ was given in ref. 18; it was later extended in ref. 19, where a Poisson approximation was presented. In other work, Chen–Stein approximations have been established for alignments with rich fractions of mismatches (20), without scoring. We now show how to rapidly obtain practical and accurate Poisson approximations for the entire logarithmic region—i.e., for alignments calculated under strong gap penalties.

In Aldous (22), the *Poisson clumping heuristic* is described. The model locates clumps by a Poisson process and then independently assigns clump scores. The ideas from the previous paragraph can be interpreted in this way. The number of clumps scores exceeding a test value t = center + c has a Poisson distribution with mean λ. In particular, the probability that at least one score exceeds t is $1 - \mathbb{P}$(no score exceeds t) $= 1 - e^{-\lambda}$. We provide numerical evidence that the Poisson clumping model holds in the entire logarithmic region. The end of an alignment (i, j) marks the clump location and its score H is the clump score. Throughout the logarithmic region the center is $\log_{1/p}(nm)$. Furthermore, we demonstrate numerically that in the logarithmic region the mean λ of the number of clumps above a threshold t has the form $\lambda = \gamma mnp^t$ as in Eq. **2**.

Given this model for the statistical behavior of alignments with gaps, it is necessary to estimate the parameters γ and p. Except for the no-gap case there is no analytic description of the parameters. However, the theory suggests two different ways of estimating γ and p by using simulations. Both assume a fixed scoring scheme ($s_{x,y}$ and a gap penalty function) and letter distribution.

The obvious method is to apply the algorithm in Eq. **1** many times to statistically independent sequences and to calculate the empirical distribution function of optimal alignment scores—i.e., the fraction of alignments with score less than t. The Poisson clumping heuristic suggests that the probability for an alignment to score below t is given by $e^{-\gamma mnp^t}$. After appropriate transformation (log[−log(data)]) the empirical distribution function is expected to form a straight line. In fact, linear regression gives a correlation coefficient above 0.99. It is then straightforward to calculate the parameters γ and p, and we call this method *direct estimation*.

Yet the true power of the theory sketched above comes to bear in the second method, which we call *declumping estimation*. From a few comparisons we calculate $H_{(1)}, H_{(2)}, \ldots, H_{(N)}$, using the declumping algorithm described above. The crucial observation is that the mean λ of the Poisson can be estimated from this data set by the average number of $H_{(i)}$ that exceed a threshold t, fitting these data by the function γmnp^t. Simulations show that plotting the empirical data on a logarithmic scale leads to an almost perfect straight line. Estimation of γ and p by regression is then straightforward. As we demonstrate, both approaches provide almost equally good estimates of statistical significance, thus by their agreement supporting the assumptions on which they are based. Declumping estimation provides these parameters dramatically faster than direct estimation.

There are several checks to perform. The first is to verify the agreement between our two proposed methods and the theory already established in the case of no indels (19). In fact, all three distribution functions agree extremely well. To test the quality of the approximation for alignments with

gaps, we simulate the empirical distribution function. Many distribution functions can be approximated satisfactorily by using a double exponential. Therefore, while direct estimation is not really challenged by approximating the empirical distribution function, for the declumping estimation to approximate the empirical distribution function well would be highly remarkable. In fact, as shown in Fig. 1, the quality of approximation is nearly perfect, with the tails of the empirical and the approximated distribution being almost indistinguishable.

Taking the tests a step further, we want to know whether the dependence of the score on sequence length is modeled correctly. To this end we applied the approximation to sequences of length different from that used to estimate the parameters. Fig. 1 shows that both direct and declumping estimates perform well. We generally found the direct estimation to result in conservative estimates of significance when derived from long sequences. Not only does this demonstrate that the quality of approximation is sufficient to be used on a routine basis in sequence comparison, but the agreement between the two methods confirms that the Poisson clumping model adequately describes the data.

Our assumptions are also supported by the failure of the Poisson model to work for scoring systems in the linear region where alignments are essentially global instead of local. While the direct estimation appears to result in good fit even in the linear region, it immediately fails when we apply the derived parameters to accommodate length changes. In addition, counting clumps above a threshold as in the declumping simulation does not lead to an exponential function.

The output of a data base search frequently seems distorted by the fact that matches to very long sequences score high and are near the top of the hit list. Based on one simulation done under the parameter settings of the data base search, one can now account for sequence lengths and sort the hit list on the basis of significance rather than score. If the data base is treated as one long sequence as in BLAST, then ranking by score is appropriate, and correct but conservative values of statistical significance are assigned. Instead we rank each individual comparison by estimated significance, usually obtaining a different ranking of comparisons. Since we are taking the most statistically significant comparison out of the large number of comparisons, the significance values should not be considered as correct for the overall data base comparison. Instead we believe the hit list sorted by significance rather than by score is more useful to understanding biology.

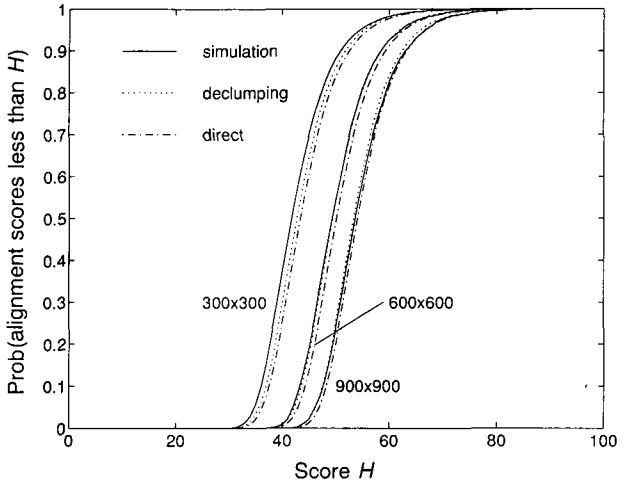

FIG. 1. Approximating the empirical distribution function (solid line) by using parameters derived by direct and declumping estimation. Parameters are derived from comparing length 900 random sequences and then applied to pairs of lengths 300, 600, and 900. Alignment used the PAM250 matrix (23) and gap(k) = 12 + 3k.

We implemented this procedure and achieved a substantial improvement in data base searches. For example, when the Protein Identification Resource PIR1 data base is searched with the α chain of human hemoglobin, the farthest known relative one finds is lupin leghemoglobin. In terms of score we found 25 nonglobins scoring better than it. After sorting by significance instead of score, leghemoglobin still was the last globin on the list, but it moved up and only 10 nonglobins ranked better.§ In a similar experiment PIR1 was searched with an azurin sequence, and we looked how far below the last azurin that plastocyanins are found. When sorted by score, plastocyanins occur between positions 37 and 262; when sorted by significance they are between positions 18 and 107.¶

In the logarithmic region we can give statistical significance estimates for the order statistics $H_{(1)} \geq H_{(2)} \geq H_{(3)} \geq \ldots$ associated with the clump scores (25). The kth best (declumped) alignment score $H_{(k)}$ has the approximate distribution of the kth order statistic of a sample of Poisson mean γmnp^t random variables. That is

$$\mathbf{P}(H_{(k)} - \log_{1/p}(nm) \leq c) \approx e^{-\gamma mnp^t} \sum_{i=0}^{k-1} \frac{(\gamma mnp^t)^i}{i!}. \quad [3]$$

Notice that $\mathbf{P}(H_{(1)} - \log_{1/p}(nm) \leq c) \approx e^{\gamma mnp^t}$, consistent with Eq. 2.

We tested this approximation by calculating the first-, second-, and third-best alignments for pairs of randomly generated sequences. Approximating the resulting three empirical distribution functions using γ and p derived from the optimal alignments and the above formula resulted in excellent fit for the suboptimal curves as well (Fig. 2). Again, this strongly reinforces our assumptions and also has direct applications to sequence comparison. We applied it to analyze the sequence of hemopexin, which has been claimed to contain internal repeats. Apart from the self-match, the highest scores are 142 (significance level better than 0.0001), 62 (significance level 0.04), 57 (0.0366), and 56 (0.012). This pattern of increasing significance of the suboptimal solutions clearly supports the claim.‖ Another consequence of our ability to assess the significance of suboptimal alignments is that we become less vulnerable to a wrong choice of gap penalties. Suppose two sequences are aligned under strong gap penalties and only part of the correct alignment is found. The rest is likely to come up as a high-scoring suboptimal solution which can readily be identified by using its significance. As an example we used a comparison of two immunoglobulins, where the optimal alignment contains only about half of the desired alignment and has a p value of 0.04. The second-best solution, with a p value of 0.001, fills in the remaining part of the alignment.**

§The search was performed with the PAM250 matrix (23), a gap penalty function gap(k) = 12 + 3k, and the distribution of amino acids given by McCalden and Argos (24).

¶Sorting by score also leads to many sequences sharing the same score. For example, the last plastocyanin scores 55 and there are 66 other hits of that same score; 262 is only the number of nonazurins scoring 56 or better. Not all plastocyanins from PIR1 are included in this list, as some of them are found only far under any reasonable cutoff for a search. The search was performed with the PAM250 (23) matrix, a gap penalty function gap(k) = 12 + 2k, and the distribution of amino acids given by McCalden and Argos (24).

‖The alignment was done with the PAM250 matrix (23) and a gap penalty function gap(k) = 12 + 3k. The parameters were derived by using direct simulation assuming uniform distribution of letters.

**Based on PAM250 matrix (23) and gap(k) = 12 + 3k. Compare to the analysis in ref. 16, where this comparison is studied in detail.

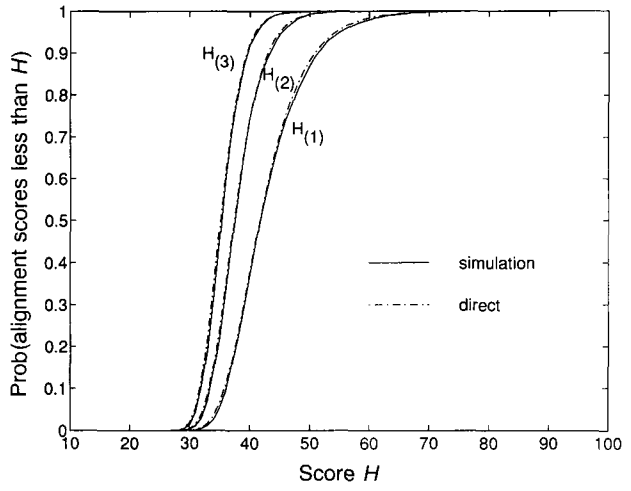

FIG. 2. Using parameters derived by direct estimation from 1000 pairs of length 300 sequences, we approximated the empirical distribution function of first-, second-, and third-best alignment. The empirical distribution function is based on 10,000 random comparisons. The PAM250 matrix and gap(k) = 12 + 3k were used.

Checking the distribution in Eq. 2 with the results of a data base search is easily accomplished. Compare the query sequence of length m to statistically independent sequences of length n_i, i = 1 to N, from a data base. The score of the ith comparison is H_i. Then $Y_i = e^{-\gamma mn_i p^{H_i}}$, i = 1 to N, is (approximately) a sample of uniform (0, 1) random variables. Order them $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(N)}$ and the average value of $Y_{(i)} = e^{-\gamma mn_i p^{H_i^*}}$ is (approximately) $i/(N + 1)$. Taking logarithms, the fit from using parameters γ and p can be evaluated by using the equations $\log(-\log i/(N + 1)) - \log(mn_i) \approx \log\gamma + H_i^* \log(p)$. We have checked the fit by running the α chain of human hemoglobin against nonhemoglobins in Swis-Prot and NEWAT [a data base from R. Doolittle that has duplicate sequences removed (21)]. In each case the fit is good with conservative estimates of significance. We also used higher-order dependence in the simulation, with identical results.

We have demonstrated how to estimate the statistical significance of optimal and suboptimal local alignments and applied this to data base searches, the identification of internal repeats, and the improvement of alignment quality. General-purpose parameters for use in data base searches can be derived by using the compositions of the query sequence and the data base. For more subtle analysis such as the assessment of suboptimal alignments it is preferable to derive parameters from simulations mimicking closely the situation under study. Software is available by anonymous ftp from hto.usc.edu.

This work was supported by grants from the National Science Foundation (DMS 90-05833 and DMS 87-20208) and the National Institutes of Health (GM-36230).

1. Doolittle, R. F., Hunkapiller, M. W., Hood, L. E., Devare, S. G., Robbins, K. C., Aaronson, S. A. & Antoniades, H. N. (1983) *Science* **221,** 275–276.
2. Barker, W. C. & Dayhoff, M. O. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2836–2839.
3. Riordan, J. R., Rommens, J. M., Kerem, B. s., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.-L., Drumm, M. L., Iannuzzi, M. C., Collins, F. S. & Tsui, L.-C. (1989) *Science* **245,** 1066–1072.
4. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 726–730.
5. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441.
6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
7. Pearson, W. R. (1991) *Genomics* **11,** 635–650.
8. Coulson, A. F. W., Collins, J. F. & Lyall, A. (1987) *Comput. J.* **30,** 420–424.
9. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
10. Collins, J. F. & Coulson, A. F. W. (1990) *Methods Enzymol.* **183,** 474–487.
11. Mott, R. F. (1992) *Bull. Math. Biol.* **54,** 59–76.
12. Gotoh, O. (1982) *J. Mol. Biol.* **162,** 705–708.
13. Waterman, M. S. & Eggert, M. (1987) *J. Mol. Biol.* **197,** 723–728.
14. Chvátal, V. & Sankoff, D. (1975) *J. Appl. Probab.* **12,** 306–315.
15. Arratia, R. & Waterman, M. S. (1994) *Ann. Appl. Probab.* **4,** 200–225.
16. Vingron, M. & Waterman, M. S. (1994) *J. Mol. Biol.* **235,** 1–12.
17. Arratia, R., Goldstein, L. & Gordon, L. (1989) *Ann. Probab.* **17,** 9–25.
18. Arratia, R., Morris, P. & Waterman, M. S. (1988) *J. Appl. Probab.* **25,** 106–119.
19. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
20. Arratia, R., Gordon, L. & Waterman, M. S. (1990) *Ann. Stat.* **18,** 539–570.
21. Doolittle, R. F. (1981) *Science* **214,** 149–159.
22. Aldous, D. (1989) *Probability Approximations via the Poisson Clumping Heuristic* (Springer, New York).
23. Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983) *Methods Enzymol.* **91,** 524–545.
24. McCaldon, P. & Argos, P. (1988) *Proteins Struct. Funct. Genet.* **4,** 99–122.
25. Goldstein, L. & Waterman, M. S. (1992) *Bull. Math. Biol.* **54,** 785–812.