Michael S. Waterman†‡ and Louis Gordon†
†Department of Mathematics, University of Southern California and ‡Department of Molecular Biology, University of Southern California

Multiple Hypothesis Testing for Sequence Comparisons

INTRODUCTION

In 1970 dynamic programming was first applied to the comparison of biological sequences by Needleman and Wunsch.[8] Their method is now called a similarity method. Since their work, many extensions and modifications have been introduced. This includes distance methods, general gap functions, multiple alignment procedures, and near-optimal methods. See Waterman[14] for a review of these approaches to sequence comparison.

When more sequences began to appear in the later 1970's, it became apparent that alignment of entire sequences was frequently not the major problem of interest. Instead it was more valuable to look for the good matching segments within longer sequences. Distance methods had become very popular, perhaps due to their mathematical relationship to metric spaces, and converting distance methods to handle segmental matching was a difficult task. See Sellers[9] and Goad and Kanehisa.[6] Similarity methods however could more easily be modified for segment comparisons.[10,12] Next we present the algorithm from the papers. Take the two sequences to be $\underline{a} = a_1a_2\ldots a_n$ and $\underline{b} = b_1b_2\ldots b_m$. They can be either DNA or protein sequences. The similarity measure between sequence letters a and b is $s(a,b)$, where $s(a,b) > 0$ if $a = b$ and $s(a,b) < 0$ for at least some cases of $a \neq b$.

Computers and DNA, SFI Studies in the Sciences of Complexity, vol. VII
Eds. G. Bell and T. Marr, Addison-Wesley, 1990

Insertions or deletions of length k receive weight $-w(k)$. The observation of Smith and Waterman[10] is that negative scoring alignments are of no interest. S($\underline{a}$, $\underline{b}$) is defined to be the best (largest) score from aligning $\underline{a}$ and $\underline{b}$. Define

$$G_{i,j} = \max\left\{0; S(a_x a_{x+1} \dots a_i, b_y b_{y+1} \dots b_j) : 1 \le x \le i, 1 \le y \le j\right\}. \tag{1}$$

$G_{i,j}$ is the best score of any alignment ending at a_i and b_j or 0, whichever is larger. The similarity algorithm is started with $G_{i,0} = G_{0,j} = 0$ for $1 \le i \le n$ and $1 \le j \le m$. Then

$$G_{i,j} = \max\left\{0, G_{i-1,j-1} + s(a_i, b_j), E_{i,j}, F_{i,j}\right\}, \tag{2}$$

where

$$E_{i,j} = \max_{1 \le k \le j}\left\{G_{i,j-k} - w(k)\right\}, \tag{3}$$

and

$$F_{i,j} = \max_{1 \le k \le i}\left\{G_{i-k,j} - w(k)\right\}. \tag{4}$$

The best scoring alignment has score max G_{ij}. Gotoh[7] showed the time for the multiple gap algorithm of Waterman et al.[11] could be reduced to $O(n^2)$ for linear gap functions $w(k) = u + vk$. For the above algorithm this is accomplished by altering the last two recursions, Eqs. (3) and (4), to:

$$E_{i,j} = \max\left\{G_{i,j-1} - (u+v), E_{i,j-1} - v\right\},$$
$$F_{i,j} = \max\left\{G_{i-1,j} - (u+v), F_{i-1,j} - v\right\}.$$

The average sequence in GenBank or EMBL is 1000 bases long. Figure 1 shows best segment alignments for independent simulations of 10 independent pairs of length 1000 DNA sequences with $P(A) = P(C) = P(G) = P(T) = .25$. The algorithm parameters are $s(a,a) = 1$, $s(a,b) = -\mu$ for $a \ne b$, and $w(k) = \delta k$, where $\mu = 1.1$ and $\delta = 2.1$. It is remarkable that these segmental matchings from random sequences are so long and score so well. Simulations such as this suggest that understanding the distribution of score ($\max G_{ij}$) under the null hypothesis of independence is an important goal. Otherwise if the analysis of "interesting" alignments proceeds on an *ad hoc* basis, it is easy to be misled by statistically insignificant alignments. As the genome projects get underway and megabases of sequence are produced, these statistical considerations will assume more importance. The examples of this paper are of DNA sequences, but the general theory allows analysis of protein and other sequences.

```
Alignment                        max G_ij   matches   mismatches   indels

GGCAG-TCTTAGAA                   10.9       13        0            1
||||| ||||||||
GGCAGCTCTTAGAA

GGCTGATCGAGCGAGGC                11.4       15        1            1
|||| ||| ||||||||
GGCT-ATCTAGCGAGGC

GATTCAAGGACCAGATAGAGT            11         17        4            0
||||  || | ||||||||||
GATTTCAGAAACAGATAGAGT

CAGCGAAAATGCAACGCC                9.9       15        2            1
||||| || | |||||||
CAGCGCAACT-CAACGCC

AGGA-CGAATATCA-GTATAACGATGACG    11.6       23        2            4
|||| |||||| || | ||||| || |||
AGGATCGAATACCATGGATAAC-AT-ACG

CGAGCCCTCCGT                     12         12        0            0
||||||||||||
CGAGCCCTCCGT

CCCGGATGCGCAGGG                  11.9       14        0            1
||||||||||| |||
CCCGGATGCGC-GGG

AACAGCTTATA                      11         11        0            0
|||||||||||
AACAGCTTATA

AGATTA-TCAATCCA-CGT-GCG          11.2       19        1            3
|||||| | |||||| ||| |||
AGATTACTGAATCCATCGTAGCG

TAGTACTCTACTGGTC                 11         14        2            0
|| |||||||||| ||
TAATACTCTACTGCTC
```

FIGURE 1 Simulation results from comparing 10 pairs of independent, identically distributed DNA sequences of length $n = 1000$ with equally likely letters. The algorithm parameters are $\mu = 1.5$ and $\delta = 2.1$.

PROBABILITY DISTRIBUTIONS

When two random sequences of length n ($\underline{a}$ and $\underline{b}$) are written in a fixed alignment, the resulting sequence of matches and mismatches can be identified as a sequence of coin tosses. The probability that the kth toss is a head is $P(x_i = y_i)$. Our object in this paper is to study long runs of matches, which in this case are long head runs. The celebrated Erdös-Rényi law[5] gave order magnitude behavior for the longest run of heads in a sequence of n independent coin tosses. Their results actually include behavior of the longest head run containing $(1-\alpha) \times 100\%$ tails where $\alpha > P(H) = p$. For length R_n of pure head runs ($\alpha = 1.0$) their result is

$$\frac{R_n}{\log_{1/p}(n)} \rightarrow 1 \text{ with probability one},$$

while for general $\alpha > p$ their result is

$$\frac{R_n}{\frac{\log(n)}{H(\alpha,p)}} \rightarrow 1 \text{ with probability one},$$

where $H(\alpha,p) = \alpha\log(\alpha/p) + (1-\alpha)\log((1-\alpha)/(1-p))$ is relative entropy. For $\alpha = 1$, $H(\alpha,p) = \log(1/p)$ and the two results are consistent. Other work[2] gives precise results for this law and gives

$$E\{R_n\} = \log_{1/p}(n) + \frac{(.577\ldots)}{\theta} - \frac{1}{2} + r_1(n) \tag{5}$$

and

$$\operatorname{Var}\{R_n\} = \frac{\pi^2}{6\theta^2} + \frac{1}{12} + r_2(n), \tag{6}$$

where .577 ... is the Euler-Mascheroni constant, $\theta = \ln(1/p)$, and the remainders $r_1(n)$ and $r_2(n)$ are of very small, but nonvanishing magnitude. For 512 fair coin tosses, the mean is approximately 9.33 while the standard deviation is about 1.93.

The probability results quoted here are for independent and identically distributed coin tosses. To carry the results over to sequence matching, the two sequences of length n are assumed to have bases chosen independently and identically with $p = P\{\text{two bases match}\} = p_A^2 + p_C^2 + p_G^2 + p_T^2$. The formulae (5) and (6) hold with n replaced by n^2.

To consider approximate matching, we allow a fixed number of mismatches k. The mean length of the longest match with k mismatches becomes, from Ref. (2), approximately

$$\log(qn^2) + k\log\log(qn^2) + k\log\left(\frac{q}{p}\right) - \log(k!) + k + \frac{.577\ldots}{\theta} - \frac{1}{2} \tag{7}$$

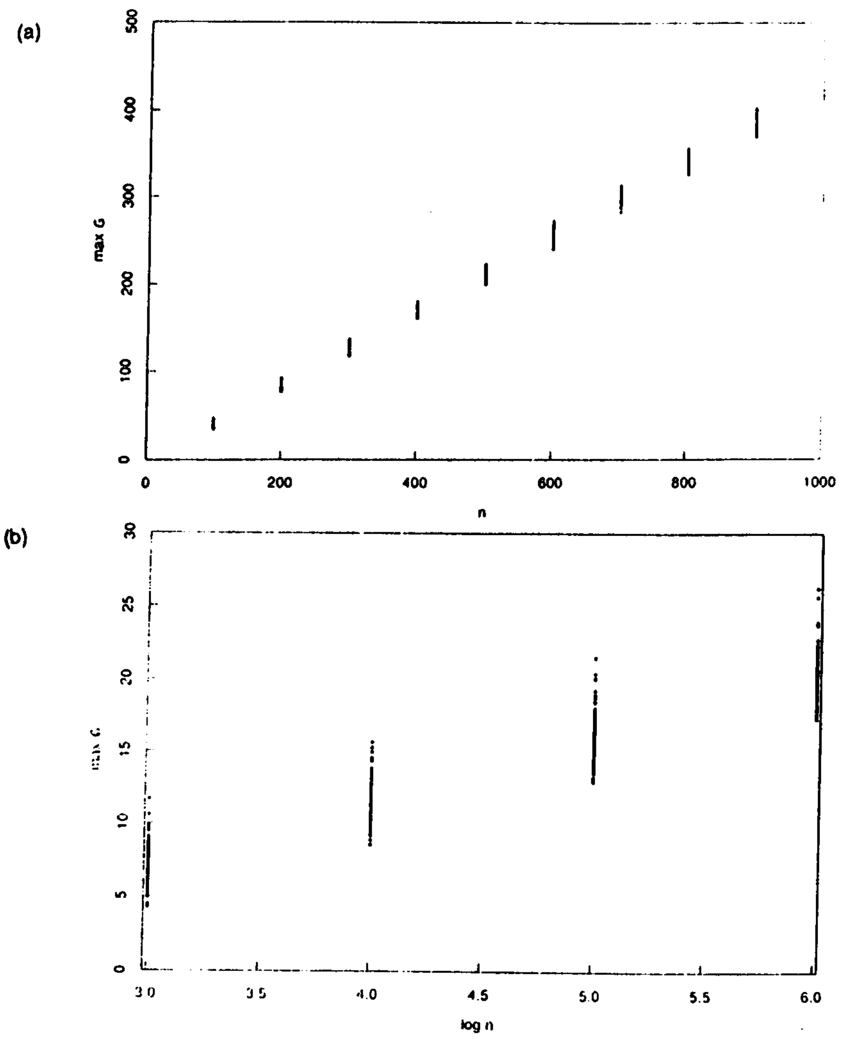

FIGURE 2 Simulation results from 100 pairs of independent, identically distributed DNA sequences of length n = 1000 with equally likely letters are compared for (a) n = 100 to 900 (growth rate is linear; the algorithm parameters are μ = 0.2 and δ = 0.5) and for (b) $n = 4^3, \ldots, 4^6$ (growth rate is logarithmic; the algorithm parameters are μ = .9 and δ = 2.0).

where $q = 1 - p$ and all log's are taken to base $1/p$. The variance remains approximately $\pi^2/(6\theta^2) + 1/12$.

The Erdös-Rényi law for the length R_n of the longest 100% head run of n coin tosses then extends to a law for the length M_n of the longest match between two sequences. We have recently shown that the length of the longest run of matches containing $(1 - \alpha) \times 100\%$ mismatches satisfies

$$\frac{M_n}{\frac{\log(n^2)}{H(\alpha,p)}} \rightarrow 1 \text{ with probability one},$$

which is the Erdös-Rényi law with n replaced by n^2. This last theorem has been obtained by use of the theory of large deviations.[4]

Considering these results, it is not surprising that $\log(n)$ laws hold far beyond the longest exact head run or match. The expected behavior of $\max G_{ij}$ is of importance in evaluating sequence comparisons. If a located match is at or below that expected from random sequences of similar composition, then the match should not be further considered without additional biological information. These distributions have been shown to fit biological sequences quite well[14] for algorithm parameters not covered by the theorems above. We have also proven that $\max G_{ij}$ undergoes a phase transition.[13] For larger values of (μ, δ), $\max G_{ij}$ grows proportional to $\log(n)$ and for smaller values $\max G_{ij}$ grows linearly with sequence length. There are only two modes of behavior at this precision. The logarithmic and linear regions of this two-dimensional parameter space have been determined numerically in a Monte Carlo study.[14] See Figure 2 for illustrations of the growth of score ($\max G_{ij}$) with sequence length. These results help those analyzing macromolecular sequences to proceed in a much less *ad hoc* manner.

A HEURISTIC CALCULATION

Our results provide substantial intuition when confronting more complicated problems as well. As an example, we provide a tentative analysis of a multiple inference procedure presented in Altschul and Erickson.[1] They propose as a measure of sequence similarity the minimal attained significance among all runs of matches, minimizing over all possible run lengths. Attained significance is computed using a binomial model, implicitly computing significance under the model of independently generated letters in each of the compared sequences. Citing a lack of theoretical development, they use a curve-fitting approach to parametrize their proposed test. In the discussion below, we will show how the previously discussed results suggest a framework in which an alternative, theoretically motivated parametrization provides a better fit.

Specifically, in terms of the notation of the previous section, let $M_n(k)$ denote the length of the longest match between two sequences of length n of independently generated letters taken from the same distribution. Inspired by Cramer's treatment of large deviations theory, approximate the log-probability of observing k or fewer mismatches in match length t as

$$-tH\left(1 - \frac{k}{t}, p\right) - \frac{1}{2}\ln(2\pi k), \tag{8}$$

an approximation good to order k/t. Note that for fixed k the attained significance is increasing in t. Hence, to a crude order of approximation, the logarithm of attained significance might be expected to be

$$S_n = -M_n(K)H(1,p) - \frac{1}{2}\ln(2\pi K),$$

where the random variable K indexes the most significant of all k-interrupted matches between the two n-sequences. Of importance is only the observation that approximate significance will pick one of the longest k-interrupted matches, each of which is approximately distributed as integerized extreme value with centers given by Arratia et al.[2]

There is reason to believe that the most significant of the k-interrupted matches should appear with relatively small k. The argument depends on the approximate independence—for small k—of the lengths of the longest k-interrupted runs of matches. This is a consequence of the conditional uniform distribution of the locations of mismatches given an extremely rich pattern of matches. A common renewal-theoretic clumping argument suggests that for large k, one should expect to observe clumps of unusually long k-interrupted runs of matches. Hence, the most significant among the k-interrupted runs should typically occur for small k, and we use Eq. (7) to conjecture the form

$$S_n = \ln(n^2) + \alpha \ln\ln(n^2) + \beta\frac{\ln\ln(n^2)}{\ln(n^2)} + \gamma + V, \tag{9}$$

where V is the integerized extreme value with approximate variance $\pi^2/6 + 1/12 = 1.73$, $\alpha = E\{K\}$, and $\gamma = E\{\mu_K\}$ consolidates those terms of Eq. (7) which are not

TABLE 1 Three Models by Ordinary Least Squares

model	$\sqrt{R^2}$	1000 × MSE
1. $s - \ln(n^2) = \alpha \ln\ln(n^2) + \beta$	.95	3.02
2. $s - \ln(n^2) = \alpha \ln(n^2) + \beta$	.96	2.34
3. $s - \ln(n^2) = \alpha \ln\ln(n^2) + \beta\frac{\ln\ln(n^2)}{\ln(n^2)} + \gamma$	.98	1.44

coefficients of $\ln(t)$ or $\ln\ln(t)$. The correction term with coefficient β is of the sort used in Arratia et al.[2] to correct for finite sample size. For a theoretical discussion showing the form of such corrections, see de Bruijn,[5] section 2.4.

We present the results of fitting Eq. (9) using the data of Altschul and Erickson,[1] in which are reported means of 1000 simulation experiments realizing the minimal attained significance for nine levels of n ranging from 70 to 518. Specifically, we take the sample mean of the 1000 attained significance levels for each value of n using the formula $s = u + .5772/\lambda$, where values of u and λ are taken from Altschul and Erickson,[1] Table II. We fit the three models in Table 1 by ordinary least squares. All three models regress against the independent variable $s - \ln(n^2)$, motivated by the heuristic argument above. It is nearly impossible to select among the models for data analytic reasons alone; correlations among pairs of the three explanatory variables used all exceed .997 in absolute value.

The reassuring feature of the heuristic specification of model 3 is the mean square error after fitting. The tabled values would estimate the variance of an integerized extreme value, if the specification, Eq. (9), were correct. Note that 1.44 is closest of all three specifications to $\pi^2/6 + 1/12$, although all three models yield confidence intervals for the variance which include 1.73.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation and the National Institutes of Health.

REFERENCES

1. Altschul, S. F., and B. W. Erickson. "A Nonlinear Measure of Subalignment Similarity and Its Significance Levels." *Bull. Math. Biol.* 48 (1986):617–632.
2. Arratia, R., L. Gordon, and M. Waterman. "An Extreme Value Theory for Sequence Matching." *Ann. Statist.* 14 (1986):971–993.
3. Arratia, R. A., L. Gordon, and M. S. Waterman. "The Erdös-Rényi Law in Distribution, for Coin Tossing and Sequence Matching." Submitted to *Ann. Statist.*.
4. Arratia, R., and M. Waterman. "The Erdös-Rényi Strong Law for Pattern Matching with a Given Proportion of Mismatches." *Ann. Prob.*, to appear.
5. de Bruijn, N. G. *Asymptotic Methods in Analysis.* New York: Dover, 1981.
6. Goad, W. B., and M. I. Kanehisa. "Pattern Recognition in Nucleic Acid Sequences. 1. A General Method for Finding Local Homologies and Symmetries." *Nucl. Acids Res.* 10 (1982):247–263.
7. Gotoh, O. "An Improved Algorithm for Matching Biological Sequences." *J. Mol. Biol.* 162 (1982):705–708.
8. Needleman, S. B., and C. D. Wunsch. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins." *J. Mol. Biol.* 48 (1970):444–453.
9. Sellers, P. "The Theory and Computation of Evolutionary Distances: Pattern Recognition." *J. Algorithms* 1 (1980):359–373.
10. Smith, T. F., and M. S. Waterman. "Identification of Common Molecular Subsequences." *J. Mol. Biol.* 147 (1981):195–197.
11. Waterman, M. S., T. F. Smith, and W. A. Beyer. "Some Biological Sequence Metrics." *Adv. Appl. Math.* 20 (1976):367–387.
12. Waterman, M. S., and M. Eggert. "A New Algorithm for Best Subsequence Alignments with Application to tRNA-rRNA Comparisons." *J. Mol. Biol.* 197 (1987):723–728.
13. Waterman, M. S., L. Gordon, and R. Arratia. "Phase Transitions in Sequence Matching and Nucleic Acid Structure." *Proc. Nat. Acad. Sci.* 84 (1987):1239–1243.
14. Waterman, M. S. *Mathematical Methods for DNA Sequences.* Boca Raton, FL: CRC Press, 1989.