Mathematics and the Genome Projects

Warren J. Ewens

Abstract. The human genome project, and the parallel genome projects for other species, will soon produce data that will require entirely novel mathematical and statistical analyses, as well as completely new computer algorithms. Here I review mathematical and statistical aspects of what has become by far the most frequently used statistical analysis used in genome sequence data analysis so far, namely the BLAST (Basic Local Alignment Search Tool) statistical analysis.

1. Background

BLAST is a method for

(i) finding high-scoring similarity segments of two aligned DNA or protein sequences,

(ii) assessing the statistical significance of any given segment score.

Our interest here is in (ii).

Although BLAST is most frequently used to compare two protein (that is, amino acid) sequences, it is convenient to nucleotide (DNA) sequences first, to establish the principles involved.

Consider the simple case of the two aligned DNA sequences given in (1):

$$\begin{array}{cccccccccccccccccccccc} g & g & a & g & a & c & t & g & t & a & g & a & c & a & g & c & t & a & a & t & g & c & t & a & t & a \\ c & a & a & c & g & c & c & c & t & a & g & c & c & a & c & g & a & g & c & c & c & t & t & a & t & c \end{array} \tag{1}$$

Suppose we give a score $+1$ if the two nucleotides in corresponding positions match (i.e. are the same) and a score of -1 if they do not match. Then as we go along comparing the two sequences, starting at the left, the accumulated score performs a two-dimensional random walk. In the above example, this walk is depicted in figure 1.

The null hypothesis asserts that there is no similarity between the two sequences, that is that one is generated at random with respect to the other. It is assumed that the null hypothesis probability p of a match at any site is less than the probability $q = 1 - p$ of a mismatch. (If all four nucleotides are equally frequent, $p = 1/4$, so that in this case the inequality requirement certainly holds.) If the null hypothesis is true, the walk then generally drifts downwards to the right. As it does so it passes through a sequence of ladder points, shown in figure 1 as

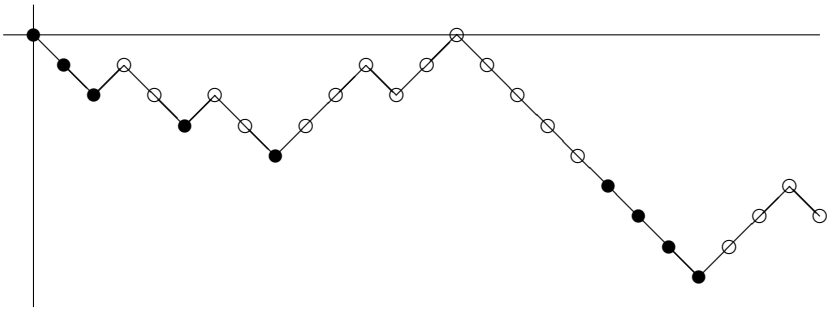

FIGURE 1

filled-in circles, and indicating a sequence of new minima sequentially achieved by the walk.

After reaching any ladder point the walk has the potential of making an "upwards to the right" excursion, and such an excursion will arise for any sub-sequence of the DNA where there is some similarity between the two sequences being compared. Any such excursion will achieve a maximum height Y above the ladder point from which it starts, perhaps after some zig-zags, and we denote the height of the upward excursion after ladder point i by Y_i. (We denote the origin as ladder point 0.) If the walk proceeds immediately from ladder point i to ladder point $i+1$, the value of Y_i is zero. In figure 1, for example, the observed values of Y_0, Y_1, Y_5, Y_6 and Y_7 are all zero, while $Y_2 = Y_3 = 1$ and $Y_4 = 4$.

BLAST theory assesses whether there is a significant similarity betwen the two sequences by using as test statistic the maximum $Y_{\max}$ of these excursion heights, and referring the observed value of $Y_{\max}$ to its null hypothesis probability distribution for significance.

In the case just described, this null hypothesis distribution is found by considering the properties of a simple random walk with step sizes ± 1, the theory for which is well known. It is however useful to review some of this theory. We denote the size of any step taken in the walk by S. Then $S = +1$ with probability p and $S = -1$ with probability $q = 1 - p$, so that the mean $E(S)$ of S is $p - q$ and the moment-generating function of S is

$$m(\theta) \;=\; qe^{-\theta} \,+\, pe^{+\theta}\,. \tag{2}$$

The (unique real positive) value of θ for which this moment-generating function takes the value 1 is denoted by λ. It is immediate in this case that

$$\lambda \;=\; \log\frac{q}{p}\,. \tag{3}$$

It is a standard result of random walk theory that if Y is the maximum height achieved by the walk after reaching any ladder point and before reaching the next, then asymptotically

$$\text{Prob}\,(Y \le y-1) \sim 1 - Ce^{-\lambda y} \tag{4}$$

where $C = 1 - e^{-\lambda}$. The asymptotic relation (4) is in a form similar to that of a geometric distribution, and we call the distribution of Y a "geometric-like" distribution.

Since the test statistic used in BLAST is $Y_{\max}$, it is necessary to find the null hypothesis mean number of ladder points in the walk, since the null hypothesis probability distribution of $Y_{\max}$ will depend on this mean. This mean depends on the length of the two sequences being compared, and we denote this length by N.

To find the mean number of ladder points we first find the mean number of steps in the walk (or comparisons of the two DNA sequences) between one ladder point and the next. This calculation is best approached using Wald's identity. Suppose that j is the (random) number of steps from one ladder point to the next, and define T_j as the sum of the sizes of these steps. Differentiating throughout in Wald's identity

$$E\left(e^{\theta T_j}(m(\theta))^{-j}\right) = 1 \tag{5}$$

with respect to θ and then setting $\theta = 0$, we obtain the equation

$$E(T_j) = E(j)E(S)\,. \tag{6}$$

Equation (6) enables us to find $E(j)$ immediately for the random walk. The eventual total displacement T_j in the walk is identically -1, since any one ladder point is unit distance below the previous one. Since $E(S) = p - q$, $E(j) = (q-p)^{-1}$. This implies that the mean number μ of ladder points along the entire sequence comparison of length N is given by

$$\mu = N/E(j) = N(q-p)\,. \tag{7}$$

Thus $Y_{\max}$ can be taken as the maximum of μ random variables, each having the asymptotic distribution given by (4).

Unfortunately there is no limiting ($N \to +\infty$) probability distribution for $Y_{\max}$. However bounds for the distribution of Y can be found by bounding the properties of the geometric-like distribution by those of two continuous exponential-like random variables. From this, if y is the observed value of $Y_{\max}$, then anticipating the statistical use of these limiting distributions,

$$1 - e^{-\mu C\, e^{-\lambda y}} \le \text{P-value} \le 1 - e^{-\mu C\, e^{-\lambda(y-1)}} \tag{8}$$

where the P-value is that associated with an observed value y of the random variable $Y_{\max}$. Often a conservative P-value is wanted, that is to say a number known to be greater than or equal to the true P-value. The inequalities (8) show immediately that, for large values of y,

$$\text{conservative P-value} \approx 1 - e^{-\mu C\, e^{-\lambda(y-1)}}\,. \tag{9}$$

In many applied cases $\mu C\, e^{-\lambda(y-1)}$ is quite small. When this is so,

$$\text{approximate conservative P-value } \approx \mu C\, e^{-\lambda(y-1)}\,. \tag{10}$$

Once the values of C and λ are known, this P-value can be calculated immediately.

2. Protein Sequences

In practice BLAST theory relates to cases that are much more complicated than the simple DNA example described above. First it is normally applied to the comparison of two protein sequences, that is sequences of amino acids, and uses scores other than the simple scores $+1$ and -1 for matches and mismatches discussed above. There are 20 amino acids and thus 400 possible amino acid comparisons at any position in the comparison of two protein sequences. If the 20 amino acids are numbered as amino acids 1, 2, ..., 20 in some agreed order, these scores are described by the entries in a 20×20 "substitution matrix", with a score $S(j,k)$ being allocated at any position if amino acid j occurs in the first sequence and amino acid k occurs in the second sequence at that position. It is always true in such a matrix that any main diagonal entry $S(j,j)$ is positive, and for comparatively rare amino acids this score is usually comparatively large. Mismatch scores are usually negative, although mismatch scores for two amino acids having similar properties are sometimes small and positive. Since the score for any amino acid comparison is used to define to upward or downward movement of the random walk describing the accumulated score for that protein comparison, they lead to random walks more complex than that depicted in figure 1. For example, if

$$\begin{aligned} &S(1,4) = -2\,, \quad S(1,8) = +1\,, \quad S(5,9) = -4\,, \\ &S(11,11) = +15\,, \quad S(14,14) = +5\,, \quad S(6,6) = +7\,, \end{aligned}$$

and if two (short) aligned amino acid sequences are (in terms of the agreed amino acid enumeration system)

$$\begin{array}{cccccc} 14 & 1 & 1 & 11 & 5 & 6 \\ 14 & 4 & 8 & 11 & 9 & 6 \end{array} \tag{11}$$

the accumulated score performs a random walk successively going through the points

$$(1,5),\ (2,3),\ (3,4),\ (4,19),\ (5,15),\ (6,22)\,. \tag{12}$$

To discuss BLAST it is necessary to consider arbitrary scoring schemes (that is, arbitrary 20×20 substitution matrices), and thus aspects of the general theory of random walks. The question of the choice of the scoring scheme will be taken up later, and for the moment we take the scoring scheme as given.

3. The BLAST Protein Sequence Random Walk

As just noted, the comparison of two protein sequences induces a random walk similar to, but much more complicated than, the DNA sequence comparison shown

in figure 1. Nevertheless the basic features described in the DNA sequence comparison continue to hold. The walk goes through a sequence of ladder points, and after reaching any ladder point has the potential to make an upward excursion. As above, we define the maximum height achieved by the excursion starting at ladder point i and before reaching ladder point $i+1$ by Y_i, with Y_i being taken as zero if the walk proceeds immediately from ladder point i to ladder point $i+1$. In this more general case the asymptotic distribution of each Y is again given by the relation (4), but with new definitions of the parameters C and λ.

It is also necessary to generalize the parameter μ of equation (7). In practice the sequence comparison is usually between a comparatively short "query" sequence of length N_1 which is of interest to the investigator and a very long database sequence of length N_2. All possible alignments of the short sequence to the long one are made, resulting in some $N_1 N_2$ amino acid comparisons. The random walk thus continues for approximately $N_1 N_2$ steps, and the test statistic is again taken as $Y_{\max}$, the maximum of the Y_i values in this extremely long random walk.

4. Parameter Determination

The parameter λ for the protein sequence random walk is found from a direct generalization of the procedure that led to (3) for the DNA case. If the frequency of amino acid j in the query sequence is p_j and the frequency of amino acid k in the database sequence is p'_k, λ is defined as the unique non-zero solution of the equation

$$\sum_{j,k} p_j p'_k e^{\lambda S(j,k)} \;=\; 1\,. \tag{13}$$

We require the mean score $\sum_{j,k} p_j p'_k S(j,k)$ to be negative, and this implies that $\lambda > 0$. It also implies that the random walk induced by the sequence comparison generally drifts down to the right, passing as it does so through a sequence of ladder points. In this more general case it is no longer necessary that ladder point $i+1$ is unit distance below ladder point i.

The parameter C is more difficult to determine. Suppose that the possible step sizes in the random walk, that is the set of scores in the substitution matrix, are

$$-c,\, -c+1,\, \ldots,\, 0,\, \ldots,\, d-1,\, d\,, \tag{14}$$

and that these steps have respective probabilities

$$p_{-c},\, p_{-c+1},\, \ldots,\, p_d\,. \tag{15}$$

All the analyses below consider a walk starting at the origin. It is convenient to start by considering an unrestricted random walk with no stopping points. Such a walk must eventually drift down to $-\infty$, since the mean step size is negative. Before doing so, however, it might visit various positive values. We let Q_k be the probability that the walk visits the positive value $+k$ before reaching any other positive value and before eventually drifting down to $-\infty$. Since it is possible

that the walk never visits any positive value, $\sum_{k=1}^{+\infty} Q_k < 1$, and we write $\bar{Q} = 1 - Q_1 - Q_2 - Q_3 - \ldots$. Wald's identity (5) with $\theta = \lambda$ can be used to show that

$$\sum_{k=1}^{d} Q_k e^{k\lambda} = 1\,. \tag{16}$$

We next consider a restricted walk where the walk which stops once it reaches one or other of the possible ladder points, that is one or other of the points $-c$, $-c+1$, $\ldots$, -1. Let R_{-j} be the probability that the first ladder point is at $-j$. Then it can be shown, after considerable analysis, that

$$C \;=\; \frac{\bar{Q}\left(1 \,-\, \sum_{j=1}^{c} R_{-j} e^{-j\lambda}\right)}{(1-e^{-\lambda})(\sum_{k=1}^{d} k Q_{+k}\, e^{k\lambda})}\,, \tag{17}$$

the sums in the numerator and the denominator extending over the finite set of values of j and k for which R_{-j} and Q_{+k} are respectively non-zero.

We next calculate A, the mean number of steps that the walk takes before first reaching either $-c$, $-c+1$, or $\ldots$ or -1. To do this we once again use equation (6). The mean net dispacement when the walk stops is $\sum_{j=1}^{c} -jR_{-j}$, where R_{-j} is the probability that the walk finishes at $-j$. The mean step size $E(S)$ is $\sum_j jp_j$, assumed to be negative. Thus A, the mean number of steps taken before the walk reaches the first ladder point, is given by

$$A \;=\; \frac{\sum_{j=1}^{c} jR_{-j}}{-\sum_j jp_j}\,. \tag{18}$$

The calculation of both C and A requires calculation of the R_{-j} values and that of C requires also the calculation of the Q_{+k} values. Fortunately various methods exist for these calculations, including rapidly converging series approximations.

The mean number μ of ladder points is then equal to N_1N_2/A, and the test statistic $Y_{\max}$ is then the maximum of μ random variables, each having the geometric-like probability distribution (4), with λ now defined in (13) and C now defined in (17). Bounds and conservative approximations for P-values associated with an observed value y of $Y_{\max}$ are given by (8), (9) and (10), with these new definitions of μ, C and λ.

5. The BLAST Testing Procedure

The above calculations are all that is needed to define the statistical testing involved in a BLAST testing procedure if $Y_{\max}$ is used as test statistic. The null hypothesis being tested is that the two protein sequences are random with respect to each other and the alternative hypothesis, for the moment, is that there is a tendency for matching between identical amino acids. Fortunately some simplifications are possible in the calculations. Thus if $K = Ce^{-\lambda}/A$, the approximate

conservative P-value given in equation (10) becomes

$$\text{approximate conservative P-value} \;=\; N_1 N_2 K\, e^{-\lambda y}\,. \tag{19}$$

K can often be calculated quite quickly, thus obviating the need to compute C and A.

It can be argued that use of $Y_{\max}$ as test statistic ignores the information provided by high-scoring sub-subsequences other than that associated with $Y_{\max}$. A generalization of the testing procedure is available by using a "normalized score". It is clear that the value of $Y_{\max}$ would be doubled if all entries in the substitution matrix were doubled. The expression

$$\lambda\, Y_{\max} \;-\; \log(N_1 N_2 K)\,, \tag{20}$$

denoted here by S', is called the normalized score, and is not subject to the same comment, since as the definition of λ implicit in equation (13) shows that doubling the entries in the substitution matrix halves the value of λ. An approximate conservative P-value associated with an observed value s' of S' is, from (19),

$$\text{approximate conservative P-value} \;=\; e^{-s'}\,. \tag{21}$$

Denote the j^{th} highest normalized score by $S'_{(j)}$, so that $S' = S'_{(1)}$, and write $T_N = S'_{(1)} + S'_{(2)} + \ldots + S'_{(N)}$. The density function $f(t)$ of T_N is, approximately,

$$f(t) \;=\; \frac{e^{-t}}{N!(N-2)!} \int_0^{+\infty} y^{(N-2)} \exp(-e^{(y-t)/N})\, dy\,. \tag{22}$$

This density function can be used to find the approximate expression

$$\text{Prob}\,(T_N \;\geq\; t_N) \;\approx\; \frac{e^{-t_N} t_N^{N-1}}{N!(N-1)!} \tag{23}$$

for the P-value for any observed value t_N of T_N. In the case $N = 1$, this is identical to the approximation for the P-value of the observed value s' of S' given in (21).

6. The Substitution Matrix

So far we have taken the elements in the substitution matrix as given. It is possible to view the testing procedure as a non-parametric one and to regard these entries as simply reasonable scores. However it is also possible to derive these entries from likelihood ratio principles. A stochastic evolutionary model is set up and from this one can calculate, for some agreed degree of evolutionary time divergence between the two sequences examined, a log likelihood ratio

$$\log \frac{q(j,k)}{p_j p'_k}\,, \tag{24}$$

where $q(j,k)$ is the probability of amino acid j in the query sequence and amino acid k in the database sequence under the evolutionary model chosen. Then the (j,k) entry in the substitution matrix is proportional to this log likelihood ratio.

Under this approach it is possible to assess the implications of making incorrect assumptions in the evolutionary model chosen.

7. Further Work

The BLAST model described in above is the simplest one possible, and the versions of BLAST that are used in practice are more sophisticated than that described here. Further, extensions even of these more sophisticated versions continue to appear. An important generalization allows gaps in the sequence alignments. Some of the parameter calculations for these do not follow from explicit theory but follow rather from simulations and approximations derived from statistical regression theory. A full generalization of the theory to cover such cases is one central area of current research. The development of gapped BLAST makes the use of sum scores such as T_N, derived from two or more well-matching subsequences, of lesser importance than before. A second generalization is to PSI (position specific iterated) BLAST, again having very complex theory. Another generalization is to the case of Markov-dependent protein sequences, a necessary generalization since Markov dependence does appear to be the case. All of these extensions will require further difficult theoretical analysis and support.

Department of Biology,
University of Pennsylvania,
221 Leidy Laboratories,
415 S. University Avenue,
Philadelphia PA 19104, USA
E-mail address: wewens@sas.upenn.edu