

Estimation and Reliability of Molecular Sequence Alignments

Jeffrey L. Thorne¹, Gary A. Churchill

*Biometrics Unit, 337 Warren Hall
Cornell University
Ithaca, NY 14853
U.S.A.*

BU-1201-M

April 1993

Send Proofs to: Jeffrey L. Thorne

¹ Current Address: *Program in Statistical Genetics, Department of Statistics, Box 8203, Raleigh NC 27695-8203, U.S.A.*

Summary

The problem of estimating the relatedness of a pair of biological sequences is addressed. A stochastic model of sequence evolution is described that allows insertion and deletion as well as replacement of amino acid residues (or substitution of nucleotides) over time. An expectation-maximization (EM) algorithm that obtains maximum likelihood estimates of the model parameters is introduced. The method assumes the sequences are related by descent from a common ancestor but the alignment (i.e., the precise evolutionary correspondence between residues in each sequence) is unknown. Results from the E-step of the EM algorithm are used to assess the likelihood that any pair of residues are related by direct descent from a common ancestor.

1 Introduction

Genes that encode proteins with the same or similar biological functions are often related by descent from a common ancestral gene. This common origin can be reflected, at the DNA level, by similarity of their nucleotide sequences and, at the protein level, by similarity of their amino acid sequences. When two protein sequences A and B are related by descent, an amino acid residue in sequence A may have a corresponding residue in sequence B . Alternatively, it may lack a corresponding residue in sequence B because either an insertion event occurred in the lineage from the common ancestral sequence to A or a deletion event occurred in the lineage from the common ancestral sequence to B . If a replacement¹ of

¹In the standard terminology of molecular evolution, a change of one amino-acid type to another in a protein is a replacement and a change of one nucleotide type to another in a DNA or RNA molecule is a substitution. In this paper, we will emphasize protein evolution and hence use the term replacement. DNA sequences can be analyzed using nearly identical methods.

one residue type for another has occurred in either of the lineages leading to *A* or *B*, corresponding residues may not be the same amino acid type. The problem of estimating rates of insertion, deletion and replacement are fundamental to evolutionary biology as is the problem of establishing the correspondence between the residues in two (or more) related sequences. In this paper, we apply model-based statistical methods to address these problems.

A sequence alignment is a hypothesis about the evolutionary correspondence between the residues in a pair of sequences. A common representation of an alignment (Figure 1A) is to exhibit the residues of one sequence on a line above those of the other sequence. Corresponding residues appear stacked one above the other. When two residues in a column are the same type, the alignment position is termed a match. When corresponding residues are different, the alignment position is termed a mismatch and at least one replacement event must have occurred. In this “stacked” alignment representation, a residue that has no corresponding residue in the other sequence is said to be opposite a gap. If we consider sequence *A* to be ancestral and sequence *B* to be its descendant, gaps in sequence *A* are the result of insertion events and gaps in sequence *B* are the result of deletion events.

Pairwise sequence alignment has generally been treated as a computational optimization problem (e.g., Waterman 1984) and dynamic programming solutions are available to solve this problem in time proportional to the product of the sequence lengths. However, the focus on computation has drawn attention away from the biological process of evolution. As a result, many alignment algorithms in widespread use make little or no reference to the underlying biological model.

Bishop and Thompson (1986) were the first to consider pairwise sequence alignment in a likelihood framework. Estimates of parameters in their model could be obtained and subsequently used to find an optimal alignment. The

Bishop and Thompson method has several limitations. An exact treatment of the likelihood is not feasible and the approximations used become less accurate for comparison of more distantly related sequences. Also, their model is restricted to allow only single-residue insertions and deletions. Thorne, Kishino, and Felsenstein (1992) introduced a model that overcomes some of these limitations by allowing multiple-residue insertion and deletion events as well as regional heterogeneity in the replacement process. Furthermore, its mathematical tractability eliminates the need for approximations. This model forms the basis of the present work.

The likelihood of two sequences is a function of their alignment and the probabilities assigned to the different types of evolutionary events. In general, the true alignment will be unknown. Therefore, we will consider the marginal likelihood – a sum over all possible alignments of the conditional likelihood for a given alignment. The main result of this work is an expectation-maximization (EM) algorithm (Dempster, Laird and Rubin 1977) that maximizes the marginal likelihood. In section 2, we describe the evolutionary model. We first present a simple model and then discuss generalizations that enhance its realism. In section 3, we consider the likelihood function and, in section 4, the EM algorithm is described. For the sake of clarity, the EM algorithm is presented in the context of the simple evolutionary model. In the E-step, the probabilities of each possible correspondence between residue pairs is computed. We discuss how these probabilities can be used to assess the reliability of specific correspondences within an alignment. In section 5, an example is provided to illustrate the method. We close with a brief discussion of the relevance of this work in the context of statistical methods for molecular evolution. Details of the likelihood computations, the E-step of the simple EM algorithm, and computational aspects pertinent to the more general model are included in the Appendices.

2 The Evolutionary Model

Our evolutionary model is composed of two independent processes, an insertion-deletion process and a substitution process. If we let Z denote the ancestral sequence and assume independence in the two lines of descent, the likelihood can be expressed as

$$\Pr(A, B) = \sum \Pr(Z) \Pr(A | Z) \Pr(B | Z) \quad (1)$$

where the summation runs through all possible ancestral sequences. This is clearly an intractable problem. However, if we assume that the two processes are jointly reversible then one sequence may be considered as ancestral and the other as its descendant (Felsenstein 1981). The likelihood can be expressed as

$$\Pr(A) \Pr(B | A) = \Pr(B) \Pr(A | B) , \quad (2)$$

where $\Pr(A)$ and $\Pr(B)$ are the equilibrium probabilities of sequences A and B . The reversibility property yields computational feasibility and we will restrict our attention to reversible models below.

2.1 The replacement process

The process by which one amino acid residue replaces another is modelled as a continuous time Markov chain on the state space of possible residues. Let $f_{ij}(t)$ denote the probability that a residue of type i is replaced by a residue of type j after an amount of evolution t . The transition probabilities can be expressed in the form

$$f_{ij}(t) = e^{Qst}, \quad (3)$$

where Q is the known rate matrix, s is a rate constant and t corresponds to time. Because the values of s and t cannot be independently estimated, the amount of replacement is measured by the value of the composite parameter st .

An empirically derived amino acid replacement model was introduced by Dayhoff, Schwartz, and Orcutt (1978) and has recently been refined (Gonnet, Cohen and Benner 1992; Jones, Taylor and Thornton 1992). Empirically derived amino acid replacement models are widely used in practice. The Dayhoff model measures the amount of replacement in units called “PAMs” (accepted point mutations per 100 residues). An amount $st = N$ PAMs means an average of N amino acid replacements per 100 residues have occurred in the evolutionary lineages that separate two sequences. The empirical data of Dayhoff et al. can be used to generate a transition probability matrix for any specified value of st . Kishino, Miyata, and Hasegawa (1990) explain how the rate matrix Q can be derived from empirical data. Implicit in this rate matrix is, for each amino acid type i , the value of its equilibrium frequency π_i .

As mentioned earlier, the techniques introduced here can also be employed for the analysis of DNA sequences. A number of reversible substitution processes have been proposed (see Tavare 1986 for a review) that could be used in conjunction with these techniques.

2.2 The insertion–deletion process

The process by which residues are inserted into and deleted from a protein sequence is modelled as a birth-death chain (Feller, 1967). In its simplest form, the process acts on single residues in the sequence and has been described by Thorne, Kishino, and Felsenstein (1991). The model is summarized here for completeness.

It is convenient to imagine the birth–death process as acting on links that connect the residues in the sequence. There are two types of links: normal and immortal. Each residue in the sequence is associated with a normal link to its immediate right. An immortal link is assumed to exist at the extreme left of the sequence.

The immortal link is not associated with any residue and, as its name implies, it cannot die. When a normal link dies, its associated residue is deleted from the sequence. The death rate per normal link is μ . When a new link is born, the link and its associated residue are inserted into the sequence to the immediate right of the parent link. The probability that the associated residue is type i is the equilibrium frequency of the type (π_i). All newborn links are normal and each link in the sequence gives birth to new links with rate λ .

There are three categories of transition probabilities for this process: $p_n(t)$ is the probability that after a time of duration t , n links are descended from a normal link and one of them is the original link; $p'_n(t)$ is the probability that after a time of duration t , n links are descended from a normal link and the original link has died; and $p''_n(t)$ is the probability that after a time of duration t , the immortal link has n descendants including itself. A system of differential equations governing these transition probabilities can be formulated and the solutions are known (see Thorne et al. 1991). By definition, $p_0(t) = p''_0(t) = 0$. The remaining transition probabilities are

$$\begin{aligned}
 p_n(t) &= e^{-\mu t}(1 - \lambda\beta(t))(\lambda\beta(t))^{n-1}, \quad n > 0; \\
 p'_n(t) &= (1 - e^{-\mu t} - \mu\beta(t))(1 - \lambda\beta(t))(\lambda\beta(t))^{n-1}, \quad n > 0; \\
 p'_0(t) &= \mu\beta(t); \\
 p''_n(t) &= (1 - \lambda\beta(t))(\lambda\beta(t))^{n-1}, \quad n > 0,
 \end{aligned} \tag{4}$$

where

$$\beta(t) = \frac{1 - e^{(\lambda-\mu)t}}{\mu - \lambda e^{(\lambda-\mu)t}}.$$

Under this model, the equilibrium probability of a sequence of length n is

$$\gamma_n = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n,$$

subject to the constraint $0 < \lambda < \mu$.

2.3 Alignments

An alternative to the “stacked” representation of a pairwise alignment is a directed graph displayed as a $(n_A + 1)$ by $(n_B + 1)$ grid where n_A and n_B are the lengths of the sequences (Figure 1B). Sequence A is shown along the left margin and sequence B is shown along the top. Nodes in the graph correspond to pairs of links, one from each sequence. An alignment path α is a connected sequence of arcs that traverse the matrix from the upper left node $(0,0)$ to the lower right node (n_A, n_B) by a series of eastern (\rightarrow), southeastern (\searrow) and southern (\downarrow) moves. Correspondence between the j^{th} base of sequence A and the k^{th} base of sequence B is represented as a southeastern arc that connects nodes $(j-1, k-1)$ and (j, k) . The reversible nature of the evolutionary model allows us to view sequence A as ancestral and sequence B as its descendant. From this perspective, a southern arc represents a deletion of a base, and an eastern arc represents an insertion of a base. An advantage of the directed graph form of representation is the ability to display multiple alternative alignments. Notice that the alignment path α contains information about the insertion-deletion process but not the replacement process.

Because of the convention that newborn links are inserted to the right of their parental link, the following two alignments represent two distinct evolutionary histories

$$(1) \quad \begin{array}{l} A \quad - \quad C \\ A \quad G \quad - \end{array}$$

$$(2) \quad \begin{array}{ccc} A & C & - \\ & A & - G \end{array} .$$

In Alignment (1), the A link is the parent of the G link whereas in alignment (2) the C link is the parent of the G link. Alignment (2) implies that the G link was inserted before the C link was deleted. Alignment (1) does not allow chronological ordering of the two events. The difference in interpretation of the two alignments results in the alignments having distinct probabilities. It is convenient for computational reasons to consider each case of a deletion immediately to the left of an insertion as a single southeastern arc in the directed graph instead of as a southern arc followed by an eastern arc. This type of composite event will (usually) be relatively improbable and will be termed a “special” southeastern arc. A southeastern arc that indicates a match or a mismatch will be termed a “normal” southeastern arc.

The following notation will be used below. Let α denote the entire alignment path and let $\alpha(i, j, k)$ denote an arc of type i that enters node (j, k) where:

$$i = \begin{cases} 0 & \text{if a southern arc enters node } (j, k) \\ 1 & \text{if a normal southeastern arc enters node } (j, k) \\ 2 & \text{if a special southeastern arc enters node } (j, k) \\ 3 & \text{if an eastern arc enters node } (j, k) \end{cases}$$

Let a_m and b_n respectively denote the observed residues at the m^{th} position of sequence A and the n^{th} position of sequence B. Let A_m and B_n respectively represent the subsequence consisting of the first m bases of A and the subsequence consisting of the first n bases of B.

2.4 Extensions of the Simple Model

Multiple-residue insertion and deletion events Insertions and deletions involving two or more adjacent residues are commonly observed in protein (and DNA) sequences. It is desirable to model these as single events. In the general insertion-deletion process, normal links can be associated with one or more amino acid residues. A normal link and its associated residues are termed a fragment. The birth-death process of links is the same as for the single residue model except that the death of a link causes the deletion of all associated residues. Fragment boundaries are fixed so that, if a group of adjacent residues is inserted as a single evolutionary event, a subsequent event that deletes at least one of the residues must delete them all. We will assume the number of associated amino acid residues per normal link is geometrically distributed. Thus, the probability that a normal link is associated with exactly n residues is

$$h(n) = (1 - r)r^{n-1} \quad 0 \leq r < 1 \quad n = 1, 2, \dots .$$

The relationships between fragments of the ancestral sequence and fragments of the descendant sequence can be represented by a path graph. There are many possible fragment configurations that are consistent with most alignments and a directed graph that contains fragment boundary information requires the definition of additional arc types. An arc corresponding to the leftmost residue in a fragment will be referred to as a “beginning arc” and all other arcs will be “continuing arcs”. Beginning arcs may be normal southern, normal eastern, normal southeastern, or special southeastern. When a link from an ancestral sequence dies but leaves a descendant, the fragments associated with the ancestral link and the descendant link may have different lengths. For this reason, continuing arcs may be special eastern or special southern in addition to normal southern, normal eastern, normal

southeastern, and special southeastern. If the ancestral fragment that was deleted has a length of m and its descendant has a length of n then, for $m \geq n$, the relationship between these two fragments would be represented on the path graph by n special southeastern arcs followed by $m - n$ special southern arcs. For $m \leq n$, the relationship would be represented by m special southeastern arcs followed by $n - m$ special eastern arcs. Path graph representation of the other possible types of relationships between fragments is straightforward.

Regional heterogeneity of replacement rates Different regions of a typical protein are subject to different functional and structural constraints. To improve the realism of our model, we can allow different fragments to experience replacements at two (or more) distinct rates, fast and slow. The expected proportion of fragments that are fast will be denoted by p_f . If slow fragments evolve at rate s , then fast fragments are assumed to evolve at rate $s \cdot k_f$ where $k_f \geq 1$.

The fragment length distribution and the birth-death process are assumed to be independent of whether the replacement rate of a fragment is fast or slow. For example, p_f is the probability that a newborn fragment is fast regardless of whether the parental link is associated with a fast fragment or with a slow fragment. With these assumptions, the replacement process provides information about whether a region is fast or slow but the insertion-deletion process does not. With regional heterogeneity of replacement rates, normal southeastern arcs are further categorized as slow southeastern arcs or fast southeastern arcs. Because other arc types reflect only insertion or deletion events, they do not need to be categorized as fast or slow.

Terminal gaps It is often desirable for biological reasons to treat gaps at the ends of an alignment differently from gaps in the interior of an alignment. Terminal gaps can arise from factors that are less prone to produce interior gaps (e.g., data

collection, frame-shift mutations). We will assume that additional residues are independently appended to one of the two sequences at both the left end and the right end of a pairwise alignment and that the number of appended residues has a geometric distribution with parameter τ . For example, the probability that the residues *ILV* are appended to the right end of sequence *A* is

$$0.5(1 - \tau)\tau^3\pi_I\pi_L\pi_V \quad .$$

The factor of 0.5 is present because the subsequence is equally likely to be appended to either of the two sequences in the pairwise alignment. Arcs representing terminal gaps will be referred to as terminal southern arcs or terminal eastern arcs. Also, the terms “left terminal” and “right terminal” will be used to specify the end of an alignment to which the terminal gap is attached.

3 The Likelihood of a Sequence Pair

3.1 The Marginal Likelihood

Let $\theta = (\mu t, \lambda t, st)$ denote the model parameters. When the observed data consists of sequences *A* and *B*, the likelihood $L_\theta(A, B)$ can be expressed as

$$P_\theta(A, B) = \sum_{\alpha} P_\theta(\alpha, A, B)$$

where the summation runs through all possible alignments. For most sequence pairs, the total number of possible alignments is enormous (Waterman, 1984). However, it is not necessary to separately enumerate each alignment and the likelihood $L_\theta(A, B)$ can be computed in time proportional to the product of the sequence lengths.

Define the partial likelihood $L_\theta(i, j, k)$ to be $P_\theta(\alpha(i, j, k), A_j, B_k)$. For the simple evolutionary model, these partial likelihoods can be computed by a recursive

procedure outlined in Appendix 1. The full likelihood $P_\theta(A, B)$ is

$$L_\theta(A, B) = \sum_{i=0}^3 L_\theta(i, n_A, n_B) . \quad (5)$$

3.2 The augmented data likelihood

When the alignment α between sequences A and B is known, the likelihood is an exponential family distribution and can be factored as follows

$$P_\theta(A, B, \alpha) = P_\theta(B|A, \alpha)P_\theta(A|\alpha)P_\theta(\alpha). \quad (6)$$

The first term in this corresponds to the replacement process. The second term is a product of the equilibrium frequencies of the amino acids that specify sequence A . The third term corresponds to the insertion-deletion process and (if allowed) terminal gaps.

With a known alignment, sufficient statistics for the replacement process parameter st under the simple model are the counts, n_{ij} , of the numbers of occurrences of $a_m = i$ and $b_n = j$ among corresponding amino-acid residues in the given alignment. The equilibrium frequency parameters π_i do not need to be estimated as they are prespecified by the Dayhoff model. The insertion-deletion process parameters of the simple model have sufficient statistics given by x_i , the total number of type i arcs in the alignment α .

4 The EM Algorithm

The augmented data likelihood suggests an EM algorithm for estimation of evolutionary parameters. This algorithm assumes the true alignment is unknown. The arc probabilities, $P_\theta(\alpha(i, j, k)|A, B)$, are computed by a “reverse-pass” algorithm as outlined in Appendix 3. The arc probabilities are used in the E-step

of the EM algorithm to compute expectations of the statistics that are sufficient for parameter estimation when the alignment is known. The data is augmented by replacing the actual (unknown) value of these statistics by their expectations. In the M-step, the augmented data likelihood is maximized.

4.1 The E-step

Starting with an initial estimate $\theta^{(0)} = (\mu t^{(0)}, \lambda t^{(0)}, st^{(0)})$, the expectation step of the algorithm for the simple model proceeds as follows,

$$E(x_i|\theta^{(s)}) = \sum_{j,k} \Pr(\alpha(i, j, k)|A, B, \theta^{(s)}) \quad i = 0, 1, 2, 3.$$

Expectations of n_{hi} are

$$E(n_{hi}|\theta^{(s)}) = \sum_{h=a_j, i=b_k} \Pr(\alpha(1, j, k)|A, B, \theta^{(s)}) .$$

Note that:

$$E(x_1|\theta^{(s)}) = \sum_i E(n_{ii}|\theta^{(s)}) + \sum_{h \neq i} E(n_{hi}|\theta^{(s)}).$$

4.2 The M-step

In the M-step, we set

$$x_i^{(s)} = E(x_i|\theta^{(s)}) \quad i = 0, 1, 2, 3$$

and

$$n_{hi}^{(s)} = E(n_{hi}|\theta^{(s)}).$$

For the simple model, the value of $st^{(s+1)}$ can be found by maximizing

$$\sum_i n_{ii}^{(s)} \log(f_{ii}(t)) + \sum_{h \neq i} n_{hi}^{(s)} \log(f_{hi}(t))$$

The maximization is not difficult because of its one dimensional nature. Our implementation performs a golden section search. The values of $\mu t^{(s+1)}$ and $\lambda t^{(s+1)}$ can be jointly found by maximizing

$$\log\left(\left(1 - \frac{\lambda}{\mu}\right)p_1''(t)\right) + x_1^{(s)}\log\left(\frac{\lambda}{\mu}p_1(t)\right) + x_2^{(s)}\log\left(\frac{\lambda}{\mu}p_1'(t)\right) + (x_0^{(s)} + x_3^{(s)})\log(\lambda\beta(t)) .$$

Its two dimensional nature makes this maximization slightly more difficult. We find it convenient to reparameterize and estimate μt and λ/μ . Our implementation finds the joint estimate of μt and λ/μ by a naive maximization technique that seems to work well in practice. The technique repeats cycles consisting of maximizing μt while holding λ/μ constant and then maximizing λ/μ while holding the new value of μt constant. The cycles are repeated until the estimates converge.

4.3 Interpreting arc probabilities

Even for closely related sequences, the true pairwise alignment is unlikely to be the optimal or most probable alignment (Thorne et al. 1991). Each individual position within an alignment may be inferred with a relatively high degree of reliability but the joint probability that all alignment positions are correctly inferred can be quite low. Inability to determine the true alignment does not imply absence of evolutionary information. Different positions in an alignment will be inferred with different degrees of reliability. Those with the highest degree of reliability may be the most interesting from the standpoint of molecular biology because these positions are likely to be slowly evolving and hence functionally important. Detection of putative functionally important regions can aid the molecular dissection of a protein.

There are previously proposed dynamic programming methods for detection of the most reliable regions within an alignment (Waterman 1983; Vingron and Argos 1990; Zuker 1991). These are modifications of widely-used algorithms that recover

an alignment that is optimal according to a prespecified scoring system. A weakness of the widely-used algorithms is the lack of a clear relationship between evolutionary assumptions and the scoring system. Specifically, previous suggestions deem an arc reliable if and only if the best score among all alignments that include the arc is within a prespecified range of the score of the optimal alignment. The previous suggestions can be adapted to the maximum likelihood context. The advantage of a likelihood approach to this problem is that it can provide statistical meaning to the prespecified range. Let α_{\max} be an alignment between A and B that satisfies

$$P_{\theta}(\alpha_{\max}|A, B) = \max_{\alpha} P_{\theta}(\alpha|A, B) .$$

The prespecified range can be defined to include only arcs satisfying the condition that the most probable alignment containing the arc has probability greater than $\delta P_{\theta}(\alpha_{\max}|A, B)$ where $0 < \delta < 1$.

We choose another approach – an approach that is different in terms of biological assumptions and statistical methodology but similar in spirit to the approach of Allison, Wallace, and Yee (1992). Instead of considering only the single most likely alignment that contains the arc, we determine the reliability of an arc by summing the probabilities of all alignments that contain it. This is exactly what the reverse-pass algorithm accomplishes by calculating $P_{\theta}(\alpha(i, j, k)|A, B)$. Within a single optimal alignment, it is usually the case that some positions are more likely to be correctly inferred than others. This fact is ignored if the reliability of an arc is determined solely by the score of the best single alignment that contains it. Summing over all alignments that contain an arc can capture more evolutionary information. As discussed in the example, visual display of arc probabilities can assist in their interpretation.

The arc probabilities as obtained here are conditional on maximum likelihood estimates of the evolutionary parameters. Although we believe the approach

described here is superior to widely-used approaches, a poorly estimated parameter might result in misleading assessment of arc reliability. A solution to this potential shortcoming may be to consider alignment inference in a Bayesian framework (Churchill and Thorne, in prep).

5 Example

To illustrate the methods, the partial amino acid sequences of chicken hepatic lectin (Drickhamer 1981) and dog pulmonary surfactant (Benson et al. 1985) that were previously examined by Zuker (1991) are compared. The maximum likelihood parameter estimates shown in Table 1 are for the generalized model (both with and without special treatment of terminal events). The large standard errors in Table 1 are due to the relatively short length of the sequences that were compared. With the special treatment of terminal gaps, the average number of residues per insertion-deletion event is $1/(1 - \tau) \doteq 3.5$. The average length of a terminal gap is $\tau/(1 - \tau) \doteq 33$ residues. Clearly, the handling of terminal gaps can have a major effect on parameter estimates. Without the special treatment, estimates of both the number and length of insertion-deletion events are greater.

Arc probabilities for this example can be visualized in several ways. Arcs can be drawn on the grid with thickness proportional to their probability. This type of depiction is sometimes helpful but representation of arc probabilities by the thickness of arcs may not be very practical for most sequence pairs of biological interest due to the relatively long length of these sequences. An alternative is to depict only arc probabilities that exceed a prespecified threshold. In Figures 2A-2C, slow and fast southeastern arcs are considered separately and given different degrees of shading. This helps distinguish fast and slow fragments and may be relevant when attempting

to locate functionally important regions of the proteins. Unfortunately, slow arcs superimposed on fast arcs will obscure the fast arcs. The appropriate choice for the depiction of arc probabilities depends on the biological questions being investigated.

6 Discussion

The evolutionary process, the biological constraints that define it, and the evolutionary pattern that it has produced can be studied via examination of sequence data. A model-based approach to this study has the advantages of explicit assumptions and statistical interpretability. Obviously, insufficient knowledge makes current models of molecular evolution, including those used here, suspect. The value of model-based methods is not solely determined by whether their assumptions are met. An important consideration is robustness. Those assumptions that are found to be the weakest can be replaced. Our hope is that model-based approaches will enhance our understanding of molecular evolution and can thereby lead to more realistic models.

Alignment inference, the search for evolutionary correspondence among the amino acids or nucleotides of sequences, is among the fundamental problems of molecular evolution. Previously, much of the research in the area of molecular evolution has solely considered nucleotide substitutions or amino acid replacements. Lack of attention to insertions and deletions has retarded both the development of alignment inference techniques and our ability to measure the amount of evolution that separates a pair of sequences. Both of these shortcomings are addressed here. The parameter estimates obtained by the EM algorithm are a more a complete description of pairwise evolutionary relationships than are measures based solely on nucleotide substitution or amino acid replacement. The assessment of alignment

position reliability increases our ability to study evolutionary correspondence in a well-defined statistical framework.

A limitation of this work is the inability to consider more than two sequences simultaneously. To overcome this limitation, more powerful computers and a more complex statistical treatment will be required. Fortunately, progress on both of these fronts is expected. The techniques introduced here are amenable to parallelization. Even more importantly, the development of statistical resampling techniques (e.g., Hastings 1970) may make consideration of more than two sequences feasible.

REFERENCES

- Allison, L., Wallace, C.S., and Yee C.N. (1992). Finite-state models in the alignment of macromolecules. *Journal of Molecular Evolution* **35**, 77–89.
- Bishop, M.J. and Thompson, E.A. (1986). Maximum likelihood alignment of DNA sequences. *Journal of Molecular Biology* **190**, 159–165.
- Benson, B., Haywood, S., Schilling, J., Clements, J., Damm, D., Cordell, B., and White, R.T. (1985). Structure of canine pulmonary surfactant apoprotein: cDNA and complete amino acid sequence. *Proceedings of the National Academy of Sciences, U.S.A.* **82**, 6379–6383.
- Churchill, G. and Thorne, J.L. (in prep). The probability distribution of a molecular sequence alignment.
- Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978). A model of evolutionary change in proteins. In *Atlas of protein sequence structure, vol. 5, suppl. 3*, M.O. Dayhoff (ed), 345–352. Washington D.C.: National Biomedical Research Foundation.

- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- Drickhamer, K. (1981). Complete amino acid sequence of a membrane receptor for glycoproteins. *Journal of Biological Chemistry* **256**, 5827–5839.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.
- Gonnet, G.H., Cohen, M.A., and Benner, S.A. (1992). Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
- Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *CABIOS* **8**, 275–282.
- Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution* **31**, 151–160.
- Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**, 57–84.
- Thorne, J.L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* **33**, 114–124.
- Thorne, J.L., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: An improved likelihood model of sequence evolution. *Journal of Molecular Evolution* **34**, 3–16.
- Vingron, M. and Argos, P. (1990). Determination of reliable regions in protein sequence alignments. *Protein Engineering* **3**, 565–569.

- Waterman, M.S. (1983). Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proceedings of the National Academy of Sciences, U.S.A.* **80**, 3123–3124.
- Waterman, M.S. (1984). General methods of sequence comparison. *Bulletin of Mathematical Biology* **46**, 473–500.
- Zuker, M. (1991). Suboptimal sequence alignment in molecular biology: Alignment with error analysis. *Journal of Molecular Biology* **221**, 403–420.

Appendix 1: The Forward-Pass Algorithm

Define the partial likelihood $L_\theta(i, j, k)$ to be $P_\theta(\alpha(i, j, k), A_j, B_k)$. For $1 \leq j \leq n_A$ and $1 \leq k \leq n_B$, we carry out the following recursion (Thorne et al. 1991)

$$L_\theta(0, j, k) = \frac{\lambda}{\mu} \pi_{a_j} p'_0(t) \sum_{i=0}^3 L_\theta(i, j-1, k)$$

$$L_\theta(1, j, k) = \frac{\lambda}{\mu} \pi_{a_j} f_{a_j b_k}(t) p_1(t) \sum_{i=0}^3 L_\theta(i, j-1, k-1)$$

$$L_\theta(2, j, k) = \frac{\lambda}{\mu} \pi_{a_j} \pi_{b_k} p'_1(t) \sum_{i=0}^3 L_\theta(i, j-1, k-1)$$

$$L_\theta(3, j, k) = \pi_{b_k} \lambda \beta(t) \sum_{i=1}^3 L_\theta(i, j, k-1).$$

The boundary conditions for this “forward-pass” algorithm are

$$L_\theta(i, 0, 0) = 0 \quad i \neq 1$$

$$L_\theta(1, 0, 0) = \gamma_0 p''_1(t)$$

$$L_\theta(0, j, 0) = \gamma_j p''_1(t) \prod_{i=1}^j \pi_{a_i} p'_0(t) \quad 1 \leq j \leq n_A \quad (7)$$

$$L_\theta(i, j, 0) = 0 \quad 1 \leq j \leq n_A, \quad i \neq 0$$

$$L_\theta(i, 0, k) = 0 \quad 1 \leq k \leq n_B, \quad i \neq 3$$

$$L_\theta(3, 0, k) = \gamma_0 p''_{k+1}(t) \prod_{i=1}^k \pi_{b_i} \quad 1 \leq k \leq n_B.$$

The likelihood of the two sequences is

$$L_\theta(A, B) = \sum_{i=0}^3 L_\theta(i, n_A, n_B) . \quad (8)$$

Appendix 2: The Forward Pass Algorithm, General Case

As before, let $\alpha(i, j, k)$ be the event that an arc of type i enters node (j, k) where the extended definition of i is

$$i = \begin{cases} 0 & \text{if a normal southern arc enters node } (j, k) \\ 1 & \text{if a special southern arc enters node } (j, k) \\ 2 & \text{if a slow southeastern arc enters node } (j, k) \\ 3 & \text{if a fast southeastern arc enters node } (j, k) \\ 4 & \text{if a special southeastern arc enters node } (j, k) \\ 5 & \text{if a normal eastern arc enters node } (j, k) \\ 6 & \text{if a special eastern arc enters node } (j, k) \\ 7 & \text{if a left terminal southern arc enters node } (j, k) \\ 8 & \text{if a left terminal eastern arc enters node } (j, k) \\ 9 & \text{if a right terminal southern arc enters node } (j, k) \\ 10 & \text{if a right terminal eastern arc enters node } (j, k) \end{cases}$$

Define the partial likelihood $L_\theta(i, j, k)$ to be $P_\theta(\alpha(i, j, k), A_j, B_k)$. Also, let the superscripts s and f on the amino acid replacement probabilities $f_{ij}^s(t)$ and $f_{ij}^f(t)$ respectively denote whether replacements are occurring in a slowly or quickly evolving region. For $1 \leq j \leq n_A$ and $1 \leq k \leq n_B$, the forward-pass algorithm is

$$L_\theta(0, j, k) = \pi_{a_j}(rL_\theta(0, j-1, k) + \lambda\beta(t)(1-r) \sum_{i=0}^8 L_\theta(i, j-1, k))$$

$$L_\theta(1, j, k) = \pi_{a_j}r \sum_{i=1,4} L_\theta(i, j-1, k)$$

$$L_\theta(2, j, k) = \pi_{a_j} f_{a_j b_k}^s(t) (r L_\theta(2, j-1, k-1) + (1-p_f) p_1(t) (1-r) \frac{\lambda}{\mu} \sum_{i=0}^8 L_\theta(i, j-1, k-1))$$

$$L_\theta(3, j, k) = \pi_{a_j} f_{a_j b_k}^f(t) (r L_\theta(3, j-1, k-1) + p_f p_1(t) (1-r) \frac{\lambda}{\mu} \sum_{i=0}^8 L_\theta(i, j-1, k-1))$$

$$L_\theta(4, j, k) = \pi_{a_j} \pi_{b_k} (r^2 L_\theta(4, j-1, k-1) + p_1'(t) (1-r)^2 \frac{\lambda}{\mu} \sum_{i=0}^8 L_\theta(i, j-1, k-1))$$

$$L_\theta(5, j, k) = \pi_{b_k} (r L_\theta(5, j, k-1) + \lambda \beta(t) (1-r) \sum_{i=1}^8 L_\theta(i, j, k-1))$$

$$L_\theta(6, j, k) = \pi_{b_k} r \sum_{i=4,6} L_\theta(i, j, k-1)$$

$$L_\theta(7, j, k) = L_\theta(8, j, k) = 0 .$$

The recursions for the right terminal arcs are:

$$L_\theta(9, j, k) = 0 \quad k < n_B$$

$$L_\theta(9, 0, n_B) = 0$$

$$L_\theta(9, j, n_B) = \pi_{a_j} \tau (L_\theta(9, j-1, n_B) + 0.5 \sum_{i=0}^8 L_\theta(i, j-1, n_B)) \quad 1 \leq j \leq n_A$$

$$L_\theta(10, j, k) = 0 \quad j < n_A$$

$$L_\theta(10, n_A, 0) = 0$$

$$L_\theta(10, n_A, k) = \pi_{b_k} \tau (L_\theta(10, n_A, k-1) + 0.5 \sum_{i=0}^8 L_\theta(i, n_A, k-1)) \quad 1 \leq k \leq n_B.$$

The boundary conditions are

$$L_\theta(0, 0, 0) = 0$$

$$L_\theta(7, 0, 0) = (1 - \frac{\lambda}{\mu})(1 - \tau)^2 p_1''(t)$$

$$L_\theta(7, 1, 0) = \pi_{a_1} 0.5 \tau L_\theta(7, 0, 0)$$

$$L_\theta(0, j, 0) = \pi_{a_j} (r L_\theta(0, j-1, 0) + \lambda \beta(t) (1-r) \sum_{i=0,7} L_\theta(i, j-1, 0)) \quad j \geq 1$$

$$L_\theta(7, j, 0) = \pi_{a_j} \tau L_\theta(7, j-1, 0) \quad j \geq 2$$

$$L_\theta(i, j, 0) = 0 \quad i \neq 0, 7$$

$$L_\theta(5, 0, 1) = \pi_{b_1} \lambda \beta(t) (1-r) L_\theta(7, 0, 0)$$

$$L_\theta(8, 0, 1) = \pi_{b_1} 0.5 \tau L_\theta(7, 0, 0)$$

$$L_\theta(5, 0, k) = \pi_{b_k} (r L_\theta(5, 0, k-1) + \lambda \beta(t) (1-r) \sum_{i=5,8} L_\theta(i, 0, k-1)) \quad k \geq 2$$

$$L_\theta(8, 0, k) = \pi_{b_k} \tau L_\theta(8, 0, k-1) \quad k \geq 2$$

$$L_\theta(i, 0, k) = 0 \quad k \geq 1, \quad i \neq 5, 8.$$

Appendix 3: The Reverse-Pass Algorithm

The arc probabilities, $P_\theta(\alpha(i, j, k)|A, B)$ for node (n_A, n_B) are

$$P_\theta(\alpha(i, n_A, n_B)|A, B) = \frac{L_\theta(i, n_A, n_B)}{P_\theta(A, B)}.$$

The remaining arc probabilities are obtained recursively by proceeding in a northwestern direction from the node (n_A, n_B) to the node $(0, 0)$. Let $\alpha^r(i, j, k)$ denote an arc of type i that leaves node (j, k) . For the simple model, the two notations for arcs are related as follows:

$$\begin{aligned}\alpha(0, j, k) &= \alpha^r(0, j-1, k) \\ \alpha(1, j, k) &= \alpha^r(1, j-1, k-1) \\ \alpha(2, j, k) &= \alpha^r(2, j-1, k-1) \\ \alpha(3, j, k) &= \alpha^r(3, j, k-1) \quad .\end{aligned}$$

It follows that,

$$\begin{aligned}P_\theta(\alpha(i, j, k)|A, B) &= \sum_{m=0}^3 P_\theta(\alpha(i, j, k)|\alpha^r(m, j, k), A, B)P_\theta(\alpha^r(m, j, k)|A, B) \\ &= \sum_{m=0}^3 P_\theta(\alpha(i, j, k)|\alpha^r(m, j, k), A_{j^*}, B_{k^*})P_\theta(\alpha^r(m, j, k)|A, B) \\ &= \sum_{m=0}^3 \frac{P_\theta(\alpha(i, j, k), \alpha^r(m, j, k), A_{j^*}, B_{k^*})}{\sum_{n=0}^3 P_\theta(\alpha(n, j, k), \alpha^r(m, j, k), A_{j^*}, B_{k^*})} P_\theta(\alpha^r(m, j, k)|A, B)\end{aligned}$$

where $j < n_A, k < n_B$,

$$j^* = \begin{cases} j+1 & m = 0, 1, 2 \\ j & m = 3 \end{cases}$$

and

$$k^* = \begin{cases} k & m = 0 \\ k+1 & m = 1, 2, 3 \quad . \end{cases}$$

$$P_\theta(\alpha(i, j, k), \alpha^r(m, j, k), A_{j^*}, B_{k^*}) = \begin{cases} L_\theta(i, j, k) \frac{\lambda}{\mu} \pi_{a_{j^*}} p'_0(t) & m = 0 \\ L_\theta(i, j, k) \frac{\lambda}{\mu} \pi_{a_{j^*}} f_{a_{j^*} b_{k^*}}(t) p_1(t) & m = 1 \\ L_\theta(i, j, k) \frac{\lambda}{\mu} \pi_{a_{j^*}} \pi_{b_{k^*}} p'_1(t) & m = 2 \\ L_\theta(i, j, k) \lambda \beta(t) \pi_{b_{k^*}} & m = 3, i = 1, 2, 3 \\ 0 & m = 3, i = 0 \end{cases} .$$

Boundary conditions for the reverse-pass algorithm ensure that no arcs visit any node (j, k) with $j > n_A$ or $k > n_B$. They are straightforward and are not detailed here. Similarly, the general version of the reverse-pass algorithm is a straightforward generalization of this simple version and is not described here.

Table 1: Maximum Likelihood Parameter Estimates

	r	τ	μt	λ/μ	st	p_f	k_f
with	0.72	0.97	0.13	0.97	93	0.39	7.6
	(0.10)	(0.02)	(0.08)	(0.03)	(30)	(0.20)	(3.9)
without	0.88		0.29	0.95	125	0.40	3.0
	(0.04)		(0.16)	(0.05)	(42)	(0.32)	(2.3)

Parameter estimates, with and without special treatment of terminal gaps, from the comparison of partial amino acid sequences for chicken hepatic lectin and dog pulmonary surfactant. Approximate standard errors are in parentheses.

FIGURE CAPTIONS

Figure 1: Two representations of an alignment.

A. The “stacked” representation of the alignment. A “-” denotes a gap.

B. The directed graph representation of the alignment.

Figure 2: Comparison of the partial amino acid sequences for dog pulmonary surfactant and chicken hepatic lectin. Arc probabilities were calculated by setting the evolutionary parameter values at their maximum likelihood estimates. The dog pulmonary surfactant sequence is along the horizontal axis and the portion of the protein sequence analyzed is 121 amino acids in length. It begins with the amino acids *LHESL* and ends with the amino acids *LAICE*. The chicken hepatic lectin sequence is along the vertical axis and the portion of the protein sequence analyzed is 202 amino acids in length. It begins with the amino acids *MDEER* and ends with the amino acids *YYVCE*. For A-C, slow southeastern arcs are darkly shaded. All other arcs types are lightly shaded. For B-D, special treatment is given to terminal gaps.

A: All arcs with probability greater than 0.01 are depicted. No special treatment is given to terminal gaps.

B: All arcs with probability greater than 0.01 are depicted.

C: All arcs with probability greater than 0.1 are depicted.

D. All arcs types are darkly shaded. Probabilities of fast southeastern arcs and slow southeastern arcs are summed. All probabilities greater than 0.1 are depicted.

Figure 1A

G - C - A C A
- T G T - C -

Figure 2B

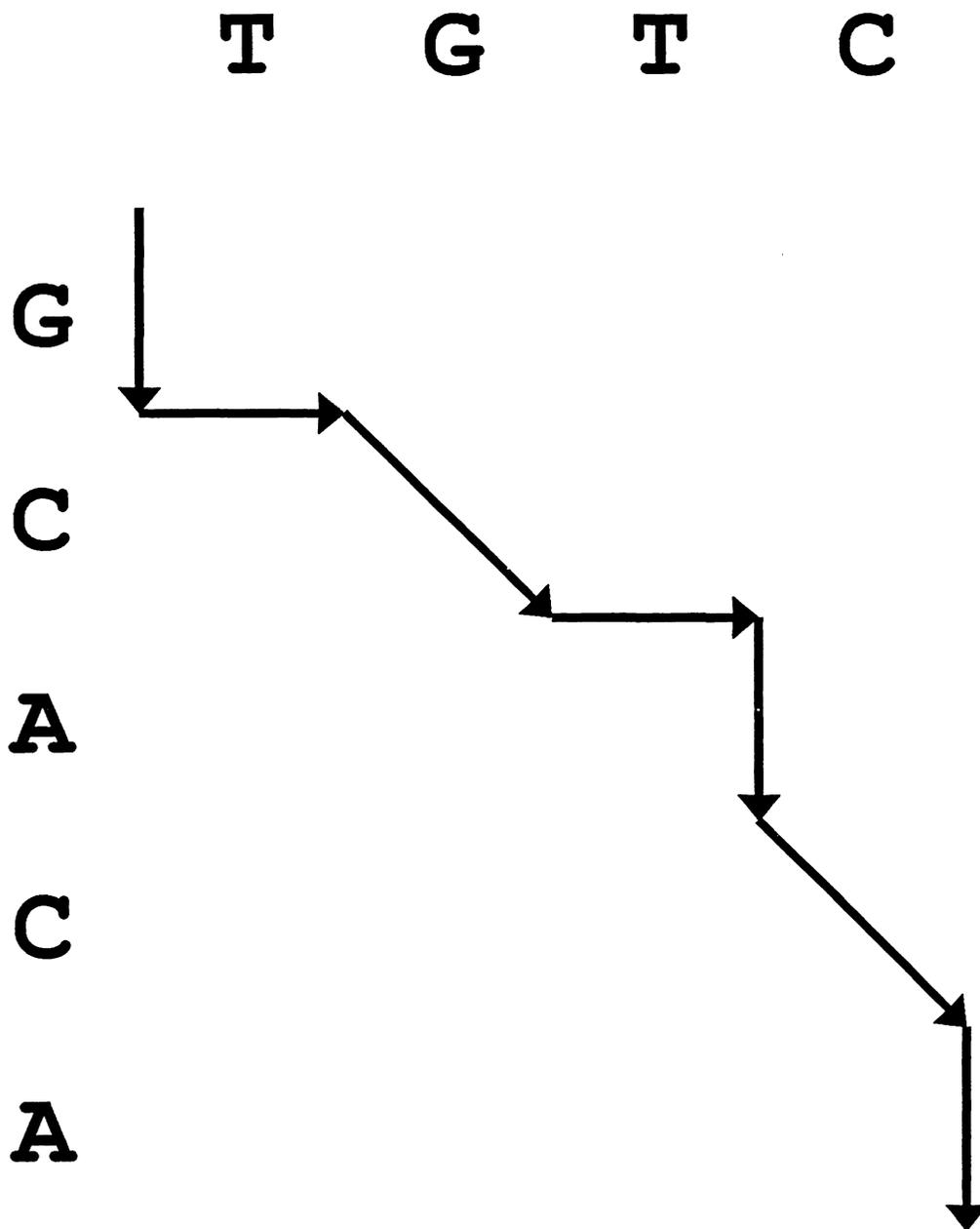


Figure 2A

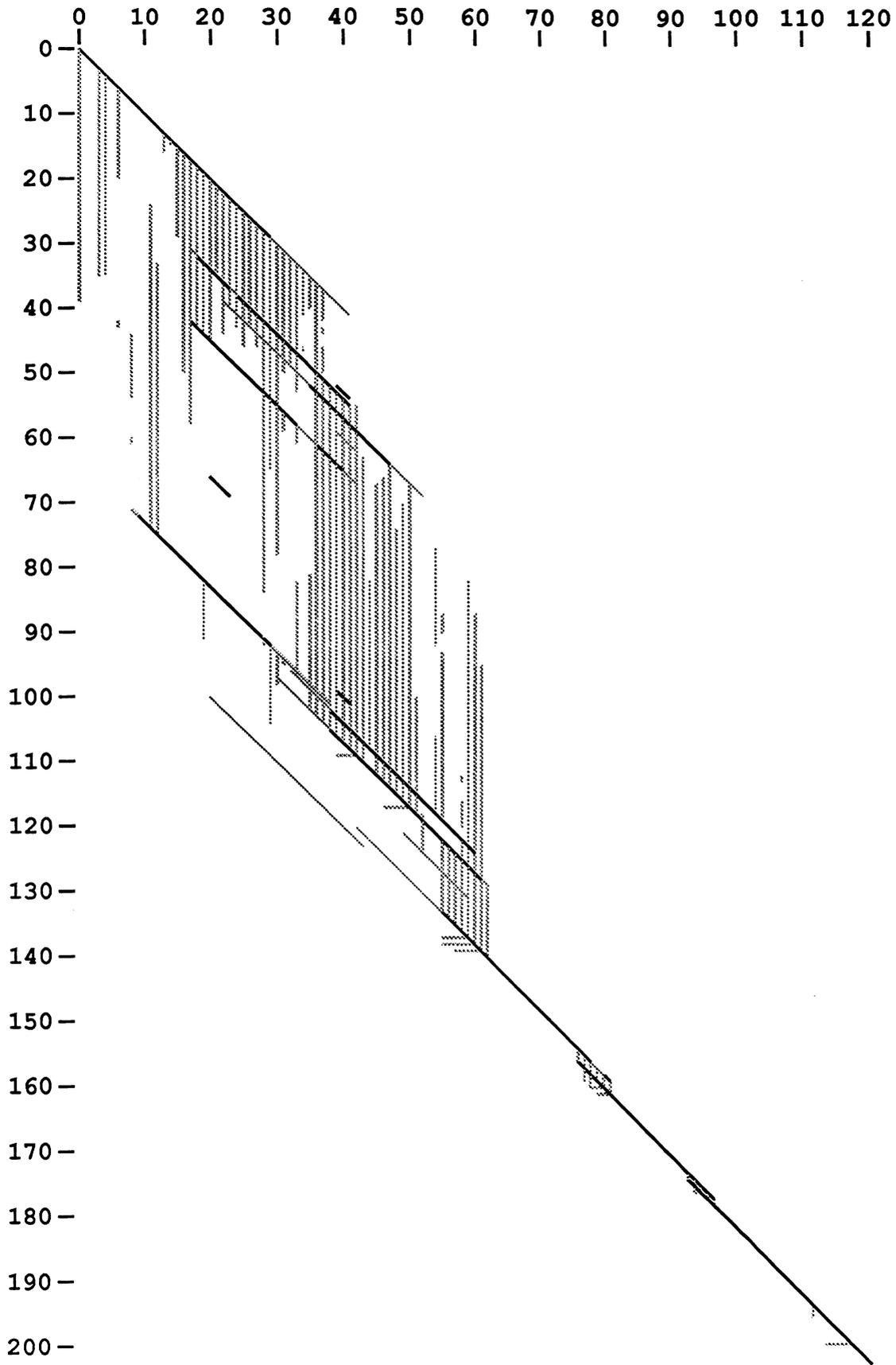


Figure 2C

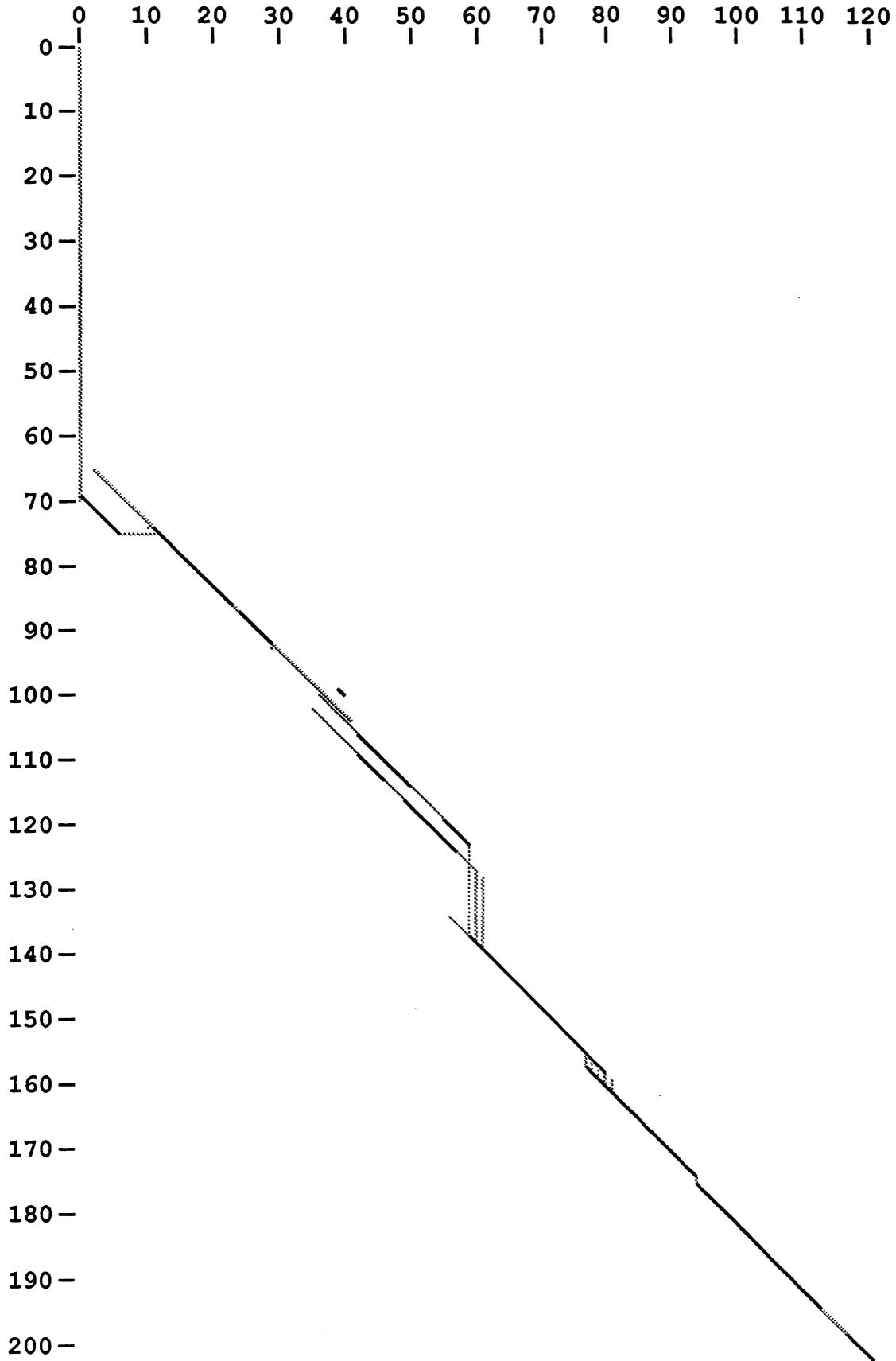


Figure 2D

