LA-UR- 01 - 3 2 7 0

Approved for public release;
distribution is unlimited.

Title: AN IMPROVED MODEL FOR STATISTICAL ALIGNMENT

Author(s): I. Miklos, Eotvos University
Z. Toroczkai, CNLS/T-Division

Submitted to: Proceedings: 1st Workshop on Algorithmus in
BioInformatics, BRICS, University of ASarhus, Denmark,
August 28-31, 2001. The proceedings of the workshop will
be published in the Lecture Notes in Computer Science

Los Alamos

NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

An Improved Model for Statistical Alignment

István Miklós¹ and Zoltán Toroczka²

¹ Department of Plant Taxonomy and Ecology, Eötvös University, Ludovika tér 2, H-1083
Budapest, Hungary

i_miklos@hotmail.com

² Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory,
Los Alamos, NM87545, USA
toro@clns.lanl.gov

Abstract. The statistical approach to molecular sequence evolution involves the stochastic modeling of the substitution, insertion and deletion processes. Substitution has been modeled in a reliable way for more than three decades by using finite Markov-processes. Insertion and deletion, however, seem to be more difficult to model, and the recent approaches cannot acceptably deal with multiple insertions and deletions. A new method based on a generating function approach is introduced to describe the multiple insertion process. The presented algorithm computes the approximate joint probability of two sequences in $O(l^3)$ running time where l is the geometric mean of the sequence lengths.

1 Introduction

The traditional sequence analysis [1] needs proper evolutionary parameters. These parameters depend on the actual divergence time, which is usually unknown as well. Another major problem is that the evolutionary parameters cannot be estimated from a single alignment. Incorrectly determined parameters might cause unrecognizable bias in the sequence alignment.

One way to break this vicious circle is the maximum likelihood parameter estimation. In the pioneering work of Bishop and Thompson, [2] an approximate likelihood calculation was introduced. Several years later, Thorne, Kishino and Felsenstein wrote a landmark paper [3], in which they presented an exact maximum likelihood algorithm, which estimates the evolutionary distance between two sequences involving all possible alignments in the likelihood calculation. Their 1991 model (frequently referred to as the TKF91 model) considers only single insertions and deletions, but this consideration is rather unrealistic [4],[5]. Later it was improved by allowing longer insertions and deletions [4] in the model, which is usually coined as the TKF92 model. However, the improved model assumes that sequences contain unbreakable fragments, and only whole fragments are inserted and deleted. As it was shown [4], the fragment model has a flaw: considering unbreakable fragments, there is no possible explanation for overlapping deletions with a scenario of just two events. This problem is solvable by assuming that the ancestral sequence was fragmented independently on both branches immediately after the split, and sequences evolved

since then according to the fragment model [6]. However, this assumption does not solve the problem completely: fragments do not have biological realism. The lack of the biological realism is revealed when we want to generalize this split model for multiple sequence comparison. For example, consider that we have proteins from humans, gorillas and chimps. When we want to analyze the three sequences simultaneously, two pairs of fragmentation are needed: one pair at the gorilla-(human and chimp) split and one at the human-chimp split. When only sequences from gorillas and humans are compared, the fragmentation at the human-chimp split is omitted. Thus, the description of the evolution of two sequences depends on the number of the introduced splits, and there is no sensible interpretation to this depending.

1.1 The Thorne-Kishino-Felsenstein Model

Since our model is related to the TKF91 model we describe it briefly. Most of the definitions and notations are introduced in here.

The TKF model is the fusion of two independent time-continuous Markov processes, the substitution and the insertion-deletion process.

The Substitution Process Each character can be substituted independently for another character dictated by one of the well-known substitution processes [7],[8]. The substitution process is described by a system of linear differential equations

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{Q}\mathbf{x}(t) \quad (1)$$

where $\mathbf{Q}$ is the rate matrix. Since $\mathbf{Q}$ contains too many parameters, it is usually separated into two components, $\mathbf{Q}_0$ s, where $\mathbf{Q}_0$ is kept constant and is estimated with a less rigorous method than maximum likelihood [4]. The solution of (1) is

$$\mathbf{x}(t) = e^{\mathbf{Q}_0 t} \mathbf{x}(0) \quad (2)$$

The Insertion-Deletion Process The insertion-deletion process is traditionally described not in terms of amino acids or nucleotides but in terms of imaginary links. A mortal link is associated to the right of each character, and additionally, there is an immortal link at the left end of the sequence. Each link can give birth to a mortal link with birth rate λ . The newborn link always appears at the right side of its parent. Accompanying the birth of a mortal link is the birth of a character drawn from the equilibrium distribution. Only mortal links can die out with death rate μ , taking their character to the left with them. Assuming independence between links, it is sufficient to describe the fate of single mortal link and the immortal one. According to the possible histories of links, three types of functions are considered. Let $p_k^{(1)}(t)$ denote the probability that after time t , a mortal link has survived, and has exactly k descendants including itself. Let $p_k^{(2)}(t)$ denote the probability that after time t , a

mortal link died, but it left exactly k descendants. Let $p_k(t)$ denote the probability that after time t , the immortal link has exactly k descendants, including itself.

Calculating the Joint Probability of Two Sequences The joint probability of two sequences A and B is calculated as the equilibrium probability of sequence A times the probability that sequence B evolved from A under time $2t$, where t is the divergence time.

$$P(A, B) = P_{\infty}(A)P_{2t}(B|A) \quad (3)$$

A possible transition is described as an alignment. The upper sequence is the ancestor; the lower sequence is the descendant. For example the following alignment describes that the immortal link (o) has one descendant, the first mortal link (*) died out, and the second mortal link has two descendants including itself.

o - A* U* -

o G* - C* A*

The probability of an alignment is the probability of the ancestor, times the probability of the transition. For example, the probability of the above alignment is

$$\gamma_2 \pi(A) \pi(U) p_2(t) \pi(G) p_0^{(2)}(t) p_2^{(1)}(t) f_{UC}(2t) \pi(A) \quad (4)$$

where γ_n is the probability that a sequence contains n mortal links, $\pi(X)$ is the frequency of the character X, and $f_{ij}(2t)$ is the probability that a character i is of j at time $2t$.

The joint probability of two sequences is the summation of the alignment probabilities.

2 The model

Our model differs from the TKF model in the insertion-deletion process. The difference is that every character has an own mortal link in the long insertions. Thus, this model allows long insertions without considering unbreakable fragments. It is possible that a long fragment is inserted into the sequence first and some of the inserted links die and some of them survive after then. A link gives birth to a block of k mortal links with rate λ_k , where

$$\lambda_k = \lambda r(1-r)^{k-1}, \quad k = 1, 2, \dots, \quad 0 < r < 1, \quad \lambda > 0 \quad (5)$$

Only mortal links can die with rate μ , $\mu > 0$.

2.1 Calculating the Generating Functions

The Master Equation First, the probabilities of the possible fates of the immortal link is computed. Collecting the gain and loss terms for this birth-death process, a Master equation is obtained:

$$\frac{dp_n}{dt} = \sum_{j=1}^{n-1} (n-j)\lambda_j p_{n-j} + n\mu p_{n+1} - \left[n \sum_{j=1}^{\infty} \lambda_j + (n-1)\mu \right] p_n \quad (6)$$

Using $\sum_{j=1}^{\infty} \lambda_j = \lambda$, and $\sum_{j=1}^{n-1} (n-j)\lambda_j p_{n-j} = \sum_{k=1}^{n-1} k\lambda_{n-k} p_k$:

$$\frac{dp_n}{dt} = \lambda r \sum_{k=1}^{n-1} k(1-r)^{n-k-1} p_k + n\mu p_{n+1} - [n\lambda + (n-1)\mu] p_n \quad (7)$$

Due to the immortal link, $p_0(t)=0, \forall t$. For $n=1$, the sum in (7) is void. The initial conditions are given by:

$$p_n(0) = \delta_{n,1} \quad (8)$$

Next, we introduce the generating function [9]:

$$P(\xi; t) = \sum_{n=0}^{\infty} \xi^n p_n(t) \quad (9)$$

Multiplying (7) with ξ^n , then summing over n , we obtain a linear PDE for the generating function:

$$\frac{\partial P}{\partial t} - (1-\xi) \left[\mu - \frac{\lambda \xi}{1-\xi(1-r)} \right] \frac{\partial P}{\partial \xi} = -(1-\xi) \frac{\mu}{\xi} P \quad (10)$$

with initial condition $P(\xi; 0) = \xi$.

Solution to the PDE for the Generating Function We use the method of Lagrange:

$$\frac{dt}{1} = \frac{d\xi}{-(1-\xi) \left[\mu - \frac{\lambda \xi}{1-\xi(1-r)} \right]} = \frac{dP}{-(1-\xi) \frac{\mu}{\xi} P} \quad (11)$$

The two equalities define two, one-parameter family of surfaces, namely $v(t; \xi; P)$ and $w(t; \xi; P)$. After integrating the first and the second equalities in (11) the following surfaces are obtained:

$$v(\xi; t) = \frac{(1 - \xi)^r}{(\mu - a\xi)^{\lambda/a}} e^{-t(\mu - \lambda)} = c_1 \quad (12)$$

$$w(\xi; t; P) = P \frac{(\mu - a\xi)^{\lambda/a}}{\xi} = c_2 \quad (13)$$

where $a \equiv \lambda + \mu(1-r) > 0$. The general form of the solution is an arbitrary function of $w=g(v)$. This means:

$$P(\xi; t) = \xi(\mu - a\xi)^{-\lambda/a} g\left(\frac{(1 - \xi)^r}{(\mu - a\xi)^{\lambda/a}} e^{-t(\mu - \lambda)}\right) \quad (14)$$

The function g is fixed from the initial condition $P(\xi, 0) = \xi$:

$$g(z) = [\mu - af^{-1}(z)]^{\lambda/a} \quad (15)$$

where

$$f(x) = \frac{(1 - x)^r}{(\mu - ax)^{\lambda/a}} \quad (16)$$

Thus the exact form for the generating function becomes:

$$P(\xi; t) = \xi \left[\frac{\mu - af^{-1}(f(\xi)e^{-(\mu - \lambda)t})}{\mu - a\xi} \right]^{\frac{\lambda}{a}} \quad (17)$$

The Probabilities for the Fate of the Mortal Links The Master Equations for the probabilities $p_n^{(1)}(t)$ and $p_n^{(2)}(t)$ are given by

$$\frac{dp_n^{(1)}}{dt} = \sum_{j=1}^{n-1} (n-j)\lambda_j p_{n-j}^{(1)} + n\mu p_{n+1}^{(1)} - \left[n \sum_{j=1}^{\infty} \lambda_j + n\mu \right] p_n^{(1)} \quad (18)$$

$$\begin{aligned} \frac{dp_n^{(2)}}{dt} = & \sum_{j=1}^{n-1} (n-j)\lambda_j p_{n-j}^{(2)} + (n+1)\mu p_{n+1}^{(2)} + \\ & + \mu p_{n+1}^{(1)} - \left[n \sum_{j=1}^{\infty} \lambda_j + n\mu \right] p_n^{(2)} \end{aligned} \quad (19)$$

We have the following conditions to be fulfilled:

$$p_0^{(1)}(t) = 0, \quad \forall t \geq 0 \quad (20)$$

and the initial conditions:

$$p_n^{(1)}(0) = \delta_{n,1}, \quad p_n^{(2)}(0) = 0, \quad \forall n \geq 0 \quad (21)$$

The corresponding partial differential equations for the generating functions $P^{(i)}(\xi, t) = \sum_{n=0}^{\infty} \xi^n p_n^{(i)}(t)$, $i=1,2$, are given by:

$$\frac{\partial P^{(1)}}{\partial t} - (1-\xi) \left[\mu - \frac{\lambda \xi}{1-\xi(1-r)} \right] \frac{\partial P^{(1)}}{\partial \xi} = -\frac{\mu}{\xi} P^{(1)} \quad (22)$$

$$\frac{\partial P^{(2)}}{\partial t} - (1-\xi) \left[\mu - \frac{\lambda \xi}{1-\xi(1-r)} \right] \frac{\partial P^{(2)}}{\partial \xi} = \frac{\mu}{\xi} P^{(1)} \quad (23)$$

Solution to the PDE-s for the Generating Functions of the Mortal Links First, we solve Eq. (22) using the method of Lagrange

$$\frac{dt}{1} = \frac{d\xi}{-(1-\xi) \left[\mu - \frac{\lambda \xi}{1-\xi(1-r)} \right]} = \frac{dP^{(1)}}{-\frac{\mu}{\xi} P^{(1)}} \quad (24)$$

The two, one-parameter family of surfaces are $v(t; \xi; P^{(1)})$ and $w(t; \xi; P^{(1)})$. Since v comes from the integration of the first equality in (22) it is the same as (12). Integrating the second equality yields:

$$w(\xi; t; P^{(1)}) = P^{(1)} \frac{(1-\xi)^{\mu r / (\mu r - \lambda)}}{\xi(\mu - a\xi)^{\lambda / (\mu r - \lambda)}} = c_2 \quad (25)$$

Proceeding as in the previous section, we have:

$$P^{(1)}(\xi; t) = \xi \left[\frac{\mu - a\xi}{\mu - af^{-1}(f(\xi)e^{-(\mu r - \lambda)t})} \right]^{\frac{\lambda}{\mu r - \lambda}} \times \left[\frac{1 - f^{-1}(f(\xi)e^{-(\mu r - \lambda)t})}{1 - \xi} \right]^{\frac{\mu r}{\mu r - \lambda}} \quad (26)$$

with f given by (16). To calculate $P^{(2)}(\xi; t)$ we first define $Q(\xi; t) = P^{(1)}(\xi; t) + P^{(2)}(\xi; t)$. Summing Eqs. (22) and (23) the following equation is obtained for Q :

$$\frac{\partial Q}{\partial t} - (1 - \xi) \left[\mu - \frac{\lambda \xi}{1 - \xi(1 - r)} \right] \frac{\partial Q}{\partial \xi} = 0 \quad (27)$$

This is easily solved with the method of characteristics. First, we integrate the characteristic equation, which is the first equation in (24), to obtain the family of characteristic curves, given by $v(\xi; t) = c_1$ as in (12). Thus, $Q(\xi; t) = g(v)$ is the general solution, where $g(x)$ is an arbitrary, differentiable function, to be set by the initial conditions. Using (20) and (21) we have $Q(\xi; 0) = \xi$. This leads to:

$$Q(\xi; t) = f^{-1}(f(\xi)e^{-(\mu - \lambda)t}) \quad (28)$$

with f given by (16), and therefore:

$$P^{(2)}(\xi; t) = f^{-1}(f(\xi)e^{-(\mu - \lambda)t}) - P^{(1)}(\xi; t) \quad (29)$$

with $P^{(1)}(\xi; t)$ given by Eq. (26).

2.2 The equilibrium length distribution

The generating function of the equilibrium length distribution can be obtained from (17) by considering the limit $t \rightarrow \infty$. Since $f^{-1}(0) = 1$ and due to the immortal link, the generating function becomes

$$\Gamma(\xi) = \left[\frac{\mu - a}{\mu - a\xi} \right]^{\frac{\lambda}{a}} \quad (30)$$

Calculating the Taylor-series of $\Gamma(\xi)$ around 0, we get for the equilibrium probabilities:

$$\gamma_n = (\mu - a)^{\lambda/a} \frac{\prod_{i=0}^{n-1} (\lambda + ia)}{n! \mu^{(\lambda + (n-1)a)/a}} \quad (31)$$

From $\frac{d\Gamma(\xi)}{d\xi}$ in the limit of $\xi \rightarrow 1^-$, the expected value of the sequence length is obtained as:

$$E(\gamma) = \frac{\lambda}{\mu r - \lambda} \quad (32)$$

3 The Algorithm

3.1 Calculating the transition probabilities

Unfortunately, the inverse of f given by (16) does not have a closed form. Thus a numerical approach is needed for calculating the transition probability functions $p_n(t)$, $p_n^{(1)}(t)$ and $p_n^{(2)}(t)$. We calculate the generating functions $P(\xi; t)$, $P^{(1)}(\xi; t)$ and $P^{(2)}(\xi; t)$ in l_1+1 points around $\xi=0$, where l_1 is the length of the shorter sequence. For doing this, the following equation must be solved for x numerically where ξ , μ , λ , r , t and a are given.

$$f(\xi)e^{-(\mu r - \lambda)t} = \frac{(1-x)^r}{(\mu - ax)^{\lambda/a}} \quad (33)$$

Given l_1+1 points, the functions are partially derived l_1 times. After this

$$p_n(t) = \frac{\partial^n P(\xi, t)}{\partial \xi^n} \frac{1}{n!} \quad (34)$$

and similarly for $p_n^{(1)}(t)$ and for $p_n^{(2)}(t)$. Thus, the transition probability functions can be calculated in $O(l_1^2)$ time.

3.2 Dynamic Programming for the Joint Probability

Without loss of generality we can suppose that the sorter sequence is sequence B. The equilibrium probability of sequence A is

$$P_\infty(A) = (\mu - a)^{\lambda/a} \frac{\prod_{i=0}^{l(A)-1} (\lambda + ia)}{l(A)! \mu^{(\lambda + (l(A)-1)a)/a}} \prod_{i=0}^{l(A)-1} \pi(a_i) \quad (35)$$

where a_i is the i th character in A and $l(A)$ is the length of the sequence.

Let A_i denote the i -long prefix of A , and let B_j denote the j -long prefix of B . There is a dynamic programming algorithm for calculating the transition probabilities $P_t(A_i|B_j)$. The initial conditions are given by:

$$P_t(A_0|B_j) = p_{n+1}(t) \prod_{k=1}^j \pi(b_k) \quad (36)$$

To save computation time, we calculate $\prod_{k=1}^j \pi(b_k)$ for every $k < j$ before the recursion. Then the recursion follows

$$\begin{aligned}
P_i(A_i|B_j) &= \sum_{l=0}^j P_i(A_{i-1}|B_l) p_{j-l}^{(2)}(t) \prod_{k=l+1}^j \pi(b_k) + \\
&+ \sum_{l=0}^{j-1} P_i(A_{i-1}|B_l) p_{j-l}^{(1)}(t) f_{a,b_{j+1}} \prod_{k=l+2}^j \pi(b_k)
\end{aligned} \tag{37}$$

The dynamic programming is the most time-consuming part of the algorithm, it takes $O(l^3)$ running time.

3.3 Finding the Maximum Likelihood Parameters

As mentioned earlier, the substitution process is described with only one parameter, st . (A general phenomena is that the time and rate parameters can not be estimated individually, only their product.) The insertion-deletion model is described with three parameters, λt , μt and r , which however, can be reduced to two, if the following equation is taken under consideration

$$\frac{\lambda}{\mu r - \lambda} = \frac{l(A) + l(B)}{2} \tag{38}$$

namely, the mean of the sequence lengths is the maximum likelihood estimator for the expected value of the length distribution.

The maximum likelihood values of the three remaining parameters can be obtained using one of the well-known numerical methods (gradient method, etc.).

4 Discussion and Conclusions

There is an increasing desire for statistical methods of sequence analysis in the bioinformatics community. The weak point of the statistical approach is the lack of an appropriate evolutionary model. A new model and an associated algorithm for computing the joint probability were introduced. This new model is superior to the Thorne-Kishino-Felsenstein model: it allows long insertions without considering unbreakable fragments. However, it is only a small inch to the reality, as it contains at least two unrealistic properties. It cannot deal with long deletions, and the rates for the long insertions form a geometric series. The elimination of both these problems seems to be rather difficult but not impossible. Other rate functions for long insertions lead to more difficult PDE-s whose characteristic equations may not be integrated without a rather involved computational overhead. The same situation appears when long deletions are allowed. Moreover, in this case calculating only the fates of the individual links is not sufficient. Thus, for achieving more appropriate models, numerical calculations are needed in an earlier state of the procedure. We hope that the generating function approach will open some novel avenues for further research.

5 Acknowledgments

Charsten Wiuf is thanked for useful discussions and suggestions. Z.T. was supported by the DOE, under contract W-7405-ENG-36.

References

1. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* 48. (1970) 443-453
2. Bishop, M. J., Thompson, E.A.: Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* 190. (1986) 159-165
3. Thorne, J.L., Kishino, H., Felsenstein, J.: An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* 33. (1991) 114-124
4. Thorne, J.L., Kishino, H., Felsenstein, J.: Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* 34. (1992) 3-16
5. Hein, J., Wiuf, C., Knudsen, B., Moller, M.B., Wiblig, G.: Statistical alignment: computational properties, homology testing and goodness-of-fit. *J. Mol. Biol.* 302. (2000) 265-279
6. Miklos, I.: Irreversible likelihood models, European Mathematical Genetics Meeting, 20-21. April, 2001, Lille, France.
7. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: A model for evolutionary change in proteins, matrices for detecting distant relationships. In: Dayhoff, M.O. (ed.): *Atlas of Protein Sequence and Structure*, Vol. 5. Cambridge University Press, Washington DC. (1978) 343-352
8. Tavaré, S.: Some probabilistic and statistical problems in the analysis of DNA sequences. *Lec. Math. Life Sci.* 17. (1986) 57-86
9. Feller, W.: *An introduction to the probability theory and its applications*, Vol. 1. McGraw-Hill, New York (1968) 264-269