The Statistics of Semi-Probabilistic Alignment

Yi-Kuo Yu[1] **Terence Hwa**[2]
yyu@fau.edu hwa@ucsd.edu

[1] Department of Physics, Florida Atlantic University, Boca Raton, FL 33431-0991, USA
[2] Department of Physics, U.C. San Diego, La Jolla, CA 92093-0319, USA

Keywords: sequence alignment, statistical significance, maximum likelihood, hidden Markov model

Introduction

Computer-assisted sequence comparison has become an integral part of modern molecular biology. Two types of algorithms have been used: those which search for the *optimal* alignment (as exemplified by the Smith-Waterman algorithm [1]), and those which identify *likely* alignments (as exemplified by the HMM-based "Sequence Alignment Modules" [2]). In each case, the quality of alignment is summarized by an alignment score S; the latter is typically taken to be the logarithm of the total likelihood in the probabilistic approaches. An important goal common to the study of all algorithms is to understand the score distribution $P(\mathsf{S})$ for the appropriate null models. This distribution gives the probability that a high scores could have arisen by chance and is therefore more meaningful to homology detection than the alignment scores themselves.

Rigorous results on such background statistics are known only for the gapless alignment [3], whose score distribution follows the so-called Gumbel form, $P(\mathsf{S}) = KMN\lambda e^{-\lambda \mathsf{S} - KMN \exp(-\lambda \mathsf{S})}$, for long sequence lengths M and N. There exist explicit formulae relating the hundreds of alignment parameters to the two Gumbel parameters λ and K. For the gapped Smith-Waterman alignment, ample empirical evidences suggest that the null score distribution still obeys the Gumbel form. But the dependences of the two Gumbel parameters on the alignment score functions are very complicated and largely unknown. For probabilistic alignments, the log-likelihood score does not even satisfy Gumbel distribution as was recently shown [4].

Combining the optimal and the probabilistic approaches to sequence alignment, we developed a new "hybrid" algorithm for which the alignment score is still Gumbel-distributed, and the Gumbel parameters can be computed accurately and rapidly, for a large class of scoring functions (including the position-specific ones), and for a wide range of sequence lengths (down to $\sim$ 100 amino acids), without the need of extensive simulation as is commonly done. We have independently checked using sequences generated by simple evolution models that the *fidelity* of homology detected by the hybrid algorithm is comparable to or better than that of the Smith-Waterman algorithm [4].

Algorithm and Results

For clarity, we describe here only the simplest algorithm with linear gap cost. All the numerical results are however obtained for the more commonly used affine gap function which is described in detail in Ref. [4]. Given two sequences $\vec{a} = \{a_1, a_2, \cdots, a_M\}$ and $\vec{b} = \{b_1, b_2, \cdots, b_N\}$ to be aligned, we first compute the likelihood $Z_{m,n}$ of aligning the subsequences $\hat{a}_{m',m} = \{a_{m'+1}, a_{m'+2}, \cdots, a_m\}$ and $\hat{b}_{n',n} = \{b_{n'+1}, b_{n'+2}, \cdots, b_n\}$ for all possible "starting point" $1 \le m' < m$ and $1 \le n' < n$. $Z_{m,n}$ can be efficiently computed using dynamical programming:

$$Z_{m,n} = w(a_m, b_n) \cdot Z_{m-1,n-1} + \nu \cdot [Z_{m-1,n} + Z_{m,n-1}] + 1, \tag{1}$$

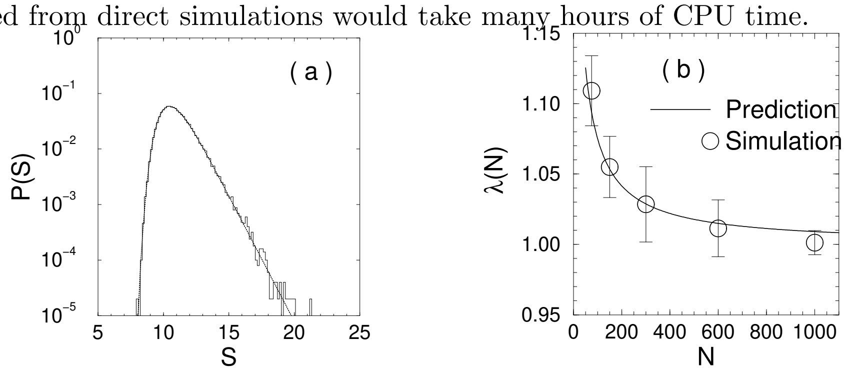
for $1 < m < M$ and $1 < n < N$, with $w(a_i, b_j)$ being the prescribed probability of pairing (a_i, b_j), ν the prescribed probability of indels, and the "boundary condition" $Z_{0,n\ge 0} = Z_{m\ge 0,0} = 1$. Note that without the +1 term on the right hand side of Eq. (1), $Z_{m,n}$ is just the total likelihood of the *global* alignment of the subsequences $\hat{a}_{1,m}$ and $\hat{b}_{1,n}$, provided the probability conservation condition

$$\sum_{a,b} w(a,b)p(a)p(b) + 2\nu = 1, \tag{2}$$

where $p(a)$ is the background frequency of the amino acid a. The condition (2) is easily satisfied, for example, by choosing $w(a,b) = (1-2\nu)\cdot e^{s(a,b)}$, for the class of (rescaled) PAM substitution scores $s(a,b)$. The additional term of +1 in Eq. (1) turns the algorithm into probabilistic *local* alignment [4]. In the fully probabilistic approach (as in SAM [2]), the alignment score is taken to be the log of the total likelihood, e.g., $\ln\left[\sum_{m,n} Z_{m,n}\right]$. We instead take

$$\mathsf{S} \equiv \max_{m,n}\{\, \ln Z_{m,n}\,\} \tag{3}$$

as the alignment score. Eqs. (1) and (3) define the semi-probabilistic or hybrid alignment algorithm we used in this study. This is an $O(M\cdot N)$ algorithm as in Smith-Waterman or probabilistic alignments. It however offers a tremendous advantage in that the probability distribution $P(\mathsf{S})$ can be computed without large scale numerical simulation, for a large class of scoring functions (including position-specific ones) as long as they obey the probability conservation condition such as (2). Our theory predicts that $P(\mathsf{S})$ is Gumbel distributed, with the important Gumbel parameter λ being exactly 1 for long sequences. A simple recipe for correction to λ due to finite sequence lengths is also given and can be efficiently computed using a dozen or so alignments of appropriately chosen correlated sequences. In the following figures, we illustrate results obtained from the hybrid algorithm, using the PAM-120 substitution matrix, and an affine gap function. Figure (a) shows the distribution $P(\mathsf{S})$ obtained from 50,000 pairwise alignment of random sequences of lengths 300. It agrees very well with the expected Gumbel distribution (smooth line). Figure (b) shows the dependence of the parameter λ on the sequence length N. The circles are data from direct simulation and the line is the theoretical prediction, which agrees with the data to within one to a few percents. With our approach, such accurate values of Gumbel parameters can be obtained within a few seconds, while comparable statistics obtained from direct simulations would take many hours of CPU time.



References

[1] Smith, T.F. and Waterman, M.S., Identification of common molecular subsequences, *J. Mol. Biol.* 147:195–197, 1981.

[2] Hughey, R. and Krogh, A., Hidden Markov models for sequence analysis: extension and analysis of the basic method, *CABIOS*, 12:95–107, 1996.

[3] Karlin, S. and Altschul, S.F., Applications and statistics for multiple high-scoring segments in molecular sequences, *Proc. Natl. Acad. Sci. USA*, 90:5873–5877, 1993.

[4] Yu, Y.-K. and Hwa, T., Statistical significance of probabilistic sequence alignment and related local hidden markov models. (submit to J. Comp. Biol.), 1999.