

Improved probability calculation for DIALIGN2

Changhoon Kim and Byungkook Lee¹

Keywords: DIALIGN2, Probability, Higher order Markov assumption, Alignment weight

1 Introduction.

DIALIGN2 is a successful sequence alignment program, which does not use a gap penalty[1,2,3,4]. It considers all possible gapless aligned segments, called diagonals, calculates a score for each, and assigns a weight to each according to the probability that the corresponding diagonal selected from two random sequences will have the same calculated score. The final alignment is chosen to be the consistent combination of diagonals which will maximize the total weight. This is an attractive algorithm since it finds good alignments without using a gap penalty, which is always arbitrary and difficult to determine. However, its success depends on the ability to accurately calculate the probability with which a diagonal between two unrelated sequences will attain a certain score. For an isolated diagonal, this is not difficult to compute. However, the algorithm calls for comparing multiple, overlapping diagonals, for which the probabilities are not independent of each other and rigorously accurate probability calculations are not always possible. In the current version of DIALIGN2, the probabilities are derived from random experiments when they are greater than 10^{-5} and by a mathematical approximation when they fall below this level [3,4]. Here we report a procedure for calculating the probabilities more accurately and which eliminates the need to compute them by random experiments.

2 Results and Discussion.

Our procedure depends on expressing the probabilities as a Bayesian chain of conditional probabilities, which we approximate as an r -th order Markov chain. The value of r was chosen to be 6 as a compromise between the need for accuracy and the limitations of computing resources.

A sequence comparison matrix (SCM) is the matrix formed by the two sequences being compared [5]. A “super-diagonal” of the matrix is a diagonal line that starts from the right or upper edge of the matrix and ends at the bottom or the right edge of the matrix. A “diagonal” is a line segment within a super-diagonal. We define $P^*(S; l, L)$ and $P^*(S; l_1, l_2)$ as the probability of finding a diagonal of length l with a score $\geq S$ somewhere within a super-diagonal of length L and within the whole SCM, respectively, of two random sequences of lengths l_1 and l_2 . The final results can then be summed up by the following two formulas:

$$P^*(S; l_1, l_2) = 1 - \left[\prod_{L=l}^{l_1-1} \{1 - P^*(S; l, L)\} \right]^2 [1 - P^*(S; l, l_1)]^{l_2 - l_1 + 1} \quad (1)$$

and

$$P^*(S; l, L) \approx 1 - \left[1 - P^*(S; l, l+r) \right] \left[\frac{1 - P^*(S; l, l+r)}{1 - P^*(S; l, l+r-1)} \right]^{L-l-r} \quad (2)$$

¹ Laboratory of Molecular Biology, Center for Cancer Research, National Cancer Institute, National Institute of Health, Bethesda, Maryland 20892-4264, USA. BK@nih.gov

The newly calculated probabilities are compared with those calculated by DIALIGN2 in Fig. 1 for two different SCM sizes (100x100 for the lower three curves and 512x625 for the upper three curves). In order to see if the use of the new probabilities improves alignment quality, alignment sensitivity and specificity were measured using the *SABmark1.63 super-family set* [6]. Each structurally aligned sequence pair was realigned by DIALIGN2 (d2) and by the modified program (d3) that uses the new probabilities. The sensitivity and specificity of the alignments were averaged over group of sequence pairs with $l_1 <$ abscissa value (Fig. 2). It is clear that the sensitivity is improved substantially by using d3 over d2 without losing specificity, although the difference decreases as the sequence length increases.

3 Figures.

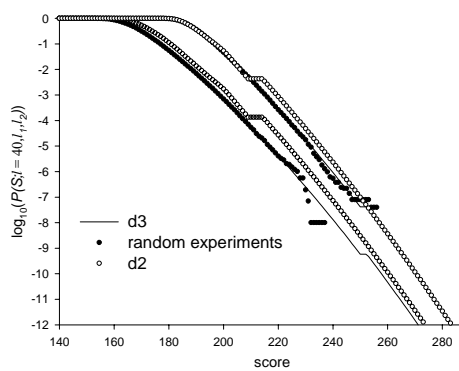


Figure 1. Comparison of probabilities.

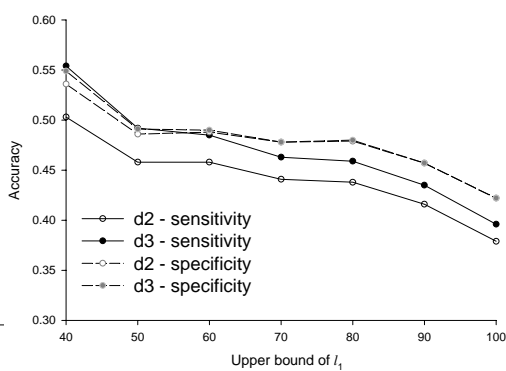


Figure 2. Alignment accuracy.

References

- [5] Argos, P. and Vingron, M. 1990. Sensitivity comparison of protein amino acid sequences. *Methods Enzymol.* 183: 352-65.
- [4] Morgenstern, B. 1999. DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* 15:211-218.
- [3] Morgenstern, B., Atchley, W.R., Hahn, K. and Dress, A. *et al.* 1998. Segment-based scores for pairwise and multiple sequence alignments. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6:115-121.
- [1] Morgenstern, B., Dress, A. and Werner, T. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. U. S. A.* 93:12098-12103.
- [2] Morgenstern, B., Frech, K., Dress, A. and Werner, T. 1998. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics.* 14:290-294.
- [6] Van Walle, I., Lasters, I. and Wyns, L. 2004. Align-m--a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics.* 20:1428-1435.