A Simple Derivation of the Distribution of Pairwise Local Protein Sequence Alignment Scores

Olivier Bastien

CNRS (UMR 5168) - INRA (UMR 1200) - CEA - Université Joseph Fourier, Laboratoire de Physiologie Cellulaire Végétale; CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France.

Abstract: Confidence in pairwise alignments of biological sequences, obtained by various methods such as Blast or Smith-Waterman, is critical for automatic analyses of genomic data. In the asymptotic limit of long sequences, the Karlin-Altschul model computes a *P-value* assuming that the number of high scoring matching regions above a threshold is Poisson distributed. Using a simple approach combined with recent results in reliability theory, we demonstrate here that the Karlin-Altshul model can be derived with no reference to the extreme events theory.
Sequences were considered as systems in which components are amino acids and having a high redundancy of Information reflected by their alignment scores. Evolution of the information shared between aligned components determined the Shared Amount of Information (SA.I.) between sequences, i.e. the score. The Gumbel distribution parameters of aligned sequences scores find here some theoretical rationale. The first is the Hazard Rate of the distribution of scores between residues and the second is the probability that two aligned residues do not lose bits of information (i.e. conserve an initial pairing score) when a mutation occurs.

Keyword: conservation function, reliability theory, Karlin-Altshul theorem

Introduction

Almost all sequence alignments methods compute a score s(*a,b*) between two compared sequences *a* and *b*. This score is a measure of similarity between the two sequences and help to distinguish biologically significant relationship from chance similarities (Smith and Waterman, 1981; Altschul et al. 1990; Waterman, 1995). These methods assign scores to insertions, deletions and replacements, and compute an alignment of two sequences that corresponds to the least costly set of such mutations. Assignment of a similarity measure begins with a matrix of similarity scores for all possible pairs of residues. Identities and conservative substitution have positive scores, while unlikely replacements have negative scores (Dayhoff et al. 1978; Henikoff and Henikoff, 1992; Bastien et al. 2005a). The score of the computed alignment is the sum of the elementary scores for each pair of aligned residues. All these methods allow the introduction of gaps in the alignment to maximize the final score and to taking account of deletion events in DNA (Waterman, 1995).

Because of the exponential increase of the number of sequence in each database and the large number of sequenced genomes, confidence in alignment score probabilities is critical to perform a rapid and accurate discrimination between alignments. The two main probability models compare the score s(*a,b*) with a score computed using random sequence A and B.

The first method proposed by Karlin and Altschul (1990) is an estimate of the probability of an observed local ungapped alignment score according to an Extreme Value Distribution, (or EVD; for review: Coles, 2001) in the asymptotic limit of long sequences. The Karlin-Altschul formula is the consequence of interpreting the number of highest scoring matching regions above a threshold by a Poisson distribution (Karlin and Altschul, 1990). As a consequence, if *s* is the score obtained after aligning two real sequences *a* and *b* (with *m* and *n* their respective lengths), the probability of finding an ungapped segment pair with a score lower than or equal to *s*, follows a particular Gumbel distribution (named EVD type I):

$$P(S(A,B) \leq s) \approx \exp(-K.m.n.\exp(-\lambda.s)) \quad (1)$$

Correspondence: Olivier Bastien, Email: olivier.bastien@cea.fr