BMC Bioinformatics

BioMed Central

Research article

Open Access

Evolution of biological sequences implies an extreme value distribution of type I for both global and local pairwise alignment scores

Olivier Bastien* and Eric Maréchal

Address: UMR 5168 CNRS-CEA-INRA-Université J. Fourier, Laboratoire de Physiologie Cellulaire Végétale; Département Réponse et Dynamique Cellulaire; CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

Email: Olivier Bastien* - olivier.bastien@cea.fr; Eric Maréchal - eric.marechal@cea.fr

* Corresponding author

Published: 7 August 2008

Received: 2 May 2007

Accepted: 7 August 2008

BMC Bioinformatics 2008, **9**:332 doi:10.1186/1471-2105-9-332

This article is available from: http://www.biomedcentral.com/1471-2105/9/332

Abstract

Background: 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. Two statistical models have been proposed. In the asymptotic limit of long sequences, the Karlin-Altschul model is based on the computation of a *P-value*, assuming that the number of high scoring matching regions above a threshold is Poisson distributed. Alternatively, the Lipman-Pearson model is based on the computation of a *Z-value* from a random score distribution obtained by a Monte-Carlo simulation. *Z-values* allow the deduction of an upper bound of the *P-value* (1/*Z-value*^{2}) following the TULIP theorem. Simulations of *Z-value* distribution is known to fit with a Gumbel law. This remarkable property was not demonstrated and had no obvious biological support.

Results: We built a model of evolution of sequences based on aging, as meant in Reliability Theory, using the fact that the amount of information shared between an initial sequence and the sequences in its lineage (*i.e.*, mutual information in Information Theory) is a decreasing function of time. This quantity is simply measured by a sequence alignment score. In systems aging, the failure rate is related to the systems longevity. The system can be a machine with structured components, or a living entity or population. "Reliability" refers to the ability to operate properly according to a standard. Here, the "reliability" of a sequence refers to the ability to conserve a sufficient functional level at the folded and maturated protein level (positive selection pressure). Homologous sequences were considered as systems 1) having a high redundancy of information reflected by the magnitude of their alignment scores, 2) which components are the amino acids that can independently be damaged by random DNA mutations. From these assumptions, we deduced that information shared at each amino acid position evolved with a constant rate, corresponding to the information hazard rate, and that pairwise sequence alignment scores should follow a Gumbel distribution, which parameters could find some theoretical rationale. In particular, one parameter corresponds to the information hazard rate.

Conclusion: Extreme value distribution of alignment scores, assessed from high scoring segments pairs following the Karlin-Altschul model, can also be deduced from the Reliability Theory applied to molecular sequences. It reflects the redundancy of information between homologous sequences, under functional conservative pressure. This model also provides a link between concepts of biological sequence analysis and of systems biology.