Distribution at Contingency of Alignment of Two Literal Sequences under Constrains

Lorentz Jantschi[1], Sorana D. Bolboaca[2]
[1] Department of Physics and Chemistry, Technical University of Cluj-Napoca, 101-103 Muncii Bvd., 400641 Cluj-Napoca, Cluj, Romania.
lorentz.jantschi@gmail.com
[2] Department of Medical Informatics and Biostatistics, "Iuliu Hatieganu" University of Medicine and Pharmacy Cluj-Napoca, 400349 Cluj-Napoca, Cluj, Romania.
sbolboaca@umfcluj.ro

Keywords: literals alignment, contingency matrix, probability distribution function (PDF), cumulative distribution function (CDF)

Sequence alignments, defined as a way of arrange DNA (deoxyribonucleic acid), RNA, (ribonucleic acid) or protein (amino-acid) sequences to identify similar regions that could reflect functional, structural or evolutionary relationships between sequences [1], is frequently used nowadays due to huge amount of already identified sequence of DNA, RNA, or proteins [2]. Several algorithms were developed and implement for global or local alignments, and each having advantages and disadvantages [3] and [4].

Our research started from the hypothesis that the distribution of alignments could provide useful information about the chance that a certain alignment occur or not by chance. We present here a statistical approach based on distribution analysis that is able to identify the thresholds for rejecting an alignment by chance under the supposition that each literal has at least one alignment in any case. For two literal sequences, we define the alignment through the frequency of matches (with 0 meaning no alignment and n meaning perfect alignment, where n is the number of nucleotides or amino-acids in the two equal length sequences). A closed form of the probability distribution function of the alignment was obtained. We provided that the cumulative distribution function have (unfortunately) no general closed form. Anyway, a series of statistics (including mode and central moments till order 4) were obtained with closed forms. By using the formula for the cumulative probability of an alignment, for the particular case of four literals alignment, thresholds to reject the alignment by chance were obtained as follow: 70% for $n > 8$; 60% for $n > 13$; 55% for $n > 21$; 50% for $n > 39$; 45% for $n > 282$; 44% for $n \to \infty$.

References

[1] D.M. Mount, Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2004.

[2] K.D. Pruitt, T. Tatusova, G.R. Brown, and D.R. Maglott, *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy*, Nucleic Acids Res **40** (Database issue), 2012, D130–D135.

[3] F. Plewniak, *Database similarity searches*, Methods in Molecular Biology **484**, 2008, 361–378.

[4] X. Qu, R. Swanson, R. Day, and J. Tsai, *A guide to template based structure prediction*, Current Protein and Peptide Science **10**(3), 2009, 270–285.