An algorithm for exact probability of pattern occurrences calculation

Evgenia Furletova[1], Mireille Regnier[2], Mikhail Roytberg[1], Viktor Yacovlev[1]

[1]*Institute of Mathematical Problems of Biology, 4, Institutskaja str., 142290, Pushchino, Moscow Region, Russia*
mroytberg@impb.psn.ru
[2]*INRIA, 78153 Le Chesnay, France*

An important aspect of studying functional fragments of biological sequences is to determine the statistical significance of their occurrences. One of the statistical significance measures is P-value i.e. probability to find at least p occurrences of a pattern (a set of words) H in a random sequence of length n. We have created an algorithm SufPref to calculate the P-value. We assume that words from the pattern have the same length (pattern length).. The algorithm calculates P-value for three types of probability models: Bernoulli, Markov models of order K, Hidden Markov models. The program that implements the algorithm is available at http://server2.lpm.org.ru/bio

Unlike the majority of existing programs that calculate exact P-value only for Markov models of order 1 or two, our program supports Markov models of any order less then length of words in the pattern. As to our knowledge at the moment there are no program computing exact P-value for Hidden Markov models. In the Bernoulli case both of the time and space complexities of SufPref are independent of the alphabet size and the time complexity in average is independent of the pattern length. We have compared our algorithm with algorithm Spatt based on minimal finite automaton [1]. In most of test examples the running time of SufPref is better than one of SPatt, especially in the cases where patterns are randomly generated. The space complexities of both algorithms are compatible. A detail analysis of complexity of the algorithm for Bernoully and Markov cases and its comparison with other algorithms are given in the paper [2].

1. G. Nuel. Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and to pattern statistics. Algorithms for Molecular Biology 2006, 1:5 doi:10.1186/1748-7188-1-5
2. M. Regnier, Z. Kirakosyan, E. Furletova, M. Roytberg. An word counting graph. London Algorithmics 2008: Theory and Practice, 2009.