

Exact distribution of local score using Finite Markov Chain Imbedding: an effective approach.

— short abstract —

GRÉGORIE NUEL *

Local score statistics are used in a wide range of problems related to biological sequences (homology, hydrophobic segments, genetic markers, G-C content, ...). On i.i.d. sequences, it has been shown that its distribution is well approximated by a Gumbel one and more recently this result has been extended to Markov dependent sequences. Beside these asymptotic approximations, combinatorial methods have been proposed to compute the exact distribution of local score (both in i.i.d. and Markov cases) using an elegant Finite Markov Chain Imbedding (FMCI) technique. Unfortunately, the practical computations in the exact case are impracticable except in toy-example cases. In this paper, we propose a new approach to deal with FMCI computations which dramatically outperform the previous ones. This new method allows for the very first time to compute exact p-values for a real scale biological study (finding hydrophobic segments in the complete SwissProt database) for which only Gumbel approximations were available before. On this problem, the approximations surprisingly appear to be reliable only for 0.5 % of the considered sequences which is very low. Exact computations require obviously more time than Gumbel approximations, but can still be very fast using our new algorithm (more than 20 p-values computed per second in our example). As a conclusion, we strongly advise to anyone dealing with local score distribution to use these new exact computations rather than approximations whenever it is possible to.

*University of Evry, CNRS (8071), INRA (1142), Laboratoire Statistique et Génome, 523, place des terrasses de l'Agora, 91000 Evry, France (nuel@genopole.cnrs.fr)