Ann Inst Stat Math (2017) 69:231–248
DOI 10.1007/s10463-015-0540-y

Faster exact distributions of pattern statistics through sequential elimination of states

Donald E. K. Martin[1] · **Laurent Noé**[2]

Received: 18 April 2014 / Revised: 1 July 2015 / Published online: 18 September 2015

Abstract When using an auxiliary Markov chain (AMC) to compute sampling distributions, the computational complexity is directly related to the number of Markov chain states. For certain complex pattern statistics, minimal deterministic finite automata (DFA) have been used to facilitate efficient computation by reducing the number of AMC states. For example, when statistics of overlapping pattern occurrences are counted differently than non-overlapping occurrences, a DFA consisting of prefixes of patterns extended to overlapping occurrences has been generated and then minimized to form an AMC. However, there are situations where forming such a DFA is computationally expensive, e.g., with computing the distribution of spaced seed coverage. In dealing with this problem, we develop a method to obtain a small set of states during the state generation process without forming a DFA, and show that a huge reduction in the size of the AMC can be attained.

Keywords Active proper suffix · Auxiliary Markov chain · Computational efficiency · Extended seed patterns · Minimal deterministic finite automaton · Overlapping pattern occurrences · Seeded alignments · Spaced seed coverage

✉ Donald E. K. Martin
donald_martin@ncsu.edu

Laurent Noé
laurent.noe@univ-lille1.fr

[1] Department of Statistics, North Carolina State University, 5116 SAS Hall, Raleigh, NC 27695, USA

[2] CRIStAL (UMR 9189 Lille University/CNRS), INRIA Lille Nord-Europe, Villeneuve d'Ascq, France