Discriminating between spurious and significant matches

Hugo Devillers, Meriem El Karoui, Sophie Schbath

Discriminating between spurious and significant matches

Hugo Devillers1*, Meriem El Karoui2, Sophie Schbath1
1INRA, UR1077, Math&#65533;matique, Informatique & G&#65533;nome. Domaine de Vilvert, F-78352, Jouy-en-Josas, FRANCE.
2INRA, UR888, Unit&#65533; Bact&#65533;rie Lactiques et Pathog&#65533;nes Opportunistes. Domaine de Vilvert, F-78352, Jouy-en-Josas, FRANCE.
* hugo.devillers@jouy.inra.fr

Word matches are widely used to compare DNA sequences, especially when the compared sequences are too long to be aligned with classical methods. Thus, for example, complete genome alignment methods often rely on the use of matches for building the alignments and various alignment-free approaches that characterize similarities between large sequences are based on word matches.

Among the matches that are retrieved between two genomic sequences, a part of them may correspond to spurious matches (SMs), which are matches obtained by chance rather than by homologous relationship. The number of SMs depends on the minimal match length (l) that has to be set in the algorithm. Indeed, if l is too small, a lot of matches are recovered but most of them are SMs. Conversely, if l is too large, fewer matches are retrieved but many smaller significant matches are probably ignored. Last, it is obvious that the subsequent analysis of the obtained matches is significantly impaired if the number of SMs is high.

To date, the choice of l mostly depends on empirical threshold values rather than robust statistical methods. To overcome this problem, we propose a statistical approach based on the use of a mixture model of geometric laws to characterize the length distribution of matches obtained from the comparison of two genomic sequences. In this work, the basic principles of our approach are presented. Its strengths and weaknesses are then discussed through examples drawn from bacterial genome comparisons.

Acknowledgments
This work was supported by the French &#65533;Agence Nationale de la Recherche&#65533; project CoCoGen (BLAN07-1_185484). We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for their help and computational resources. We thank Dr P Nicolas for valuable comments on this work.