A pipeline for automatic taxonomic community inventories on data from NGS : an example with freshwater diatoms

Lenaïg L. Kermarrec, Philippe P. Chaumeil, Frédéric F. Rimet, Jean-Marc Frigerio, Valerie Laval, Marc-Henri M.-H. Lebrun, Agnes Bouchez, Alain Franc

Fourth International Barcode of Life Conference
Adelaide, Australia

Kermarrec L. (1), Chaumeil P. (2), Rimet F. (1), Frigerio J.M. (2), Laval V. (3), Lebrun M.H. (3), Bouchez A. (1), FRANC A. (2)

1: INRA CARRTEL, Thonon, France
2: INRA BIOGECO, Cestas, France
3: INRA BIOGER, Thiverval-Grignon, France

Abstract Title:
A PIPELINE FOR AUTOMATIC TAXONOMIC COMMUNITY INVENTORIES ON DATA FROM NGS : AN EXAMPLE WITH FRESHWATER DIATOMS

Abstract Text:
Most of the work devoted to molecular taxonomy in the context of barcoding has been devoted to get the name of a query from a match of a barcode or a sequence on a high quality database. Those high quality databases (quality of sequences and quality of taxonomic assignation) have been obtained on Sanger. Currently, NGS (Next Generation Sequencing) are exponentially developing, providing large data sets of sequences. Barcode may take an interest in making taxonomic inventories from these data on environmental sequences. As the quality of data produced by NGS is still lower than for those produced by Sanger, the strategy we propose is (i) keep expanding high quality databases with taxonomic expertise and Sanger sequencing and (ii) develop tools for automatically matching reads from NGS on these databases for retrieving taxonomic inventories of communities. Two questions will be distinguished. First, to exhibit perfect matches between reads and a word on a reference sequence. We present a pipeline which implements this task without any heuristic, and retrieving with high consistency a known artificially built community. The accuracy is better than BLAST due to the absence of heuristic, and to the fact that any step is implemented in a rigorous way. Second, most of reads get imperfect matches, as they may differ from a reference sequence by a couple of snp's or length differences in homopolymers. We will present a program with still ongoing research which implements these imperfect matches, with a time longer than perfect matches, but which can take into account within species genetic variability between the reference and the queries, and some technical discrepancies to sequencing technology. This tool, called metaMatch, allows an automatic taxonomic identification just after the output of NGS. Work is still going on to improve the speed and accuracy of this program.

Abstract considered for:

- Data Analysis Working Group and the Data Portal