Algorithm for error detection in metagonomics NGS data

Dimitar Vassilev[1], Milko Krachunov[2], Ivan Popov[1], Elena Todorovska[1], Valeria Simeonova[2], Pawel Szczesny[3,4], Pawel Siedlecki[3,4], Urszula Zelenkiewicz[3], Piotr Zelenkiewicz[3]

[1]Bioinformatics group, AgroBioInstitute, Sofia, Bulgaria

[2]Faculty of Mathematics and Informatics, Sofia University "St.Kliment Ohridski", Bulgaria

[3]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland

[4]Institute of Experimental Plant Biology and Biotechnology, University of Warsaw, Poland

http://www.abi.bg/

http://www.ibb.waw.pl/

Because of the nature of metagenomics data, it is neither possible to resample the data to account for the sequencing errors that inevitably occur, nor it is possible to clearly differentiate between an error and a biological variation [6]. Small errors in the sampled data often lead to significant changes in the results of any further analyses and studies based on the data, for example during the construction of phylogenetic trees or during the evaluation of the biological diversity in the sampled environment [2]. For improving the quality of such studies, it is essential that an approach for detecting probable errors is devised.

There are numerous published methods for error detection and correction in NGS data, however none of them are designed to work with metagenomics data, but instead focus on applications such as de novo sequencing of genomes where the appearance of biological variations that are undistinguishable from the errors is not an issue [1,2,3,4]. An example of such software is SHREC (used a s a point of reference in this study), which corrects errors in short-read data using a generalized suffix tree [5].

The input data for the initial tests consists of tens of thousands of 16S RNA short-reads with lengths between 300 and 500 bases. For the proposed method to be applied, the read sets need to be filtered of obvious noise and then aligned to each other.

The basic idea behind error correction is that if a given a bit of data, such as a single base, appears too rare in the dataset it is more likely for it to be an error than a biological variation (SNP). A threshold defining "too rare" can be established using the error rate of the sequencing equipment. Higher weights assigned to reads that are locally more similar to the read in question can improve the error recognition by excluding irrelevant data from species that have diverged. . The outline of our evaluation algorithm is as follows:

1. we go over the reads evaluating each base individually;
2. for each base in question, we create a window containing the base at its centre;
3. we calculate a similarity score between the read in question and every other read in the dataset within that particular window. The score excludes the evaluated base, while the bases closest to it are assigned the highest weights;
4. we calculate an evaluation score for the base by calculating a frequency weighted with the similarity score. The result is the ratio of the sum of the similarity scores for the reads that contain the base and the sum of the similarity scores for all the reads;
5. we compare the score of the base to a threshold that has been calculated in advance and experimentally verified. Any scores below the threshold are considered errors and the bases are replaced with the base candidate that would score most using the outlined algorithm.

The biggest challenge in the implementation of this approach is the pre-processing of the data, i.e. the sequence alignment. It is both a difficult and resource intensive task. Trading alignment accuracy for speed is not desirable as alignment errors affect both the evaluation and any further studies.

References

1. Chaisson MJ, Pevzner PA. (2007) Short read fragment assembly of bacterial genomes. Genome Research 18:324-330.
2. Flicek P., Brudno M. (2009) Sense from sequence reads: methods for alignment and assembly. Nature Methods Supplement 6(11) S6-S11.