Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2021:11**
http://www.math.su.se

A Bayesian Approach to Clustering
Correcting Errors in DNA barcode reads

Nik Tavakolian*

June 2021

Abstract

DNA barcodes are short DNA sequences introduced into a population to track the relative frequencies of lineages over time. These barcode sequences are unknown to the human observer upon insertion and must be identified using next-generation sequencing technology. This process is error prone and results in a large number of error sequences. To estimate the relative frequencies of the barcodes accurately these errors must be corrected for. This error correction task can be posed as a clustering problem where the goal is to group similar sequences together. Existing methods for this task have used the observed frequency of the sequences but have disregarded the per nucleotide error rate in the clustering process. Without an accurate estimate of this error rate the distribution of error sequences cannot be inferred, limiting the error correction accuracy of these methods. Furthermore, these methods have delegated the task of parameter selection to the user, leaving room for user errors resulting from unsuitable parameter choices. In this work we set out to develop a clustering procedure that addresses these shortcoming. We estimate the per nucleotide error rate and devise a Bayesian hypothesis test for distinguishing between true barcodes and error sequences. The proposed method considers all nearby sequences before clustering a given sequence and achieves higher accuracy than the current state-of-the-art method on simulated datasets.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: nik.tavakolian@gmail.com. Supervisor: Chun-Biu Li.