An automated real-time solution for genetic disorders detection and classification

Gjorgji Madjarov[a], Lukasz Krych[b], David Galevski[c], Anne Kristine Schack[b,d], Aleksandar Nikov[c], and Chris Kyriakidis[d]

[a] University Ss Cyril & Methodius, Skopje 1000, N. Macedonia
[b] University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg, Denmark
[c] Netcetera, Zypressenstrasse 71, 8040 Zürich, Switzerland
[d] gMendel, Fruebjergvej 3, 2100 Copenhagen, Denmark

Genetic disorders affect 350 million families worldwide. Accurate diagnosis and treatment remain a long and painful process that takes many years, dozens of medical examinations and still 40% of patients are misdiagnosed creating a heavy human, economic and societal burden.

Here, we propose a solution for automated, real-time genetic disorders detection and classification from DNA sequences obtained by Oxford Nanopore Technologies GridION x5. The proposed solution enables accurate, systematic and timely diagnosis that can significantly improve the treatment process. In this research, our focus is detection and classification of six different genetic disorders (Klinefelter, Turner, Down, Edwards, Patau and Prader-Willi/Angelman syndromes). The process starts with the sequence data quality check obtained by the fastq files. Only the sequences that meet the quality criteria (Phred quality score >10 and sequence length: 900 - 1200) are used. The process continues with demultiplexing of barcodes in the sequence data (the samples are pooled with proprietary DNA barcoding kit) and chromosome classification. After that, utilizing data driven sequence modeling, we build multiple base and conceptual models for the specific genetic disorders. Following a multi model fusion strategy, our approach combines the outcomes of the individual models into a single and accurate decision.

Our approach is tested and validated on synthetically prepared samples. The sequences that met the quality criteria are 81% of the total number of sequences generated by GridION x5. Demultiplexing of the barcodes in the sequence data is performed with macro-average precision of 98.2% and unclassified reads of 21%. In terms of computational efficiency, the proposed demultiplexing algorithm is an order of magnitude faster than the guppy barcoder. The chromosome classification is performed with macro-average precision of 99.1% and unclassified reads of 4%. The ground truth chromosome classification is defined using the Smith–Waterman algorithm. In comparison to this algorithm, the proposed algorithm showed significant improvement in terms of computational efficiency. The proposed algorithm manages to align a sequence in 1.1 milliseconds which is 50 times faster than the Smith–Waterman algorithm. The initial classification on the selected genetic disorders obtained on a small subset (20%) of the synthetically prepared samples showed very promising results too (macro-average precision of 98% and computational efficiency of 0.4 milliseconds per sequence). All the experiments are performed on one referent hardware architecture (Intel i7 10th generation, 8 cores, 32 GB RAM, no CUDA) using thread parallelism of 10.

Next step will be to extend the set of genetic disorders that will be detected by the proposed solution.