Using machine learning approach to improve base calling in next generation sequencing data

D.V.Antonets, N.E. Russkikh

Novel Software Systems, LLC, russkikh.nikolay@gmail.com

New generation sequencing is rapidly turning from expensive and dedicated tool of basic research into one of the most powerful techniques aimed to revolutionize the medicine. The single molecule sequencing strategy used by SeqLL (USA) simplifies the DNA sample preparation process, avoids PCR-induced bias and errors, simplifies data analysis and tolerates degraded samples. Here we developed a new machine learning approach to solve base calling task for SeqLL sequencing platform and improved the quality and performance of base calling.

On the training set of image patches corresponding to known signals obtained from synthetic oligonucleotides sequencing we created a predictive model using the extreme gradient boosting algorithm. The developed model was integrated in our custom base calling and oligonucleotides assembly pipeline developed for SeqLL sequencing technology. Our new algorithm works ten times faster, as compared to previous solution and it also demonstrated 10 % improvement in sensitivity at the same level of specificity.

All presented results are property of SeqLL company (USA).