The application of deep learning methods in the analysis of livestock genomes

Deep learning is a subfield of machine learning methodology, which has recently and rapidly been gaining importance in many fields of science. Originally, it has been developed mainly for image recognition, but nowadays it has also been increasingly used in other fields, including genomics. According to the Editorial view of the Nature Genetics journal from January 2019, deep learning algorithms are to "to revolutionize genome analysis". Their applications range from gene expression analysis, through modelling of gene expression regulation, to proteomics. However, in livestock genomics, analyses involving deep learning remain very sparse. The next generation sequence technology for exploring individual variation in DNA sequence of whole genomes is also a relatively new field of research. This data has a very high information content and thus can be used to associate the DNA variation with the variation of phenotypes or to track the phylogenetic origin of individuals. However, it has to be kept in mind that because of a high through-output nature of the generation of whole genome DNA sequence data, it is associated with an error rate of polymorphism detection, which is higher than that of small-scale DNA variation analysis. The goal of the project is to apply deep learning algorithms within the field of livestock genomics with the emphasis on the analysis of DNA sequence of whole bovine genomes.

In particular, our project is going to use deep learning for four different aspects of whole genome DNA sequence analysis. **The first**, dimensionally the smallest, data set involves the *classification* of single nucleotide polymorphisms (SNPs) of four bulls genotyped by two technologies - next generation sequencing and an oligonucleotide microarray, *into true-positive and false-positive polymorphisms*. False-positive polymorphisms are represented by SNPs, which have genotypes discordant between both genotyping technologies. The classification algorithm will identify DNA sequence features with the highest impact on generating a false-positive SNP. **The second** data set, involves *classification* of cows *into mastitis-resistant and mastitis-prone* individuals, using point (SNPs) and structural (copy number variants, CNV) types of mutations, identified based on whole genome DNA next generation sequences of 32 Polish Holstein-Friesian cows. The classification algorithm will identify SNPs and CNVs with the highest impact on mastitis resistance. **The third** data set will be used for a *multilevel classification* problem, in which SNPs and InDels identified from whole genome DNA sequences will be used to *assign individuals to breeds*. For this purpose we will use data from 1 000 Bull Genomes Project (run7), consisting of 3 103 individuals representing various cattle breeds. The classification algorithm will identify SNPs and InDels characteristic for the breeds. The **forth** problem will use the same data set as described for problem one and apply recurrent neural networks in a context of *SNP genotype imputation*, i.e. *prediction* of the full set of SNP genotypes corresponding to the whole genome sequence, based on a subset of SNP genotypes identified by an oligonucleotide microarray.

In summary, SNP status classifier will select DNA sequence features with the highest impact on wrong polymorphism identification and can be used for prioritising SNPs obtained by the next generation whole genome DNA sequence analysis. Predictors developed for clinical mastitis will identify SNPs and CNVs associated with the risk of clinical mastitis and allow for assessing this risk for each cow with SNP or CNV genotypes available even prior to its entering of the production system. The breed classifier selects SNPs that are the most characteristic for each breed and consequently points at genomic regions contributing to breed-specific features.

Within the frame of the project, we also plan to acquire new data, independent of those used in the methodological part of the project. This data will be used to validate the classification algorithms (parts 1-3) and the SNP genotype imputation algorithm (part 4). Obtaining good validation results, manifested by a low level of incorrect class assignments, on an independent data set, significantly increases the reliability of the results and their future applications in other studies.