A Hierarchical Bayesian Model for Ranking of Genes in Metagenomics Based on Differential Abundance

Viktor Jonsson[1], **Olle Nerman**[1] **and Erik Kristiansson**[1]
[1] Mathematical Sciences, University of Gothenburg and Chalmers University ofTechnology, Sweden.

Metagenomics is a growing research field within ecology and medicine where entire communities of microbes are studied on the genome level. The majority of bacteria in the environment are unculturable and therefore difficult to study individually. Metagenomics does not rely on cultivation and is suitable for analysis of bacteria in their natural communities. The aim is to gain insight into such communities by observing differences in the abundance of genes between environmental conditions.

The statistical challenge lies in finding the genes which have a large enough difference between conditions to be statistically significant. However, the discrete and overdispersed nature of the data makes analysis harder and methods based on normality assumptions become suboptimal. In addition, the number of samples is often small while the number of genes present in a bacterial community is vast. This creates the problem of finding a few truly differentially abundant genes in a sea of noise.

We present a novel statistical model for inference in metagenomics. It is based on a generalized linear model with a canonical log-link but we extend this to include robust moderation of gene-specific variances. The moderation is achieved by putting a hierarchical structure on the variances and assuming that the gene-specific variability is sampled from a global distribution. The model is implemented in a Bayesian framework and the analyses rely on Markov Chain Monte Carlo (MCMC) sampling. Model performance has been evaluated on both simulated and real datasets. The evaluation shows that it has better performance than standard methods, including ordinary generalized linear models, t-tests and variance stabilizing transforms. We conclude that hierarchical Bayesian modeling can substantially increase the power of statistical inference in metagenomics.

Keywords: Metagenomics, Bayesian modelling, Biostatistics.