Analysis of the mutation rate in whole-genome cancer data to understand cancer development

Johanna Bertl, Aarhus University

Abstract

Understanding the mutational process in cancer cells is crucial to distinguish driver mutations, responsible for the initiation and progress of cancer, from passenger mutations that arise in high numbers due to disturbed cell division and repair processes in cancer cells. The heterogeneity of the mutation process on various levels makes this a challenging question: whole-genome analyses have shown that the mutation pattern differs fundamentally between different cancer types, but also between patients and along the genome depending on the genetic and epigenetic context.

Here, we analyse whole-genome DNA sequences of tumor and healthy tissue of 505 patients with 14 different cancer types (Fredriksson et al., Nature Genetics, 2014). We model the probabilities of different types of mutations at each position on the genome using a multinomial logistic regression model. Explanatory variables describe local genomic characteristics like the local base composition and epigenetic factors like replication timing. Interaction terms allow for cancer and sample specific effects. The increased precision of our position-specific model compared to modelling the number of mutations in a region comes with computational and numerical challenges: to be able to obtain estimates in this model with an extremely large n and a moderate to large p, we need to balance between compressing the data and loss of information, and we extend current implementations of estimation in GLMs. The size of the data also poses challenges to classical variable selection procedures.

Our results confirm cancer type specific mutation pattern. We also find that the impact of explanatory variables like replication timing varies between

different cancer types, which points to different mutational mechanisms. Finally, we use predictions from our model to construct the null model in a screen for genomic elements that drive cancer development. Assuming independence between positions, the null distribution of the number of mutations in a sequence is obtained by convolution.