Project C1
Feature selection in high dimensional data for risk prognosis in oncology
Dr. Sangkyun Lee, Prof. Dr. Alexander Schramm, Prof. Dr. Sven Rahmann

Cancer is an Evolutionary Process
Progression of cancer is driven by both natural selection and therapy, which requires adaptation to different niches and microenvironments.

Tumours are Moving Targets
- Understanding inter- and intra-tumour heterogeneity
- Identifying genomic biomarkers for risk prognosis

Resource-Efficient Feature Generation from DNA
Detailed Analysis of Variant-Tolerant Read Mapping by Min-Hashing
- Min-hashing to estimate the Jaccard similarity $J(K_1, K_2) = \frac{|K_1 \cap K_2|}{|K_1 \cup K_2|}$ between k-mer sets $K_1, K_2$
- Known variants may yield additional k-mers which may be added to the reference $K_0$
- We showed: beneficial only for variants with high population frequency. ([Schnider and Rahmann 2013])

Correspondence between specific SNVs and Methylation Level Changes
- Inter-individual methylation changes may often be traced to a single causative SNV.
- State (unmethylated, semi-methylated, fully methylated) requires discretisation of methylation level in $[0, 1]$
- Achieved with beta mixture models, computed with a novel EM-type algorithm using moment estimators in the M step. ([Schnider and Rahmann 2017])

Identification of Sparse Feature Graphs
Preliminary Model similarity between protein complexes
- Focus: Dependencies between expression levels of genes
- Estimation of inverse covariance matrix $\Sigma = \Sigma^{-1}$ with partial correlations

Penalized likelihood approach, assuming multivariate normal distribution:
$$\min_{(\Theta, \beta) \in \mathbb{R}^p} -\log \det \Sigma + (\beta^T \Theta + R(\beta))$$

LASSO:
- $R(\beta) = \lambda |\beta|_1$
- Determine $\lambda$ by controlling the family-wise error rate (restrictive) or by cross-validation (sample dependent)

SLOPE:
- $R(\beta) = \sum_{i=1}^{p^2} |\beta_i|_r$
- Determine $\lambda$ by False Discovery Rate (FDR) control
- Efficient saddle point optimisation for the Dantzig selector formalisation

Feature Extraction and Selection from High-Throughput Sequence Data
Extremely high feature dimension vs. small number of samples:
- Up to millions of genetic variants
- How to find interpretable tumour-specific variants?

Analysis of Multiple Data Types
- Next-generation sequencing, microarrays, arrayCGH
- Gene expression, methylation, copy number variations

Computational Efficiency
- Resource-efficient read mapping and alignment (using GPUs)
- Resource-efficient variant filtering

Tumour Subtype Identification
- Bioclustering of patients and variants by a modified Boolean matrix factorization (BMF)
- Solve NP-hard BMF by proximal alternating minimization of relaxed objective

Tumour Evolution
- Mutational dynamics were analyzed in one primary tumour and multiple relapse samples from a single neuroblastoma patient.
- A phylogenetic tree was built according to Hamming distances between the primary tumour and the relapse samples (Neighbor Joining).
- A stable pattern of bifurcation emerged, suggesting that at least two different subclones developed independently from the primary tumour.
- Every branch has 100% bootstrap support and is robust against various perturbations of the input data.

Intratumoural Heterogeneity
- Allele frequency shifts between matched pretreatment primary and relapse neuroblastomas.
- Most SNVs detected in relapse tumours are below the whole-exome sequencing detection limit in the primary tumour.

Table of contents:
- Methodology
- Results
- Artificial Intelligence Group
- Chair of Genome Informatics
- University of Duisburg-Essen