EXTERNAL VALIDATION AND CHARACTERISATION OF GLIOBLASTOMA (GBM) CANCER SUBTYPES USING A SURVIVAL BASED BAYESIAN CLUSTERING (SBC) MODEL

Duitama,C.[1] ; Ahmad, A.[2] ; Fröhlich,H.[3]

Life Science Data Analytics & Algorithmic Bioinformatics, Bonn-Aachen International Center for IT, Bonn

[1]s0caduit@uni-bonn.de, [2]ashar@bit.uni-bonn.de, [3]frohlich@bit.uni-bonn.de

Introduction

The background of this master thesis is the SBC, a model that infers clinically relevant cancer subtypes by jointly clustering molecular data along with survival data in a semi-supervised manner. In the original paper, the emphasis was on cluster discovery along with cluster characterisation on a Breast Cancer data set and a GBM (Glioblastoma Multiforme) data set. The goal of this project was to perform an external validation of the SBC clusters trained on the Verhaak study using the TCGA patient cohort, and characterise the obtained clusters.

Past Work

SBC (Fig.1)

- SBC is a fully bayesian approach as omics data contains a lot of noise with p >> n.
- Dirichlet Process to automatically infer the number of clusters.
- Molecular Data modelled as a Hierarchical Multivariate Gaussian Distribution (Mixture model).
- Survival time is modelled as Log-linear (Accelerated Failure Time) distribution with molecular covariates (Mixture model).
- L-1 regularization for the covariates of the Survival Model (Bayesian Lasso)

iSBC (Fig.2)

- Cluster-specific Independence assumption allows for the integration of more than one data source.

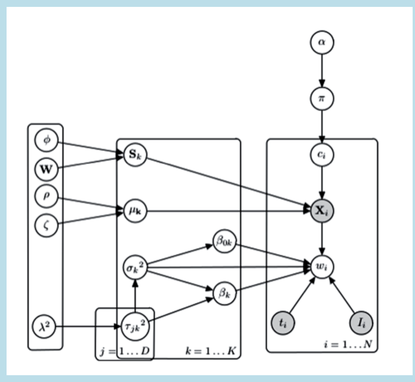

Fig. 1: Graphical Model SBC

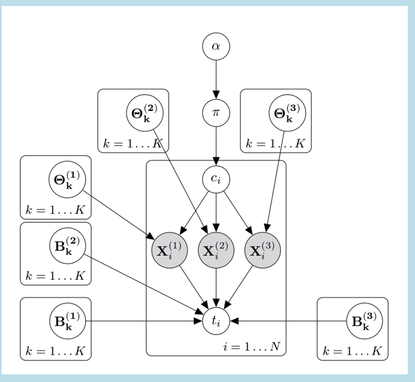

Fig. 2: Graphical Model iSBC

Preliminary Results

Training

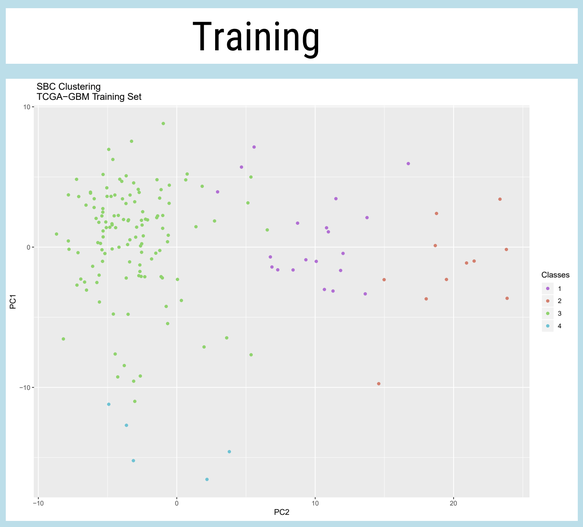

Fig.3: PCA Training Set

Testing

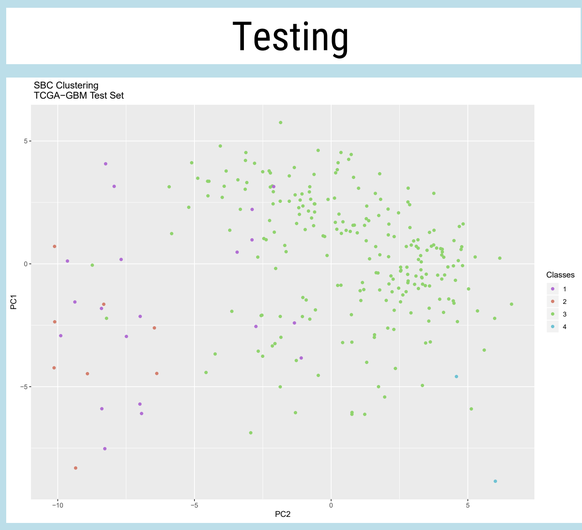

Fig.4: PCA Test Set

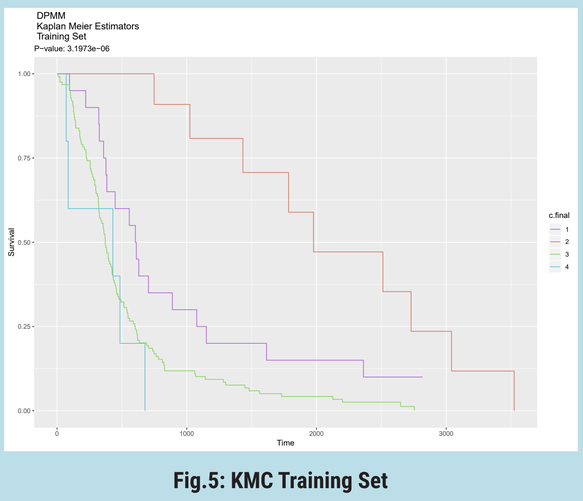

Fig.5: KMC Training Set

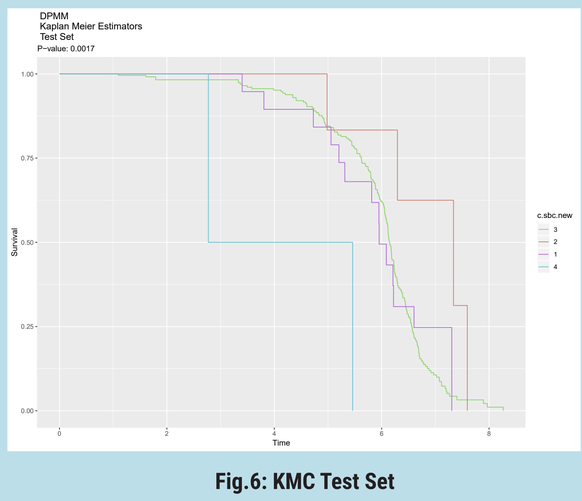

Fig.6: KMC Test Set

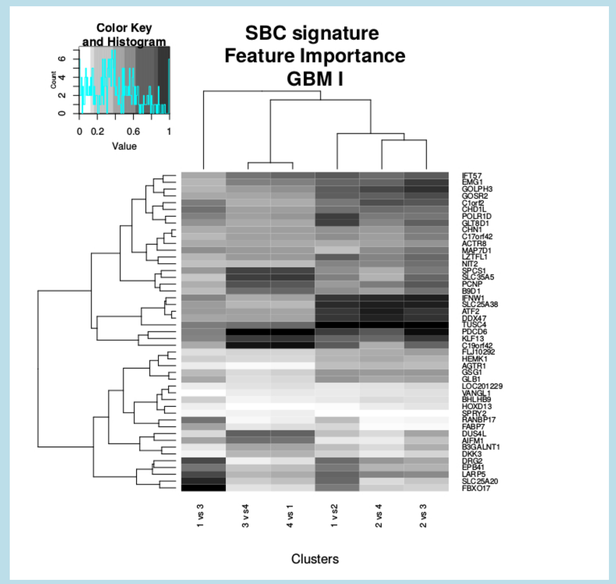

Fig.7: Feature Importance

- To ensure the clinical data for training and testing sets was similar, we matched the Karnofsky Index Distribution on both sets.
- Effect of scaling technique: ComBat to reduce batch effects, Moment Matching and independent scaling. The last one gave the best performance.
- Effect of model initialisation: K-means performed better than Flexmix.
- Effect of including background pathway information: Oncogenic Pathway gene sets performed better than KEGG Pathway gene sets.
- Karnofsky corrected sample after independent scaling and K-means initialisation gives four clinically distinct clusters (pval= 3.2E-06) on the 160 Ver haak samples (training set), and 4 clusters (pval=2.0E-03) on the external data set.
- Feature importance of the SBC model shown in Fig.7.

Future Work

- Feature preprocessing using auto encoders for pathway aggregation scores, and aggregated scores of survival and molecular data.
- Further biological characterisation of the clusters looking into other data modalities for understanding clusters obtained, such as Clinical Data, Copy Number Data, Somatic Mutation Data, Methylation data and miRNA MicroArray Data.
- Exploring iSBC clusters.

Code is available in the following GitHub repository: use QR code