SUMMARY OF PROJECTS
DREAM9 CHALLENGE, TCGA DATA

James McMurray,
Prof. Dr. Bernhard Schölkopf

Max Planck Institute for Intelligent Systems
Tübingen, Germany

DREAM9

BACKGROUND

- The DREAM (Dialogue for Reverse Engineering Assessments and Methods) project runs several challenges in bioinformatics annually
- Chose to participate in the DREAM9 AML (Acute Myeloid Leukemia) Outcome Prediction challenge
- Acute Myeloid Leukemia is a particularly lethal type of leukemia, only approximately a quarter of the patients diagnosed with AML survive beyond 5 years.
- Challenge consists of three sub-challenges related to outcome prediction

DETAILS

- Provided proteomics data and many clinical covariates (e.g. age, whether the patient has received previous chemotherapy, etc.)
- Subchallenge 1: Determine the best model to predict which AML patients will have Complete Remission or will be Primary Resistant. (Classification)
- Subchallenge 2: For patients who have Complete Remission, predict remission duration. (Regression)
- Subchallenge 3: Predict the overall survival time for each patient. (Regression)

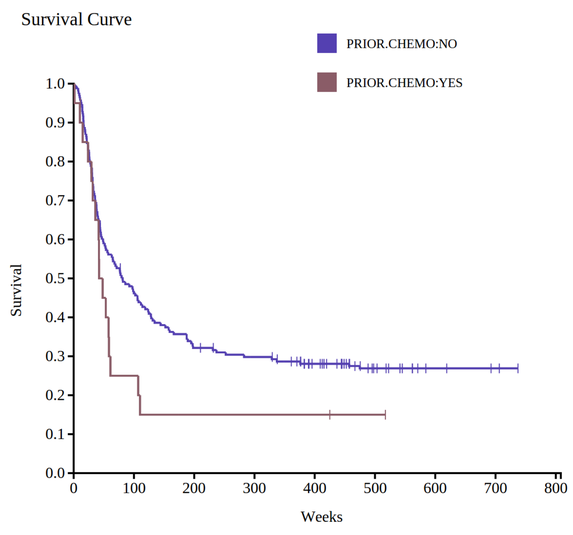

Figure 1: The survival curve for whether the patient has previously received chemotherapy.

METHODS

- Missing data was interpolated using Gaussian Processes.
- Most successful approach has been using Random Forests for classification and regression.
- Also tried dimensonality reduction via the Gaussian Process Latent Variable Model - this was useful in Subchallenge 2.

RESULTS SO FAR

Subchallenge	My best	Current best
1	0.5965, 0.6728	0.7802, 0.8148
2	0.6061, **0.6918**	0.698, 0.6557
3	0.6309, **0.7178**	0.7186, 0.5997

Table 1: The results so far - have achieved reasonable success on SC2 and SC3. For SC1, the first column is the Balanced Accuracy and the second is the Area Under Receiver Operating Characteristic Curve. For SC2 and SC3, the first column is the Concordance Index and the second column is the Pearson Correlation Coefficient.

TCGA

BACKGROUND

- The Cancer Genome Atlas (TCGA) provides a lot of public data of various types (RNASeq, DNA Methylation, MicroRNA, SNPs (restricted access), expression arrays, etc.) for many different types of cancer.
- Main aim is to discover links between the different types of cancer.
- Chose to analyse GBM (Glioblastoma multiforme) cancer data.
- Verhaak, R.G., et al. (2010) *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1*. Cancer Cell. 17(1):98-110
- Previously identified four sub-types of GBM using factor analysis and consensus clustering.

METHODS

- Use Gaussian Process Latent Variable Model in GPy software package for dimensionality reduction.
- Use K-Means for clustering in latent space.

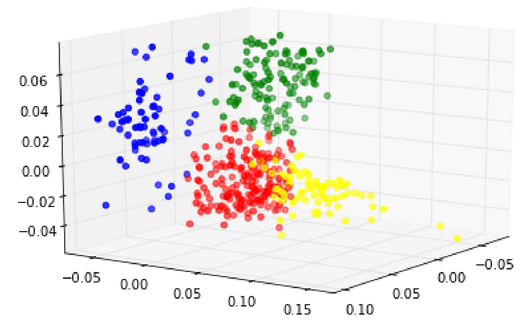

Figure 2: The clusters found by K-means in the 3D latent space produced by the GPLVM.

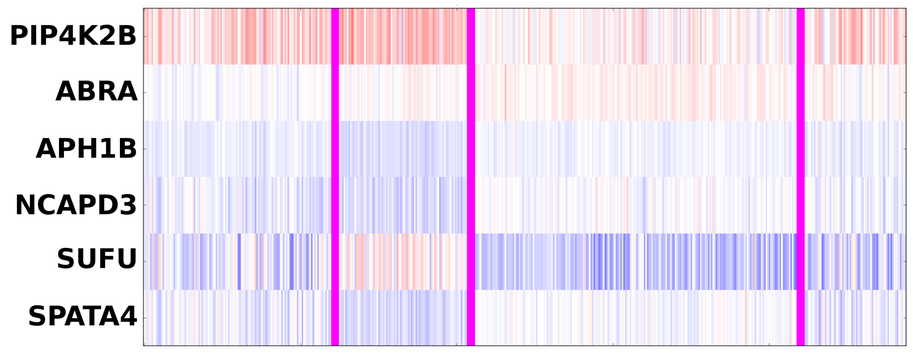

Figure 3: The expression of the genes (rows) across the GBM samples (columns). The magenta lines delineate the clusters.

Cluster	Total dead	Mean survival time of the dead (days)
0	91/120 (75.8%)	40.2
1	65/82 (79.3%)	41.8
2	139/206 (67.5%)	37.7
3	47/65 (72.3%)	**528.4**

Table 2: The mean survival time of the dead demonstrates clinical differences between the clusters.

CONCLUSIONS SO FAR

- Note we observe different significant genes.
- Still need to compare Gene Ontology enrichment analysis to test biological significance of clusters.
- Would be interested in applying Deep Gaussian Processes and hierarchical learning to these problems.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861