Understanding Glioblastoma using Bayesian Graphical Modeling

Nishant Subramani

Glioblastoma multiforme (GBM) is the most common, malignant and aggressive primary brain tumor that affects humans. Survival rates are low--the median survival rate without treatment is 4.5 months, but with standard-of-care and chemotherapy, this rises to 15 months (Johnson 2011, Van Meir 2010). The goal of this study is to relate pathology features, various clinical measurements in the study of disease, with gene expression using a Bayesian graphical model approach. This could provide invaluable information for developing an intervention method or early identification technique using gene sequencing. Studies that focus on learning from gene expression data have the possibilities of changing the way disease is thought of. Patients may have a strong indication at birth of certain diseases just due to their genetic code, but we have yet to understand gene expression data at this level for some diseases, especially Glioblastoma. The hope is that if we can, disease prediction and identification can be done at birth, and a child can be put on the proper regiment to cure the illness as early as possible.

One of Stanford's Biomedical Informatics Groups has a dataset associated with Glioblastoma that they are beginning to use for this task. The lab is willing to give me access to the data early next week. The format of the data and types of features the data contains is unclear at this juncture beyond the goal at hand as I have not seen any part of the dataset yet. I anticipate having to curate the gene expression data by identifying significant associations between certain parts of the sequence data and certain pathology features. Much of the gene expression data is almost meaningless and useless for this task and the methodologies to identify the important features must be used. These feature selection methodologies include basic Wald Chisquare statistical techniques, variable importance algorithm using Random Forests, and step-wise regression methods.

Overall, however, this project serves to utilize Bayesian graphical models and the probabilistic dependencies that Bayesian models utilize with an emphasis on the K2 algorithm as an appropriate heuristic to identify the optimal structure given the GBM pathological data. Evaluation of this model could employ simple N-fold cross-validation or a held out test set depending on the amount of the data available. Normally, however, the number of examples for gene expression data is far less than the number of features present in the dataset. This is a significant problem, and penalized logistic regression and very aggressive feature selection methodologies may be necessary. The hope is that through a graphical model approach, a structure of dependencies can be learned between various pathological features of Glioblastoma and gene expression data through which GBM in a gene expression capacity can be understood. The prospects of the study include a better identification of risk levels for GBM at birth through a genetic screening and early intervention methodologies can help mitigate the effects of the disease at a later stage in life.