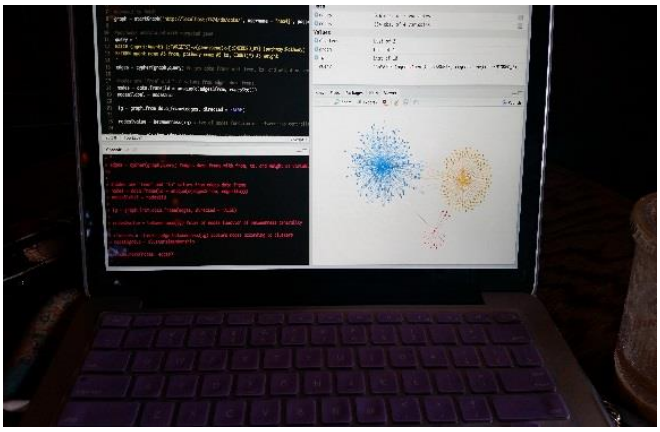

An Application of Graph Databases to Capture Complex Relationships in Translational Pediatric Cancer Research

Joyce Wang (EAS 2018)
Advisor: Deanne Taylor

During my research experience at the Children's Hospital of Pennsylvania, I worked under the mentorship of Dr. Deanne Taylor and Pichai Raman to construct a graph database that captures the complex relationships in translational pediatric cancer. Translational pediatric cancer data is very complex, as it tries to capture the relationships between genes, pathways, clinical phenotype, and drug efficacy/response. Common database structures such as relationship databases do not allow for efficient capture of theses complicated relationships that are inherent in the data. In order to meet this need I developed a graph database for the study of pediatric cancers, specifically neuroblastoma, an embryonal tumor of the autonomic nervous system and the most common extracranial solid tumor.

Graph databases take advantage of complex data connections and store data in the form of nodes and relationships, which can each contain any number of properties. They excel at managing highly connected data and relationships; connections do not have to be inferred, thereby cutting down computation time. Our initial focus was to capture neuroblastoma cell line data in a Neo4j database and use this infrastructure to discover novel findings. Neo4j is an open source NoSQL graph database that is accessible from most programming languages using a built-in REST API interface. I accessed the database through R, which includes packages that allow for effective visualization and manipulation of the data.

The particular neuroblastoma cell line being stored contains information on genes, pathways, and treatments, which will be modeled as nodes, while information on treatment response and relevant genetic lesions will be modeled as relationships between the respective nodes. I believe that this method will allow us to better capture the intrinsic connections present in the data, and consequently allow us to easily uncover interesting patterns and relationships. After performing several proof of concepts using the database I constructed, there were some promising results, but further modifications regarding the type and format of the data loaded into the database need to be implemented.

Throughout my internship, I learned about graph databases and familiarized myself with the R programming language, along with neuroblastoma treatment research. More importantly, however, I learned how to interact with others in a work environment and cooperate in team settings. I learned that there were lots of collaborations and meetings in addition to individual work involved with research, and that you need strong people skills in order to be successful. Additionally, I found that I really enjoyed the field of bioinformatics and I would like to explore this field more in my remaining time here at Penn and in my future career.