Abstract

JENNIFER M. STAAB: Systematic approaches to integrate inconsistent, noisy high-throughput data to bolster subtle relationships obscured by standard analyses. (Under the direction of Shawn M. Gomez.)

The increasing availability and decreasing cost of high throughput technologies coupled with the availability of computational tools form a basis for a shift to a more integrated approach to analyzing biological processes. In particular, classical statistical analysis techniques are designed to analyze data characterized by a single data source and are distinguished by a much higher ratio of subjects to the number of observations. In contrast, bioinformatics and systems biology applications often involve large data sets characterized by an abundance of observations spawned from a relatively small sample of subjects. The complexity of these systems coupled with the need to integrate inconsistent (noisy) data require appropriate methodologies that address these issues.

Standard analyses can proficiently identify associations within consistent data, but these approaches are not robust at identifying relationships across data sources and/or where nontrivial amounts of inconsistency (noise) are present. Such data requires approaches that account for this increasing inconsistency within the data. One technique of accounting for such inconsistency is to limit analyses to subsets of data where the desired associations are the most prominent. Challenges for this particular approach involve the determination of subsets of interest while simultaneously establishing a metric with which to judge statistical importance.

My initial work using this approach involved providing a methodology to represent Nuclear Magnetic Resonance (NMR) Spectra as hundreds of aligned peaks as opposed to thousands of unaligned points, which allows for more sophisticated means of analysis. My later work explores the development of data mining methodologies for identifying associations that exist within subsets of inconsistent, noisy data while addressing how to sensibly target subsets of interest while establishing a metric of association that provides statistical significance. Two approaches were developed, the first of which established a p-value associated metric, while the latter allowed for multiple arbitrary metrics of interest to be used to identify statistically significant patterns. This work helps to establish methodologies for the identification of rare, but significant patterns in large noisy data sets.