Using Scan Statistics for Anomaly Detection in Genetic Networks

Christopher C. Overall, J.L. Solka, J.W. Weller, Carey E. Priebe

We have applied graph-based scan statistics to genetic networks in order to detect anomalies in gene activity over time. An anomaly may manifest as a single gene with excessive connections to other genes or it may manifest as chatter, in which a neighborhood of genes centered about a particular gene exhibit excessive connections with one another.

Scan statistics are commonly used in image analysis to detect the presence of an anomalous local signal in a time-series of images. Based upon these techniques, Priebe et al. (2005) developed a theory of scan statistics for graphs and then applied the new methodology to detect anomalies in a time-series social network that they developed from the publicly available Enron email dataset.

The current work has been adapted from that of Priebe et al. (2005) and applied to datasets generated from time-series transcription profiling experiments, such as those using microarrays and reverse transcription-coupled PCR. Time-series gene expression datasets are typically represented as a matrix in which each column corresponds to a time point and each row corresponds to a gene. Hence, each element of the matrix represents the expression value of a single gene at a single time point. Prior to applying the graph-based scan statistic methodology to a gene expression dataset, each of the time points are transformed into a graph with the genes as vertices and an edge drawn between two genes if there is a relationship between them. Either univariate or multivariate model-based clustering is applied to the gene expression values at a particular time point and the posterior probabilities are then used to determine relationships. After each time point is transformed into a graph, graph-based scan statistics are then utilized to detect any anomalous gene activity. We applied this methodology to the following two datasets for analysis: the yeast *Saccharomyces cerevisiae* cell cycle dataset (microarray) from Spellman et al. (1998) and the rat central nervous system development dataset (reverse transcription-coupled PCR) of Wen et al. (1998).