Community Detection in Biological Networks: Lessons from the DREAM 2016 Module Identification Challenge

Abstract: The 2016 DREAM Disease Module Identification Challenge was developed to systematically assess the state of computational module identification methods on a diverse collection of molecular networks. Six different anonymized networks were presented with the gene names anonymized. The goal was to partition the genes into non-overlapping modules of from 3-100 genes each, based soley on the patterns of network connectivity. Collections of modules were scored based on the number of modules that were statistically significantly
enriched for a set of trait or disease-related phenotypes, according to a set of previously published GWAS datasets. For the first subchallenge, gene names were anonymized separately for each network and it asked for modules in each of the six networks considered separately; the
second subchallenge used the same identifier across networks and asked for one collection of modules integrating information together from all six networks.

We won the first DREAM Disease Module Identification subchallenge and performed in the top cohort for the second subchallenge in a field of 42 interational teams. Our winning "Double Spectral" method used the ``Diffusion State Distance" spectral graph metric of Cowen and Hescott et al. to define a matrix of pairwise dissimilarities between all nodes of the network. We then ran an off-the-shelf spectral clustering algorithm on the matrix of these distances, using the training rounds to set the target number of clusters the spectral clustering should return. (We also made an additional attempt to find dense bipartite structure in some of the networks, but it did not seem to help our performance in the end).

After the contest concluded, we partnered with the contest organizers to again run our Double Spectral method, this time on a consensus matrix that captured the co-occurence of different genes in the same clusters among all the top performing teams of the challenge. The results produced the best set of communities, better than any team's individual cluster predictions. We find many interesting modules corresponding to core disease- and trait revelant pathways, including for extreme height, rheumatoid arthritis and inflammatory bowel disease. All contest networks and team predictions are in the process of being released to the public, where they will provide an important benchmark dataset for computational researchers seeking to test methods for finding disease modules.

Bio: Lenore Cowen is a professor in the Department of Computer Science at Tufts University, with a courtesy appointment in the Department of Mathematics. After finishing her Ph.D. in mathematics at MIT in 1993, she was an NSF postdoctoral fellow and then joined the faculty of the Mathematical Sciences Department (now the Applied Mathematics and Statistics Department) at Johns Hopkins University. She joined Tufts in 2001. Her research interests span three areas: discrete mathematics, algorithms, and computational molecular biology.