One of the major goals of gene expression data analysis is to uncover genetic pathways, i.e., chains of genetic interactions. For example, a researcher may be interested in identifying the genes that contribute to a disease. This task is difficult because subgroups of genes display similar activation patterns only under certain experimental conditions. Genes that are co-regulated or co-expressed under a subset of conditions will behave differently under other conditions. Finding genetic pathways therefore could be aided by identifying clusters of genes that are co-expressed under subsets of conditions as opposed to all conditions. A high degree of correlation between the activity levels of subsets of genes under subsets of conditions does not of course necessarily imply causality relations. Further biological analysis would be required to find actual genetic pathways.

In this talk, I will describe how data is generated to find genes that are co-regulated. I discuss problems associated with microarray data collection, such as noise, noise characterization and missing data. I focus in particular on the selection of material, using the detection of ovarian cancer markers as an example. Next, I will provide an overview of biclustering techniques and their applications in genomics. Biclustering is also known as bidimensional clustering, subspace clustering and co-clustering in other application fields. Biclustering techniques produce local models whereas clustering approaches compute global models. If we use a clustering algorithm on the rows of the gene expression matrix, a given gene cluster is defined using all the conditions. In contrast, a biclustering technique will assign a gene to a bicluster based on a subset of conditions. Furthermore, when a clustering algorithm is applied to the rows of the gene expression matrix, it assigns each gene to a single cluster. Biclustering techniques on the other hand identify clusters that are not mutually exclusive or exhaustive. A gene may belong to no cluster, one or more clusters.

By simultaneously clustering the rows and columns of the gene expression matrix, one can identify candidate subsets of conditions that may be associated with cellular processes that exhibit themselves only or subsets of genes that potentially play a role in a given biological process. I will discuss the biological significance of the biclusters identified by analyzing yeast gene microarray data and gene expression data from tissues provided by the University of Minnesota Cancer Center’s Tissue Procurement Facility. Finally, I will conclude by describing differential biclustering and its application to revealing conserved and diverged co-expression patterns between related organisms.

Ahmed Tewfik received his B.Sc. degree from Cairo University, Cairo Egypt, in 1982 and his M.Sc., E.E. and Sc.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, in 1984, 1985 and 1987 respectively. Dr. Tewfik has worked at Alphatech, Inc., Burlington, MA in 1987. He is the E. F. Johnson professor of Electronic Communications with the Department of Electrical Engineering at the University of Minnesota, and is an IEEE Fellow. He served as a consultant to MTS Systems, Inc., Eden Prairie, MN, Emerson-Rosemount, Inc., Eden Prairie, MN, CyberNova, Milpitas, CA, Macrovision, Santa Clara, CA, Visionaire Technology, Fremont, CA and Ipsos, New York. He worked with Texas Instruments and Computing Devices International. From August 1997 to August 2001, he was the President and CEO of Cognicity, Inc., an entertainment marketing software tools publisher that he co-founded, on partial leave of absence from the University of Minnesota. His current research interests are in genomics and proteomics, programmable wireless networks, brain computing interfaces, healthcare safety and datanomic and pervasive computing and storage.