

Assessing Intercorrelations in Big Data Using Kernel Density Estimation and Animation With Applications to Genomic Cancer Research

Eugene Demidenko

Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA

Abstract

Genomic data is big and prone to hidden features such as internal intercorrelations and clusters/groups. These phenomena are illustrated by the classification of ovarian cancer subtypes based on the analysis of 12,000 gene expressions from 489 ovarian biopsy carcinomas obtained from The Cancer Genome Atlas (TCGA) Research Network. To understand the intercorrelation of genes in those datasets, researchers must analyze about 72 million correlation coefficients – this problem is impossible without smart condensation of the data and its visualization. To work around this problem, we will discuss the concept of Gene Rank (GR), recently introduced by the author as the maximum eigenvector, which has a close connection with the Page Rank used in the GOOGLE search engine. GR is illustrated by plotting the connectivity of ovarian cancer genes to identify the most and least connected genes, using kernel density distribution to visualize the data rather than raw observations. A surprising finding was discovered: while the known oncogene BRCA1 is at the center of typically connected genes, other oncogenes, such as TP53 and KL, are least connected (we call these genes 'lonely killers').

As confirmed by various studies, ovarian cancer is heterogeneous and can be classified by several subtypes – identifying the subtype of a particular cancer patient is the key to personalized medicine. We identified four ovarian cancer subtypes by optimally projecting the genes from 489 biopsies onto the 3D sphere shown in an animated fashion, where each frame displays the connected pair of genes with the correlation coefficient decreasing from 1 to 0. This plot clearly displays 3 to 4 subtypes, in agreement with other independent biomedical studies.

In summary, big data requires novel methodological techniques for analysis and visualization, including distribution functions and graphical animation.