Abstract:

Synonymous mutations, which represent a single nucleotide change in the codon without affecting the resultant amino acid, have largely been considered inconsequential. However, evidence has surfaced that these mutations do have consequences that affect protein folding kinetics which can lead to structural and functional changes. The purpose of this research is to explore the effects of synonymous mutations in breast cancer cells. Specifically, the analysis is focused on MUC6, a secreted mucin protein that is highly susceptible to mutations of this type, in an attempt to explore how it might propagate human breast cancer.

Using data from The Cancer Genome Atlas project and Mertins, Philipp et al., we analyzed and quantified both synonymous and missense mutations in a cohort of 69 breast cancer patients. Additionally, gene expression was explored using RNA-seq and protein production data in the form of iTRAQ ratios which were collected and analyzed for each patient. The mutation and gene expression data were analyzed using a variety of linear and non-linear methods including principal component analysis (PCA), multiple factor analysis (MFA), t-distributed stochastic neighbor embedding (t-SNE), and random forest.

Analysis of the mutation data showed extreme hypermutability in regard to both synonymous and missense mutations in the MUC6 gene across all patients. The mutation rate discovered is extremely high compared to all other affected genes in our dataset, however, many of the other hypermutated genes are also members of the mucin family including MUC4, MUC12, and MUC16. It appears that the actual production of MUC6 in our patients is decreased when compared to a pool of cancer patients. Lastly, using non-linear methods, it was discovered that

RNA-seq, the number of synonymous mutations in a gene, and the cancer stage were the most important features when using machine learning algorithms in an attempt to predict protein production values.