

# **Correlation as a function of distance in oligonucleotide space: Using the correlation to conduct more efficient inference using oligonucleotide arrays**

Cavan Reilly, Ph.D.  
University of Minnesota – Twin Cities

## **Abstract:**

Here we examine how sequence similarity among oligonucleotides on oligonucleotide arrays induces correlation in the measured intensity values. There is substantial correlation between the positive match and mismatch probes, and while some of this correlation could have a spatial component (due to adjacency of the 2 probes) we argue that a considerable portion of this correlation can be accounted for by probe sequence similarity. Our approach is similar to the approach taken in classical spatial statistics (i.e. Geostatistics), except here our space is the space of oligonucleotides. We show how to use distances on this space to derive positive definite correlation matrices for sets of oligonucleotides. We introduce a model for gene expression data that assumes the mean intensity for all positive match probes in a set is the same, as is the mean intensity for all mismatch probes, but we observe these means subject to a stochastic disturbance whose correlation structure depends on the similarity of the probe sequences. Finally we show how to use this model to exploit the correlation in sequence space to obtain more accurate estimates of gene expression from oligonucleotide arrays.