

Advancing Microarray Technology: Efficient Design of Sequence Libraries Covering All k-mers with Degenerate Characters to Improve Interaction Measurement

Kim, Ryan (School: Horace Mann School)

Microarray technologies are used to understand the driving forces in genomics and proteomics through high-throughput measurements of interactions among DNA, RNA, peptides, and proteins. Universal microarrays use sequence libraries that include all k-mers, sequence variants of length k, to enable comprehensive, universal, and unbiased measurements of these interactions. Libraries that maximize k facilitate discovery of new interactions and reveal binding properties at greater detail; however, because of the exponential growth in k of such libraries and the limited size of microarrays, developing an efficient universal library for higher k values is difficult. This project introduces a novel way to generate compact designs of sequence libraries by utilizing degenerate characters that represent all characters in the alphabet Σ . Because multiple sequences without degenerate characters can be represented by a sequence with a degenerate character, the universal library size can theoretically be significantly reduced. A greedy algorithm was developed where optimal local steps are chosen to generate the smaller library sets and a condition of at most one degenerate character in every k-mer is utilized. The new sequence libraries approach the theoretical lower bounds and achieve nearly $1/|\Sigma|$ of the original sizes for DNA, RNA, and amino acid libraries. In addition, through simulation of a protein-DNA binding experiment using the sequence libraries that incorporated the degenerate character, accurate binding scores were acquired for high-affinity k-mers. These resulting sequence libraries provide the first step towards efficiently incorporating degenerate characters into libraries and towards improved understanding of fundamental cellular processes.