Abstract

In molecular biology, it is known that proteins interact with deoxyribonucleic acid (DNA) to form protein binding sites on the DNA. It is often of interest to know the locations of these binding sites. Usually, a chromatin-immunoprecipitation sequencing (ChIP-Seq) experiment is conducted to produce reads throughout the whole genome and the reads serve as location markers for possible binding sites. The number of read occurrences in different windows of the genome gives rise to a set of count data, which is referred as read counts. Software programs known as peak callers have been created to process these count data and produce a list of predicted binding sites. However, there are always biases observed in the count data. If these biases are not properly accounted for, the ChIP-Seq analysis will be affected. Current peak callers like Model-based Analysis of ChIP-Seq (MACS) introduced by Zhang et al. (2008) have some features that take biases into account. In this thesis, a new peak caller named MACS-dispersion is modified from the original MACS to further enhance its ability to account for biases that are prevalent in ChIP-Seq experiments. A comparison between the new and the old peak callers will highlight the advantages of using MACS-dispersion to locate protein binding sites.