DNA-Chip Analyzer (dChip)

Cheng Li
Wing Hung Wong

Abstract
DNA-Chip Analyzer (dChip) is a software package implementing model-based expression analysis of oligonucleotide arrays and several high-level analysis procedures. The model-based approach allows probe-level analysis on multiple arrays. By pooling information across multiple arrays, it is possible to assess standard errors for the expression indexes. This approach also allows automatic probe selection in the analysis stage to reduce errors due to cross-hybridizing probes and image contamination. High-level analysis in dChip includes comparative analysis and hierarchical clustering. The software is freely available to academic users at www.dchip.org.

5.1 Introduction

Affymetrix oligonucleotide arrays (Lockhart et al., 1996; Lipshutz et al., 1999; Parmigiani et al., Chapter 1, this volume) have been applied to many gene expression studies in the past few years (Tavazoie et al., 1999; Cho et al., 2001; Hakak et al., 2001; Gentleman and Carey, Chapter 2, this volume; Irizarry et al., Chapter 4, this volume). Many sets of 11–20 pairs of perfect match and mismatch oligonucleotides are used to measure the underlying mRNA concentrations of genes in a sample. Besides linear normalization and average difference and signal methods provided by MAS software (Affymetrix, 2001), researchers have proposed alternative low-level analysis methods such as feature extraction (Schadt et al., 2002), normalization (Hill et al., 2001), and expression index computation (Li and Wong, 2001a; Holder et al., 2001; Irizarry et al., in press; Lazaridis et al., 2002; Zhou and Abagyan, 2002) in the attempt to improve on these aspects. Hoffmann et al. (2002) have shown that such low-level analysis methods can have a large impact on high-level results such as sample comparisons. In the rest of this chapter, we will describe the DNA-Chip Analyzer (dChip) software, which provides several low-level and high-level analysis methods.
5.2 Methods

Irizarray et al. (Chapter 4, this volume) introduce the Affymetrix array design, CEL and CDF files, and probe-level data. Interested readers can refer to Chapter 4 or Lipshutz et al. (1999) for details on oligonucleotide array data.

5.2.1 Normalization of Arrays Based on an “Invariant Set”

Since array images usually have different overall image brightnesses (Figure 5.1A, see color insert), especially when they are generated at different times and places, proper normalization is required before comparing the expression levels of genes between arrays. Model-based expression computation (Li and Wong, 2001a) requires normalized probe-level data (from Affymetrix’s DAT or CEL files). For a group of arrays, we normalize all arrays (except the baseline array) to a common baseline array having the median overall brightness (as measured by the median CEL intensity in an array).

A normalization relation can be understood as a curve in the scatterplot of two arrays with the baseline array drawn on the $y$-axis and the array to be normalized on the $x$-axis. A line running through the origin is a multiplicative normalization method (Affymetrix, 2001; the scaling method in Affymetrix Microarray Suite software (MAS)), and a smoothing spline through the scatterplot can also be used (Figure 5.1A; Schadt et al., 2001).

We should base the normalization only on probe values that belong to nondifferentially expressed genes, but generally we do not know which genes are nondifferentially expressed (control or housekeeping genes may also be variable across arrays). Nevertheless, we expect that probes of a nondifferentially expressed gene in two arrays to have similar intensity ranks (ranks are calculated in the two arrays separately).

We use an iterative procedure to identify a set of probes (called an “invariant set”), which presumably consists of points from nondifferentially expressed genes (Figure 5.1B). Specifically, we start with points of all $PM$ probes (about 140,000 for the HU6800 array). If a point’s proportional rank difference (absolute rank difference in the two arrays divided by $n = 140,000$) is small enough, it is kept for the new set. By small enough we mean $< 0.003$ when its average intensity rank in the two arrays is small and $< 0.007$ when it is large, accounting for fewer points at the high-intensity range, and this threshold is interpolated in between them. These parameters were chosen empirically to make the selected points in the “invariant set” thin enough to naturally determine a normalization relation. In this way, we may obtain a new set of 10,000 points, and the same procedure is applied to the new set iteratively until the number of points in the new set no longer decreases. A piecewise-linear running median curve is then