Chapter 2

DATA PRE-PROCESSING ISSUES IN MICROARRAY ANALYSIS

Nicholas A. Tinker, Laurian S. Robert, Gail Butler, Linda J. Harris.
ECORC, Agriculture and Agri-Food Canada, Bldg. 20, 960 Carling Ave., Ottawa, ON, K1A 0C6, Canada
e-mail: {tinkerna,robertls,butlergm,harrislj}@agr.gc.ca

1. INTRODUCTION

Microarray experimentation is a young but rapidly maturing field, and the potential complexities of microarray data are spawning a rich statistical literature. Some of the concepts presented in this literature may be new and unfamiliar to the molecular biologist, who typically gleans information in a stepwise manner from many small, carefully controlled experiments. Microarrays yield enormous quantities of data and can address many simultaneous hypotheses, but results are variable and require careful preparation and statistical analysis. Gathering, organizing, and preparing data for statistical analysis is a large and important component of microarray experimentation. These steps are referred to collectively as pre-processing.

This chapter is written for the newcomer to microarray technology. It is intended as an introduction to some of the pre-processing steps that are detailed in further chapters.Broadly defined, pre-processing includes the planning and design of experiments, the acquisition and processing of images, data transformation, data inspection, and data filtering. We cannot possibly represent every variation in terminology and procedure that will be encountered in other literature, but we have attempted to introduce a variety

---

1 The design and construction of microarrays themselves, and the extensive laboratory management that accompanies this task, are not discussed in this chapter. These steps vary substantially with the type of microarray technology being used; those who intend to prepare their own microarrays should consult literature on this subject (e.g., Allain et al., 2001; Dolan et al., 2001; Rockett et al., 2001; http://www.microarrays.org).
of terminology, and to explain the basis of some major procedural differences. A flow diagram for the topics covered in this chapter is presented in Figure 2.1. Because of the strong requirement for data management at all stages of pre-processing, we begin with a general discussion of this topic.

Figure 2.1. A diagram of preprocessing steps in microarray analysis. Numbers in parentheses indicate sections where topics are discussed.

2. DATA MANAGEMENT

For most laboratories, microarray analysis will not be a one-time experience. Even if an experiment addresses one specific question, it will likely be followed by experiments to test additional factors. Properly done, a microarray experiment has the potential to become part of a larger study. This concept has inspired some authors to address the need for common data descriptions, formats, and repositories (Becker, 2001; Fellenberg et al., 2002; Kellam, 2001). Given the success and importance of large central databases for DNA and protein, there are strong reasons to nudge microarray data in the same direction. However, it is clear that microarray data is far more complex than sequence data, and that it can be meaningless or misleading unless: (1) the context of experimental conditions is fully described, and (2) measurements are standardized such that comparisons are valid. For this reason, standards such as the “Minimal Information About Microarray Experiments (MIAME)” (Brazma et al., 2001) are being developed. Meeting these standards is a responsibility that demands good data management – even if reduced to a basic need to publish reproducible results. At a minimum, forms that provide templates for the description of an experiment and its components should be used. These forms can be based on standards such as MIAME, with adaptations that are appropriate for unique characteristics of the laboratory, the experimental organism, or the methodology being employed. Separate forms can be made to describe