Chapter VI
Computational Models for the Analysis of Modern Biological Data

Tuan D. Pham
James Cook University, Australia

ABSTRACT

Computational models have been playing a significant role for the computer-based analysis of biological and biomedical data. Given the recent availability of genomic sequences and microarray gene expression, and proteomic data, there is an increasing demand for developing and applying advanced computational techniques for exploring these types of data such as: functional interpretation of gene expression data, deciphering of how genes, and proteins work together in pathways and networks, extracting and analysing phenotypic features of mitotic cells for high throughput screening of novel anti-mitotic drugs. Successful applications of advanced computational algorithms to solving modern life-science problems will make significant impacts on several important and promising issues related to genomic medicine, molecular imaging, and the scientific knowledge of the genetic basis of diseases. This chapter reviews the fusion of engineering, computer science, and information sciences with biology and medicine to address some latest technical developments in the computational analyses of modern biological data: microarray gene expression data, mass spectrometry data, and bioimaging.

MICROARRAY GENE EXPRESSION DATA

Microarrays are a relatively new biotechnology that provides novel insights into gene expression and gene regulation (Brazma and Vilo, 2000; Whitchurch, 2002; Zhang at al, 2002; Pham et al, 2006a). Microarray technology has been applied in diverse areas ranging from genetics and drug discovery to disciplines such as virology, microbiology, immunology, endocrinology, and neurobiology. Microarray-based methods are the most widely used technology for large-scale analysis of gene expression because

they allow simultaneous study of mRNA abundance for thousands of genes in a single experiment (Kellum and Liu, 2003). The generation of DNA microarray image spots involves the hybridization of two probes labelled with a fluorescent red dye or a fluorescent green dye. The relative image intensity values of the red dye and the green dye on a particular spot of the arrays indicate the expression ratio for the corresponding gene of the two samples from which the mRNAs have been extracted. Thus, robust image processing of microarray spots plays an important role in microarray technology (Nagarajan, 2003; Liew at al, 2003; Lukac et al, 2004).

DNA microarray data consists of a large number of genes and a relatively small number of experimental samples. The number of genes on an array is in the order of thousands, and because this far exceeds the number of samples, dimension reduction is needed to allow efficient analysis of data classification techniques. Many statistical and machine-learning techniques based different computational methodologies have been applied for cancer classification in microarray experiments. These techniques include linear discriminant analysis, *k*-nearest neighbor algorithms, Bayes classifiers, decision trees, neural networks, and support vector machines (Dudoit and Fridlyand, 2003; Golub et al, 1999; Guyon et al, 2002). Nevertheless, common tasks of most classifiers are to perform feature selection and decision logic.

Based on the motivation that conventional statistical methods for pattern classification break down when there are more variables (genes) than there are samples, Nguyen and Rocke (2002) proposed a partial least-squares method for classifying human tumor samples using microarray gene expression data. Zhou et al. (2004) proposed a Bayesian approach for selecting the strongest genes based on microarray gene expression data and the logistic regression model for classifying and predicting cancer genes. Yeung et al. (2005) reported that conventional methods for gene selection and classification do not take into account model uncertainty and use a single set of selected genes for prediction, and introduced a Bayesian model averaging method, which considers the uncertainty by averaging over multiple sets of overlapping relevant genes. Furey et al. (2000) applied support vector machines for the classification of cancer tissue samples or cell types using microarrays. Lee et al. (2003) proposed a Bayesian model for gene selection for cancer classification using microarray data. Statnikov et al. (2005) carried out a comprehensive evaluation of classification methods for cancer diagnosis based on microarray gene expression data.

Recently Pham et al. (2006b) carried out cancer classification by transforming microarray data into spectral vectors. The same authors used the spectral difference or spectral distortion between the pair of spectra for pattern comparison, which appears to be a potential approach for the cancer classification using microarray gene expression data.

MASS SPECTROMETRY DATA

Current best practice for reducing human mortality rates caused by complex diseases is to detect their symptoms at early stages. By early recognition of symptoms, one can get the most effective clinical treatment for the best outcome. Recent advances in biotechnology open doors to fascinating opportunities for the better understanding of the biology of many complex human diseases at molecular levels. These advances will hopefully lead to the early detection and treatment of such diseases (Petricoin and Liotta, 2003; Wulfkuhle et al, 2003).