PREFACE

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 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.

The rapid advancement of cDNA microarray technology has revolutionized genomic research in life science. However, the enormous amount of data produced from a microarray images makes automatic computer analysis indispensable. An important first step in analyzing microarray image is the accurate determination of the cDNA spots in the image. We report in Chapter 1 a novel spot segmentation method for cDNA microarray images. The algorithm makes use of morphological operations, adaptive multi-level thresholding and statistical intensity modeling to: (i) perform automatic image block segmentation, (ii) generate the grid structure automatically within each image block, where each subregion in the grid contains only one spot, and (ii) to segment the spot, if any, within each subregion. The algorithm is fully automated, require no user intervention, robust, and can aid in the high throughput computer analysis of cDNA microarray images.

Microarray data is explosively growing every day, and there exist many platforms to generate gene expression data. A problem arises because the expression files are quite different for same genes and same tissues in different platforms. The question we can ask is that: is there significant difference or not within same platform and among the different platforms, as well as how to do the meta-analysis based on the cross-platform microarray data. There are a few papers addressing normalization, cross-platform comparison, and meta-analysis, separately. In Chapter 2, we first analyze the repeatability within one platform based on quartile normalization, and then normalize different platform using z-transformation for cross-platform comparison. In the cross-platform comparison, we review the traditional Pearson correlation, rank correlation, and the co-inertia analysis, and also propose the principle component analysis based comparison method. The results show that those methods can let us to compare the different platform from different aspects. In the meta analysis, we

review the linear model, ANOVA model and two-component model for detecting differential genes based on cross-platform microarray data. In our experiments, we analyzed Affymetrix, Agilent, Amersham, Mergen, MGH, and Operon six platforms

The false discovery rate (FDR) is a measure of error in multiple testing widely used for DNA microarray data. FDR is a particularly useful measure of error especially in exploratory microarray studies where investigators aim to screen for differentially expressed genes. Various experimental designs and statistical methods have been proposed for detecting differential expression. However, it is important to determine the statistical power of a given design and method to detect differentially expressed genes at the planning stage. In this paper, we provide a general computational framework to determine the sample size and associated power using FDR as a measure of error. In Chapter 3 we illustrate the proposed sample size calculation framework with a model of gene expression that incorporates (1) heterogeneity of variance, (2) non-constant gene expression structure, and (3) measurement (technical) errors in the technology.

Biological sequence analysis is at the core of bioinformatics, bringing together several fields, from computer science to probability and statistics. Its purpose is to computationally process and decode the information stored in biological macromolecules involved in all cell mechanisms of living organisms – such as DNA, RNA and proteins – and provide prediction tools to reveal their structure, function and complex relationship networks.

Within this context several methods have arisen that analyze sequences based on alignment algorithms, ubiquitously used in most bioinformatics applications. Alternatively, although less explored in the literature, the use of vector maps for the analysis of biological sequences, both DNA and proteins, represents a very elegant proposal to extract information from those types of sequences using an alignment-free approach.

Chapter 4 presents an overview of alignment-free methods used for sequence analysis and comparison and the new trends of these techniques, applied to DNA and proteins. The recent endeavors found in the literature along with new proposals and widening of applications fully justifies a revisit to these methodologies, partially reviewed before (Vinga and Almeida, 2003).

Chapter 5 reviews the statistical methods for segmenting DNA sequences into compositionally homogeneous domains. These methods are mainly classified as change point methods and hidden Markov model based methods. These methods either use dynamic programming algorithms to obtain an optimal segmentation or heuristic approaches to decipher the complex heterogeneities in sequences. We discuss the general theories and then present the critical assessment of the methods in light of DNA segmentation. Most segmentation methods function as exploratory tools to unravel yet unknown features in genome sequences, but they can be easily adapted to perform specific tasks, for example, detection of CpG islands and isochores or delineation of coding and non-coding regions. Moreover these methods have complementary strengths and can be used in concert to mine biologically significant entities from genome sequences.

The word design problem is an important problem in DNA computing. The goal is to design a set of single-stranded DNA molecules that are structure-free and mutually *non-crosshybridizing*. Chapter 6 presents an innovative method of constructing such large sets, so called *codesets*, whose degree of structure-freeness and noncrosshybridization is controlled by a given parameter. Using a well-known, simplified model of hybridization affinity, known as the h-distance model, our method produces sets larger than those produced under similar

models, for example, a similar (slightly less realistic) model can construct 108 8-mers, while our construction produces 256 8-mers with the same constraints in terms of the strands' mutual hybridization affinity and GC-content. We further provide a justification for the claimed large sizes of produced sets by estimating their relative sizes to those of theoretically maximal sets. The existence of such large sets of structure-free, non-crosshybridizing strands is necessary for large-scale DNA computations and approaches. Current methods that employ more realistic estimates of DNA hybridization affinity have been unable to produce sets large enough to encode more than several thousands or even hundred parameters. It seems feasible to extend our methodology to include more realistic assumptions, such as making use of known thermodynamic parameters.

Studies of drug effects on cancer cells are performed through measuring cell cycle progression such as inter phase, prophase, metaphase and anaphase in individual cells. Such studies require the processing and analysis of huge amounts of image data. Manual image analysis is very time consuming thus costly, potentially inaccurate and poorly reproducible. Stages of an automated cellular imaging analysis consist of segmentation, feature extraction, classification, and tracking of individual cells in a dynamic cellular population. Image classification of cell phases in a fully automatic manner presents the most difficult task of such analysis. In Chapter 7, we considered applying several advanced computational, probabilistic and fuzzy methods for the computerized classification of cell nuclei in different mitotic phases recorded over a period of twenty-four hours at every fifteen minutes with a time-lapse fluoresence microscopy. The experimental results have shown that the proposed methods are effective and have potential for higher performance.

Chapter 8 reviews and presents genetic algorithms and statistical methods based approach for early detection and classification of breast cancer using digital mammography. A feature selection using genetic algorithms and statistical methods has been investigated and presented. Genetic algorithms and statistical methods have been combined and a hybrid system has been developed and tested on a benchmark database. A comparative analysis of the results is included in this chapter.

Motivation of Chapter 9: WebProt is an open source software tool for text mining in molecular biology texts. It is used to collect background information about genes and proteins from online literature sources. This is useful for molecular biologists working with many unfamiliar genes, like for example in a microarray experiment with thousands of genes.

Results: A study using 42 different proteins and their official synonyms/aliases showed that 69,7% of all the results from Google were useful, i.e. containing either *related protein names* or *biological functions, locations and processes* related to the query protein. This is 10% lower than in a previous experiment, so we made a filter requiring each web page to provide at least 2 results, before being taken into consideration. This increased the precision to 82,7%, with a relative recall of 84.4%, thereby increasing the F-score of the system from 82,2% to 83,5%.

Availability: WebProt can be accessed at http://www.idi.ntnu.no/ssatre/webprot/. It is implemented in Python, using Snakelets to run as a web system. The program can also be freely downloaded for academic purposes from the same place, or by sending an e-mail request to the first author of this chapter. The advantage of using the web based system is that you get access to all the results submitted and searched for by other biologists, and this will speed up some of your searches tremendously.