# CHAPTER I

# INTRODUCTION

Research in carcinogenesis relies on the analytical platform which includes methods for observation, identification, and defining the various biological characteristics of carcinomas cells. Moreover, the trends in this field are toward finding cancer genetic profiles according to point-of-care molecular diagnostic systems. In terms of molecular and cell biology, cancer is a disease of abnormal gene expression with multi-step progression. At the time of diagnosis, especially in Thailand the cancer is too advanced which causes a high risk of death. The advancement of cancer is defined by stages. The American Joint Committee On Cancer (AJCC) has introduced the TNM staging system, which describes the extent of the cancer in the body, and how far the cancer has grown into the mucosa and whether or not it has spread to the lymph nodes or distant organs. The microarray is a proper technique since it can analyze an enormous amount of gene expressions with single experiment. Although microarray technology can be efficiently used to analyze the large amount of gene expression data, the analysis process is a complex task for this technique. Therefore, a commercial or free software is usually used to manage and analyze microarray data. The commercial analysis software needs program license thus adding high costs on analysis procedure. While a free software helps for saving cost. In this study, we analyzed microarray data by using R software, free software for statistical computing and graphics which integrates suite of software facilities for data manipulation, calculation and graphical display. R is very much a vehicle for newly developed methods for interactive data analysis, especially in Bioinformatics research. The software has been developed rapidly, and extended by a large collection of the functionality provided by additional units of software called packages. The bioinformatics field uses R packages based on Bioconductor project the development project aimed to offer tools for an analysis of genomic data (e.g. sequence, microarray, annotation and many other data types). The emphasis of R programming in bioinformatics academic discipline is on DNA microarray data analysis. The design of R and Bioconductor is modular which means that users can easily write extensions and distribute codes to other users. R is suitable for microarray and genomic research because the program can be downloaded for free. Moreover, R can run on a variety of platforms including Windows, Unix, MacOS, etc. R provides an unparalleled platform for programming new statistical methods in an easy and straightforward manner. In addition to R, another softwares that can be used for analyzing different gene expression is called a package. Limma is well known for the analysis of gene expression microarray data, specifically by using linear models to analyze the designed experiments and the assessment of gene differential expression. Statistical analysis of DNA microarray experiments is still under heavy development. The statistical method is important for analyzing the microarray data since we collect raw data with various genes expression value. Then, we will use a normalization method which is based on statistical methods. The first transformation applied to expression data adjusts the individual hybridization intensities to balance them appropriately, so that meaningful biological comparisons can be made. There are a number of reasons why data must be normalized, including unequal quantities of starting RNA, differences in labeling or detection efficiencies between the fluorescent dyes used, and systematic biases in the measured expression levels.

Normalization methods may be suitable for specific platforms based on simple assumptions. The statistical algorithms that affect for normalization include: image analysis--gridding, spot recognition of the scanned image, removal or marking of poor-quality, data processing-- background correction, using for determination of spot intensities and intensity ratios, visualization of data (e.g. see MA plot), log-transformation of ratios, and global or local normalization of intensity ratios. Identification of statistically significant changes include student's t-test, ANOVA, Bayesian method or Mann–Whitney test methods tailored to microarray data sets, which take into account multiple comparisons. These methods assess statistical power based on the variation presented in the data and the number of experimental replicates, and can help minimize Type I and type II errors in the analyses. Another method is Network-based methods: Statistical methods that take the underlying structure of gene networks into account, representing either associative or causative interactions or dependencies among gene products. The well-known normalization algorithm includes: LOWESS or Loess algorithm which measures intensity-dependent normalization and is the most universal and common. Then, we estimate each element based on the array by values of log2 (Cy5/Cy3) as function of log10 (Cy3*Cy5). Next, quantile algorithm is a popular normalization technique for single-color array while Aquantile improve from quantile algorithm to analyze two-color microarray data. Microarray data may require further processing aimed to reduce the dimensionality of the data to aid comprehension and more focused analysis. Next, the data are often filtered. Filtering microarray data is a process of selecting a subset of the available probes for exclusion or inclusion in an analysis. The filtered data use some set of objective criteria (e.g., elimination of genes with minimal variance in the samples) or statistical analyses to select genes with expression levels that correlate with particular groups of samples. Normalization and filtering transformations must be carefully applied, because they can have a profound effect on the results. Different methods of statistical analysis applied to the same data set may produce different sets of significant genes. After the normalized and filtered expression data are collected, DNA microarray analysis is next. Typically, genes are investigated with comparing cancer and normal by identification the patterns of expression correlate with cancer samples or have similar patterns of expression in multiple samples.

In the recent years, biomedical and bioinformatics community has been engaged in an intensive studies on discovery of genes involved in carcinogenesis and cancer treatment which help improve the efficiency of cancer management. One of the unsolved puzzles of the cancer is to identify tumor prognosis factor genes, which has strong correlation with the severity of the disease. With the state of the art of microarray technology it is actually possible to detect cancer genes and to identify genes referring to TNM staging. So far the approach of conceptual framework can be used for cancer screening at the early stages which can save the patient's life. Therefore, this study aim to analyze the difference in gene expression profiles between normal and rectal cancer tissues by cDNA microarray, and subsequently we use free software to manage and analyze cDNA microarray data that helping for reducing costs.

## 1.1 Research objectives

To analyze the difference in gene expression profiles between normal and rectal cancer tissues by cDNA microarray.

## 1.2 Usefulness of the research (theoretical and/or applied)

1.2.1 This study is the implementation of microarray data analysis without using commercial software.

1.2.2 The finding will help improve the efficiency of cancer management.

## 1.3 Research scope

This research is a retrospective study. Tissue samples used in this study are the biopsied tissues of primary rectal cancer available from Division of Gastroenterology, Department of Internal Medicine, Maharaj Nakorn Chiang Mai Hospital during 2009-2010. Whole Human Genome cDNA microarray expression data of 6 stage III rectal cancer patients from Cancer Treatment and Research Center at Maharaj Nakorn Chiang Mai Hospital were analyzed by R software. The clinical information of six patients were collected from medical charts of the patient including: patient characteristic, disease status (staging), treatment regimen (surgery/ radiotherapy/ chemotherapy/ hormone), pathological report, patient status (date of disease progression; local, regional, distant metastasis, date of death, date of last follow up, and date of lost to follow up).

We analyze the difference in gene expression profiles between normal and rectal cancer tissues by cDNA microarray. The study was approved by the Clinical Research Ethics Committee at Faculty of Medicine, Chiang Mai University.

## 1.4 Research location

Bioinformatics Research Laboratory (BIRL), Faculty of Science, Chiang Mai University, Chiang Mai, Thailand.

Cancer Treatment and Research Center, Department of Radiology, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand.

## 1.5 Research duration

12 months.