

## STATISTICAL AND SVM-BASED ONCOGENE DETECTION OF HUMAN CDNA EXPRESSIONS FOR OVARIAN CARCINOMA

CHING-HAO LAI<sup>1</sup>, SHYR-SHEN YU<sup>1</sup>, JUN-DONG CHANG<sup>1</sup> AND MENG-HSIUN TSAI<sup>2,\*</sup>

<sup>1</sup>Department of Computer Science and Engineering

<sup>2</sup>Department of Management Information Systems & Institute of Bioinformatics

National Chung Hsing University

250, Kuo-Kuang Road, Taichung, 402, Taiwan

{ phd9514; pyu; phd9607 }@cs.nchu.edu.tw; \*Corresponding author: mht@nchu.edu.tw

Received April 2008; revised September 2008

**ABSTRACT.** *In this paper, the human ovarian cDNA expression database is analyzed for discriminating oncogenes according to different pathological stages of ovarian carcinoma. The human ovarian cDNA expression database of this paper collects 41 patient samples, which includes 13 samples at normal ovarian tumors, 6 samples of borderline of cancers, 7 samples of ovarian cancer at stage I and 15 samples of ovarian cancer at stage III. Due to 9,600 genes of each pathological sample, a large number of genes are analyzed and discovered difficultly. For this reason, linear regression and analysis of variance (ANOVA) are used to discover and detect 21 notable oncogenes. Furthermore, these 21 notable oncogenes are divided and examined by support vector machine (SVM) with 5 different classifications according to their gene expressions of pathological stages. From the experimental results, the average accuracy of 5 classification experiments is 89% in cross validation. Moreover, the related scientific literatures also indicate these 21 discovered oncogenes are related to ovarian cancer or other cancers. It proves these discovered oncogenes are highly related to different cancers. Finally, this paper also develops a graphical user interface (GUI) bio-statistical system for gene expression analysis to assist doctors and pathologists to analyze and diagnose ovarian cancer.*

**Keywords:** Ovarian cancer, Oncogenes, Gene expression analysis, Microarray database, Support vector machine

1. **Introduction.** Ovarian cancer, one of the common gynecological cancers, is the fifth leading cause of cancer deaths in women in the western world [1] (the third leading in Taiwan). Obviously, ovarian cancer is common in industrialized country. Until now, the exact cause is still unknown. Mainly, ovarian cancer can be divided into three major pathological stages: benign, borderline and invasive stages. In the symptom of ovarian cancer, the benignancy tumor becomes to malignancy is difficult to diagnose and estimate. Moreover, the overall 5-year survival rate is about 46% in this situation [2]. Therefore, how to diagnose ovarian cancer precisely and efficiently from benign to invasive tissue is important for increasing the cure rate. Recently, lots of scientific literatures continually denote the gene expressions which are related to ovarian cancer [3-8]. Many researches of microarray analysis point out some specific genes which are detected and apparently related to cause of ovarian cancer by many biological experiments. According to the above discovered information, pathologists further detect the specific ovarian genes by analyzing the differences of genes to compare the levels of gene expression in different conditions and pathological stages. By analyzing gene expressions with biological and statistical computing could represent gene expression at each pathological stage, especially at the

---

\*The corresponding author has the same contribution as the first author.