Clinical Data Analysis Reveals Three Subytpes of Gastric Cancer

Xinxin Wang[1#], Zhana Duren[2#], Chao Zhang[3], Lin Chen[1*], Yong Wang[2*]

1 Department of General Surgery, Chinese People's Liberation Army General Hospital, Beijing 100853, China
2 National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
3 Computer Science Department, University of Missouri-Columbia, MO 65211, USA
# These authors contributed equally
* Correponding to: CHEN Lin, Email: chenlinbj@vip.sina.com, WANG Yong, E-mail:ywang@mass.ac.cn

***Abstract*—Gastric cancer is the fourth most common cancer and second leading cause of cancer-related death worldwide. Nowadays the accumulated large scale clinical data allows the clinicopathlogical review to identify the clinical factors, reveal their possible correlations, and mine the possible clinical patterns for gastric cancer. Here we analyze the clinical data of over 1500 gastric cancer patients histopathologically diagnosed and treated during 2006 to 2010. Specifically, we collect and preprocess the data by extracting 14 available clinical factors from three categories, i.e., the clinical background, immunohistochemistry data, and the caner's stage information. Then these factors are quantized and the significant factors and their correlations are calculated. Importantly, we define a distance between two patients by their clinical factors profile similarity and cluster all the patients into subgroups. We find that most of the patients fall into three major classes and we define them as three subtypes of gastric cancer. Each subtype is analyzed and characterized by its own significant factors and correlations. Our analysis may provide important insights for gastric cancer classification and diagnose.**

***Key Words*— Gastric cancer, Retrospective study, Clinical data analysis, Biomarker, Clustering**

I. INTRODUCTION

Gastric cancer, or stomach cancer, refers to tumors that develop in the lower part of the esophagus, in the stomach, or in the uppermost part of the small intestine. It is the fourth most common cancer and second leading cause of cancer-related death worldwide [1].

Thanks to the development of new technologies, clinical data can be obtained through various methods that may include, but are not limited to the clinical background, immunohistochemistry data, and the cancer's stage information. The large scale clinical information is deposited in paper, electronic medical records, or even databases [2]. It's crucial to perform clinicopathlogical review to identify the important clinical factors, reveal their possible correlations, and find the possible clinical patterns. Those valuable knowledge mined from the data will help disease diagnosis and treatment.

Traditionally, clinical data analysis was carried out by statistical methods and had remarkable successes [3-4]. Here we introduce the systems biology's viewpoint to analyze clinical data. Systems biology is a newly proposed term to describe the study of the interactions between the components of biological systems, and how these interactions give rise to the function and behavior of that system [5]. The two keywords for systems biology are network and data integration [6-7]. In this paper, we would like to introduce these two concepts into clinical data analysis. Specifically we will emphasize the relationships among patients, and the correlations among clinical factors, and the integration of various available clinical measurements.

II. METHOD AND MATERIALS

A. Collecting and preprocess the clinical data

We obtained the clinical data from the Department of Surgery of Chinese People's Liberation Army General Hospital. In total 2,752 gastric cancer patients underwent gastrectomy between January 2006 and May 2010. We retrospectively reviewed the records of those patients to extract the clinical features or factors closely related to gastric cancer. We grouped the available demographic and clinicopathological information obtained from patient records into three classes, i.e., the clinical background information, the caner's stage information, and the immunohistochemistry data.

The clinical background information contains age, gender, characteristics of primary gastric cancer, and treatment-related factors. Characteristics of the primary gastric cancer included tumor location, tumor size, macroscopic type, histological differentiation, depth of tumor invasion, lymph node metastases, and lymphovascular invasion. Treatment-related factors include resection margin (positive or negative), extent of lymph node dissection, type of reconstruction, and hepatic resection for metastatic hepatic tumors (absence or presence).

The gastric cancer's stage information was determined according to the 7th edition of the TNM classification of the American Joint Committee on Cancer (AJCC) and Union for International Cancer Control (UICC) [8]. Cancer stage is the most important prognostic factor for gastric cancer and is

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

indicated by four letters, TNM, T, N, and M. TNM is indicative to the overall tumor stage using the size and extension of the primary tumor, its lymphatic involvement, and the presence of metastases to classify the progression of cancer. There are eight stages for TNM (1a, 1b, 2a, 2b, 3a, 3b, 3c, and 4).T- refers to the size of the tumor and whether it has invaded nearby tissue (five stages as 1, 2, 3, 4a, and 4b). N- refers to whether the lymph nodes are involved (five stages as 0, 1, 2, 3a, and 3b). M- refers to whether the stomach cancer has spread to distant tissues/organs (two stages as 0 or 1).

In addition, immunohisochemical data can provide the information at molecular level. Immunohistochemistry has taken a central role in the field of pathology and is widely used in the diagnosis of abnormal cells. It provides a way to measure the protein concentrations in the tumor sample. Specifically in gastric cancer, seven proteins (HER-1, HER-2, p53, p170, Ki-67, VEGF, and p16) are measured and their concentration are detected as positive (+) or negative (-) by a community-wide threshold.

We preprocess the clinical data by excluding the patients that had a history of malignancy, underwent either laparotomy or a bypass procedure, or had substantially incomplete clinical factors. Finally, 1,525 gastric cancer patients are included in this study.

B. Categorize and quantize the clinical factors

For every gastric cancer patient, 14 clinical factors are further investigated. Three factors, age, sex, and tumor size, are clinical background information. Four factors, TNM,T, N, and M, are related to cancer stage description. The above seven factors are global phenotype features. At the molecular level, immunohisochemical measurements provide seven factors as HER-1, HER-2, p53, p170, Ki-67, VEGF, and p16.

Nine factors take binary values and can be easily quantized as 0 or 1. As a result, HER-1, HER-2, p53, p170, Ki-67, VEGF, and p16 have the values either 0 or 1 for negative and positive. Sex takes value 0, 1 for male and female respectively. M has the value 0 or 1 to denote the status for metastasis. T, N, and TNM naturally take multiple values. We use an integer to quantize it to denote the grade of the cancer. For example, integer 1-5 can denote the five grades for T. Age and tumor size have continuous values and we binned them into 5 categories and similarly use an integer to quantize it.

C. Clustering the patients

It's well known that gastric cancer is a heterogeneous disease comprising multiple subtypes that have distinct biological properties and effects in patients. Simply pooling all the samples together may fail to identify the true characters of gastric cancer. It's important to reveal possible disease subtypes as the first step.

After quantization, every patient is denoted by a vector in N dimension (N=14 in our case). Given two patients, we define a score S to measure their clinical difference by assessing the distance between vector $X=(x_1,x_2,\cdots x_N)$ and $Y=(y_1,y_2,\cdots y_N)$,

$$S = diff(X,Y) = \sum_{i=1}^{N} |x_i - y_i|$$

We then construct clusters from a hierarchical cluster tree induced from all pairwise distances of patients. Here the cluster function in MATLAB is utilized in our implementation.

III. RESULTS

A. Analyzing the data with all patients

The clinical characteristics of all the 1,525 samples are summarized in Table 1. We show the distribution of the 14 clinical factors in all patients. For example, the patients have the mean age 58.99 and about 60% patients have their ages ranging from 50 to 59.

From this table, significant factors can be easily picked out. It's clearly that Ki-67 serves as a good biomarker [9] for gastric cancer. In total 1,519 patients are positive in Ki-67 protein abundance. This fact is confirmed by further literature search. The researchers showed that tumors that have a high frequency of cells expressing Ki-67, which is associated with resistance to aromatase inhibitors, contained an elevated frequency of somatic mutations (those that arise during tumor progression, rather than being inherited) and genome-structure changes compared with tumors with a low frequency of Ki-67 positive cells. In addition, VEGF is also a pretty strong biomarker by showing 94% patients have positive value. For the cancer staging, most of the stomach cancers (1,361 in 1,525) have not spread to distant tissues/organs (M is 0).

We then calculate the significant correlations among the 14 clinical factors. Pearson correlation coefficient (PCC, and is typically denoted by r) is used to measure the correlation (linear dependence) between two patients represented by variables X and Y. It gives a value between -1 and +1. In addition, a recent novel association detecting method, the maximal information coefficient (MIC), is also used to increase the coverage by revealing possible nonlinear relationships. MIC captures a wide range of associations for functional relationships [10].

The correlation results are graphically illustrated in Figure 1. Each clinical factor is denoted by a node in the graph and an edge connects two nodes if their correlation is significant (absolute PCC is larger than 0.4 or MIC is larger than 0.16). When considering all the samples, TNM shows significant correlations with T, N, and M respectively. And

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

tumor size correlates with TNM with a PCC 0.42. These results are reasonable and can be expected.

Table 1. Baseline characteristics of all the gastric cancer patients (1,525) in our data.

All gastric cancer samples (n=1525)		
Clinical feature	**Number of patients**	**Percentage(%)**
Age (mean)	58.9948	
<40	106	6.95
40-49	206	13.51
50-59	428	28.07
60-69	472	30.95
≥70	313	20.52
Sex		
Male	1176	77.11
Female	349	22.89
HER-1		
-1	1183	77.57
1	342	22.43
HER-2		
-1	957	62.75
1	568	37.25
Ki-67		
-1	6	0.39
1	1519	99.61
p16		
-1	1033	67.74
1	492	32.26
p170		
-1	576	37.77
1	949	62.23
p53		
-1	560	36.72
1	965	63.28
VEGF		
-1	99	6.49
1	1426	93.51
T		
1	91	5.97
2	176	11.54
3	240	15.74
4a	516	33.84
4b	502	32.92
N		
0	433	28.39
1	267	17.51
2	309	20.26
3a	330	21.64
3b	186	12.20
M		
0	1361	89.25
1	164	10.75
TNM		
1a	75	4.92
1b	114	7.48
2a	119	7.80
2b	154	10.10
3a	143	9.38
3b	324	21.25
3c	432	28.33
4	164	10.75
Tumor Size		
<2	142	9.31
2—4	553	36.26
4—6	460	30.16
6—8	227	14.89
>8	163	10.69

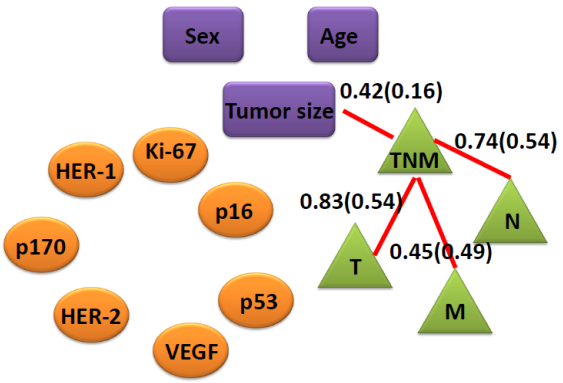

Figure 1. The clinical factors investigated and their correlations calculated by using all the patient samples. The correlations with higher Pearson correlation coefficient (absolute value larger than 0.4) are illustrated in red line. The correlation value is labeled on the edge (The first value is PCC and the second in the bracket is the MIC strength).

B. Clustering the patients into three classes

We clustered the 1,525 samples by a simple hierarchy clustering strategy. The clustering tree is shown in Figure 2. There are three major groups, which indicate three possible subtypes of gastric cancer. Thus we name the three groups of gastric cancer as Subtype I, Subtype II, and subtype III (as shown in different colors in Figure 2). Each group has 762, 187, and 176 samples respectively. We also tried different clustering methods and different ways to define similarity of patients, the results remain largely the same. Next we will go deep into each subtype to identify significant factors and associations to further characterize the subtype's clinical patterns.

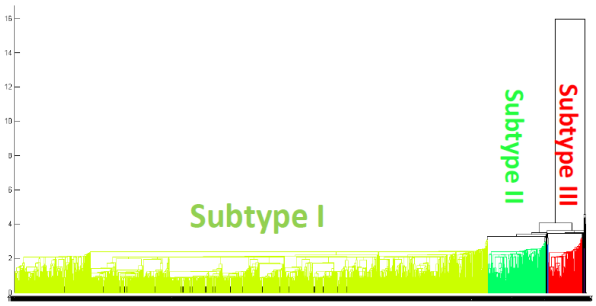

Figure 2. The 1,525 samples are hierarchically clustered. The three major branches are colored in green, blue, and red respectively. We define them as Subtype I, Subtype II, and Subtype III.

C. Analyzing Subtype I gastric cancer

The clinical characteristics of all the 762 patients are summarized in Table 2. Interestingly, two factors have distinct subtype-specific patterns. All the Subtype I patients are male and they all have negative protein expression of biomarker Her-1. Her-1 (EGFR or ErbB-1) is known as a member of the epidermal growth factor receptor tyrosine kinases subfamily, and plays a crucial role in signaling pathway in the regulation of cell proliferation, survival and differentiation. Expression of Her-1 and Her-2 is thought to be a prognostic factor and target of novel biologic agents [11]. Other factors possess similar distribution with the case when considering all the patients.

We further check the correlations among the clinical factors in Subtype I gastric cancer patients. The results in Figure 3 demonstrate that there are clear correlations among sex, Ki-67, VEGF, M, and HER-1. In particular, Her-1 is negatively correlates with sex. In summary, Subtype I is characterized by female, negative Her-1 protein expression and their strong negative correlation.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Table 2. Baseline characteristics of Subtype I gastric cancer patients (762) in our data.

Subtype I gastric cancer (n=762)		
Clinical feature	**Number of patients**	**Percentage(%)**
Age (mean)	60.4244	
<40	27	3.54
40-49	89	11.68
50-59	234	30.71
60-69	251	32.94
≥70	161	21.13
Sex		
Male	762	100.00
Female	0	0.00
HER-1		
-1	762	100.00
1	0	0.00
HER-2		
-1	496	65.09
1	266	34.91
Ki-67		
-1	0	0.00
1	762	100.00
p16		
-1	516	67.72
1	246	32.28
p170		
-1	280	36.75
1	482	63.25
p53		
-1	286	37.53
1	476	62.47
VEGF		
-1	0	0.00
1	762	100.00
T		
1	41	5.38
2	87	11.42
3	117	15.35
4a	269	35.30
4b	248	32.55
N		
0	234	30.71
1	150	19.69
2	159	20.87
3a	161	21.13
3b	58	7.61
M		
0	762	100.00
1	0	0.00
TNM		
1a	33	4.33
1b	62	8.14
2a	54	7.09
2b	83	10.89
3a	90	11.81
3b	197	25.85
3c	243	31.89
4	0	0.00
Tumor Size		
<2	71	9.32
2—4	282	37.01
4—6	224	29.40
6—8	118	15.49
>8	67	8.79

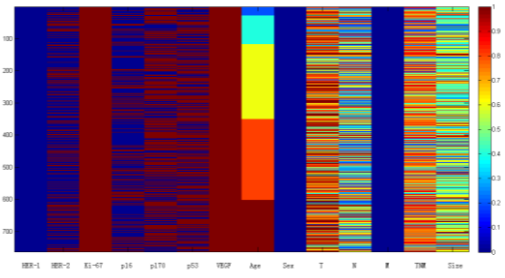

Figure 3. The heatmap to illustrate the clinical factor distribution of Subtype I gastric cancer.

D. Analyzing Subtype II gastric cancer

The clinical characteristics of all the 187 patients are summarized in Table 3. Again, two factors have distinct subtype-specific patterns. All the Subtype II patients are female and they all have negative protein expression of biomarker Her-1. Other factors possess similar distribution with the case when considering all the patients.

We further check the correlations among the clinical factors in the Subtype II. The results in Figure 4 demonstrate that there are clear correlations among sex, Ki-67, VEGF, M, and HER-1. In particular, Her-1 is positively correlates with sex. In summary, Subtype II is characterized by female, negative Her-1 protein expression and their strong positive correlation.

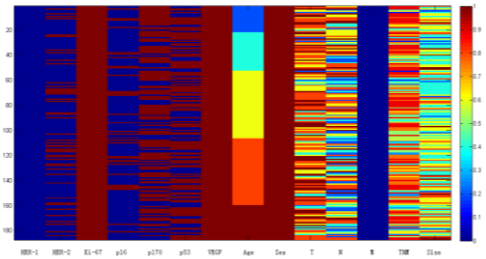

Figure 4. The heatmap of clinical factor distribution for Subtype II gastric cancer.

Table 3. Baseline characteristics of Subtype II gastric cancer patients (187) in our data.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Subtype II gastric cancer (n=187)		
Clinical feature	**Number of patients**	**Percentage(%)**
Age (mean)	56.2835	
<40	21	11.23
40-49	31	16.58
50-59	54	28.88
60-69	53	28.34
≥70	28	14.97
Sex		
Male	0	0.00
Female	187	100.00
HER-1		
-1	187	100.00
1	0	0.00
HER-2		
-1	140	74.87
1	47	25.13
Ki-67		
-1	0	0.00
1	187	100.00
p16		
-1	154	82.35
1	33	17.65
p170		
-1	66	35.29
1	121	64.71
p53		
-1	73	39.04
1	114	60.96
VEGF		
-1	0	0.00
1	187	100.00
T		
1	6	3.21
2	18	9.63
3	38	20.32
4a	74	39.57
4b	51	27.27
N		
0	46	24.60
1	30	16.04
2	42	22.46
3a	46	24.60
3b	23	12.30
M		
0	187	100.00
1	0	0.00
TNM		
1a	4	2.14
1b	10	5.35
2a	17	9.09
2b	25	13.37
3a	15	8.02
3b	47	25.13
3c	69	36.90
4	0	0.00
Tumor Size		
<2	11	5.88
2—4	74	39.57
4—6	61	32.62
6—8	30	16.04
>8	11	5.88

E. Analyzing Subtype III gastric cancer

The clinical characteristics of all the176 patients are summarized in Table 4. Again, gender and Her-1 have distinct subtype-specific patterns. All the Subtype III patients are female and they all have negative protein expression of biomarker Her-1. Other factors possess similar distribution with the case when considering all the patients.

We further check the correlations among the clinical factors in the Subtype III. The results in Figure 5 demonstrate that there are clear correlations among sex, Ki-67, VEGF, M, and HER-1. In particular, Her-1 is negatively correlates with sex. In summary, Subtype III is characterized by female, positive Her-1 protein expression and their strong negative correlation.

Table 4. Baseline characteristics of Subtype III gastric cancer patients (176) in our data.

Subtype III gastric cancer (n=176)		
Clinical feature	**Number of patients**	**Percentage(%)**
Age (mean)	61.3	
<40	3	1.70
40-49	23	13.07
50-59	43	24.43
60-69	63	35.80
≥70	44	25.00
Sex		
Male	176	100.00
Female	0	0.00
HER-1		
-1	0	0.00
1	176	100.00
HER-2		
	80	45.45
	96	54.55
Ki-67		
-1	0	0.00
1	176	100.00
p16		
-1	88	50.00
1	88	50.00
p170		
-1	56	31.82
1	120	68.18
p53		
-1	18	10.23
1	158	89.77
VEGF		
-1	0	0.00
1	176	100.00
T		
1	15	8.52
2	22	12.50
3	32	18.18
4a	50	28.41
4b	57	32.39
N		
0	56	31.82
1	35	19.89
2	42	23.86
3a	24	13.64
3b	19	10.80
M		
0	176	100.00
1	0	0.00
TNM		
1a	12	6.82
1b	15	8.52
2a	20	11.36
2b	20	11.36
3a	15	8.52
3b	38	21.59
3c	56	31.82
4	0	0.00
Tumor Size		
<2	24	13.64
2—4	62	35.23
4—6	59	33.52
6—8	17	9.66
>8	14	7.95

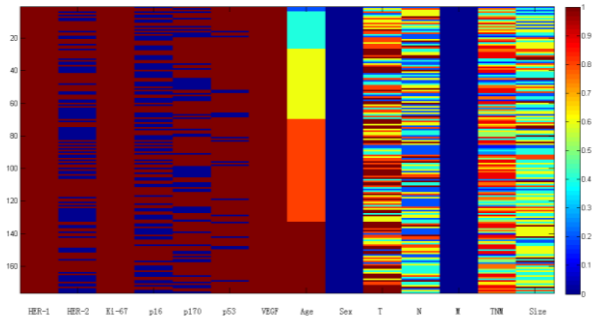

Figure 5. The heatmap of clinical factor distribution for Subtype III gastric cancer.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

IV. DISCUSSIONS AND CONCLUSIONS

In this paper we analyze the clinical data of over 1,500 gastric cancer patients. For every patient, 14 clinical factors are investigated varying from clinical background information, cancer stage diagnose, and protein expression from immunohistochemistry measurement. When calculating using all the samples, we identify several factors and reveal some interesting correlations among these factors for gastric cancer. We then cluster all the patients by defining a score of patient pairs by their clinical factor profiles similarity. Three major groups are identified. Surprisingly we find that two clinical factors, Her-1 and gender, can clearly characterize and differentiate these three groups. Thus we call the revealed groups as subtypes of gastric cancer.

It's interesting to find that gender is an important factor for gastric cancer subtype characterization. Actually several types of cancer, including stomach, liver, and colon, are far more common in men than in women. Some scientists have hypothesized that differences in lifestyle, such as diet and smoking, may account for the role of gender factor. On the other hand, growing evidence also suggests that the differences are rooted in basic biological differences between men and women. For example, recent research indicates that estrogen protects against gastric cancer [11]. This fact provides evidence for the rationality of our subtype definition by gender and Her-1. Importantly, our subtype definition for gastric cancer could help scientists find better drug targets against the disease.

Gastric cancer is a heterogeneous disease comprising multiple subtypes that have distinct biological properties and effects in patients. Here, we systematically analyze the clinical data to identify new, intrinsic subtypes of gastric cancer. These subtypes need further validation. One way is to analyze high-throughput gene expression data or multiple layer data integration to support our findings [12,13,14]. Another way is to test if these subtypes might be associated with differences in patient survival times and responses to various standard-of-carecytotoxic drugs.

One advantage of our analysis is that we provide a new way to integrate heterogeneous data. Clinical factors are form different categories and have different meanings from clinical diagnosis to molecular level. By identifying significant correlations for the pairwise factors, we connect those separate factors with each other to find interesting patterns. For example, the combination of patient background information (gender) and molecular biomarker (Her-1 expression) can distinguish three subtypes. This data integration strategy is hopefully to reveal gastric cancer biomarkers with high accuracy.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (NSFC) under Grant 61171007 and 11131009. This research was also supported in part by the Project of Knowledge Innovation Program (PKIP) of Chinese Academy of Sciences, Grant No. KJCX2.YW.W10

REFERENCES

[1] Thun MJ, DeLancey JO, Center MM, Jemal A, Ward EM. The global burden of cancer: priorities for prevention. Carcinogenesis 2010; 31: 100-110.

[2] Spilker B.L. Schoenfelder J. Data Collection Forms in Clinical Trials Raven Press, New York.1991.

[3] Wu, CY and Kuo, KN and Wu, MS and Chen, YJ and Wang, CB and Lin, JT. Early Helicobacter pylori eradication decreases risk of gastric cancer in patients with peptic ulcer disease, European Journal of Gastroenterology & Hepatology, 22, 1, 125--128, 2010.

[4] Zhou, Y. and Zhuang, W. and Hu, W. and Liu, G.J. and Wu, T.X. and Wu, X.T. Consumption of large amounts of Allium vegetables reduces risk for gastric cancer in a meta-analysis, Gastroenterology, 141:80-89, 2011.

[5] Chen L, Wang R-S, Zhang X-S: Biomolecular Networks: Methods and Applications in Systems Biology. John Wiley & Sons, Hoboken, New Jersey; 2009.

[6] Y. Wang, X. S. Zhang and L. Chen. A network biology study on circadian rhythm by integrating various omics data. OMICS: A Journal of Integrative Biology, 2009, 13(4).

[7] Y. Wang, X.-S. Zhang, and Y. Xia. Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. Nucleic Acids Research, 2009, 37(18):5943--5958.

[8] Kay Washington, 7-th edition of AJCC cancer staging manual: stomach, Annals of Surgical Oncology, 17(6), 1471-1474, 2010.

[9] Yong Wang, Qiao-Feng Wu, Chen Chen, Ling-Yun Wu, Xian-Zhong Yan, Shu-Guang Yu, Xiang-Sun Zhang, and Fan-Rong Liang. Revealing metabolite biomarkers for acupuncture treatment by linear programming based feature selection. BMC Systems Biology, 6(Suppl 1):S15, 2012.

[10] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting Novel Associations in Large Data Sets, Science, 1518-1524, 2011.

[11] Sheh A et al. 17(lowercase beta)-estradiol and Tamoxifen prevent gastric cancer by modulating leukocyte recruitment and oncogenic pathways in Helicobacter pylori -infected INS-GAS male mice, Cancer prevention research, 2011.

[12] Jørgensen JT, Hersom M. HER2 as a Prognostic Marker in Gastric Cancer-A Systematic Analysis of Data from the Literature. J Cancer 2012; 3:137-144.

[13] Joe Gray, Brian Druker. Genomics: the breast cancer landscape. Nature, 486, 328–329, 2012

[14] Iain Beehuat Tan, Tatiana Ivanova, Kiathon Lim, et al, Intrinsic Subtypes of Gastric Cancer, Based on Gene Expression Pattern, Predict Survival and Respond Differently to Chemotherapy, Gastroenterlogy, 2011;141:476-485.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE