Gene Discovery from Hepatoma Microarrays

Chaur-Chin Chen [*]
Angiogenesis Research Center
National Taiwan University
Taipei, Taiwan 100
E-mail: cchen@cs.nthu.edu.tw
Tel/Fax: +886 3 573 1078 / +886 2 2362 8167
http://angiogenesis.mc.ntu.edu.tw

September 3, 2004

Abstract

Microarray images used to study gene expression in cancer diseases has recently attracted a variety of researchers including medical doctors, computational biologists, and bioinformaticians. A set of microarray images were acquired by a sequence of biological experiments which were scanned via a high resolution scanner. For each spot corresponding to a gene, the ratio of Cy3 and Cy5 fluorescent signal intensities was obtained and which may be normalied based on piecewise linear regression such as lowess proposed by Terry Speed. In this study, we have collected, from 44 patients of Hepatoma, 44 microarray images based on which an $M \times N$ *genematrix,* A*, with* $N = 44$ *patients and* $M = 13574$ *effected genes in each microarray. We start with our gene discovery from a genematrix* $A \in R^{M \times N}$, $M = 13574$, $N = 44$ *formed from 44 microarray data sets including* $N_1 = 12$ *patients of hepatitis C virus (HCV),* $N_2 = 27$ *patients of hepatitis B virus (HBV), 1 patient clinically diagnosed to be infected with HCV as well as HBV, and 4 patients were infected with neither HCV nor HBV. There are 13574 spot features computed from the effected genes in each microarray which were provided by a local company Welgene, Inc. in Nankang, Taipei city. Three problems under investigation are listed as follows.*

(1) *Detect the differentially expressed, either up-regulated or down-regulated genes among 13574 genes.*

(2) *Select a subset of genes among 13574 genes which "best" distinguishes HCV patients from HBV ones from 39 patients.*

(3) *Select a subset of genes from 13574 genes among 44 patients which "best" distinguishes the patients with vascular invasion and those without vascular invasion.*

Keywords: *Clustering, DNA, Dsicrimination, Fisher, Microarray*

[*] This work was supported in part by 92-EC-17-A-19-S1-0016

1 Introduction

Microarray images used to study gene expression in cancer diseases has recently attracted a variety of researchers including medical doctors, computational biologists, and bioinformaticians. A set of microarray images were acquired by a sequence of biological experiments which were scanned via a high resolution scanner. For each spot corresponding to a gene, the ratio of Cy3 and Cy5 fluorescent signal intensities was obtained and which may be normalied based on piecewise linear regression such as lowess proposed by Terry Speed. In this study, we have collected, from 44 patients of Hepatoma, 44 microarray images based on which an $M \times N$ genematrix, A, with $N = 44$ patients and $M = 13574$ effected genes in each microarray. We start with our gene discovery from a genematrix $A \in R^{M \times N}$, $M = 13574$, $N = 44$ formed from 44 microarray data sets including $N_1 = 12$ patients of hepatitis C virus (HCV), $N_2 = 27$ patients of hepatitis B virus (HBV), 1 patient clinically diagnosed to be infected with HCV as well as HBV, and 4 patients were infected with neither HCV nor HBV. There are 13574 spot features computed from the effected genes in each microarray which were provided by a local company Welgene, Inc. in Nankang, Taipei city. Two problems under investigation are listed as follows.

(1) Detect the differentially expressed, either up-regulated or down-regulated genes among 13574 genes.

(2) Select a subset of genes among 13574 genes which "best" distinguishes HCV patients from HBV ones.

(3) Select a subset of genes from 13574 genes among 44 patients which "best" distinguishes the patients with vascular invasion and those without vascular invasion.

2 Most Differentially Expressed Genes

Let A[1:13574, 1:44] be the genematrix with each entry being the lowess normalized $Cy3/Cy5$ *ratio*, where the tumor tissues dyed with the fluorescence light of wavelength 532 nanometers and the normal tissues dyed with the fluorescence light of wavelength 635 nanometers.

A gene k is said to be up regulated if

$$log_2(A[k,j]) \;\geq\; T \quad for\ 1 \leq j \leq N = 44$$

A gene k is said to be down regulated if

$$log_2(A[k,j]) \;\leq\; -T \quad for\ 1 \leq j \leq N = 44$$

The following genes are detected as *differentially expressed* when the threshold $T = \mathbf{2.0}$ is chosen.

3 Discriminative Genes to Distinguish HCV from HBV

Let $X[1:K, 1:N]$ be derived from $A[1:13574, 1:44]$ with $N = N_1 + N_2 = 12 + 27 = 39$ patients including $N_1 = 12$ HCV patients and $N_2 = 27$ HBV patients with K genes being selected to distinguish HCV from HBV. The selection of K genes are based on the condition $C_k > T_c$ for each gene k, where the Fisher's ratio for gene k C_k is defined below: the larger, the more separable.

$$C_k = (\mu_1(k) - \mu_2(k))^2/(p_1 s_1^2(k) + p_2 s_2^2(k)),$$

$$where\ p_1 = N_1/N, \quad p_2 = N_2/N$$

$$\mu_1(k) = \frac{1}{N_1}\sum_{j=1}^{N_1} X[k,j]$$

$$\mu_2(k) = \frac{1}{N_2}\sum_{j=12}^{N} X[k,j]$$

$$s_1^2(k) = \frac{1}{N_1}\sum_{j=1}^{N_1}(X[k,j] - \mu_1(k))^2$$

$$s_2^2(k) = \frac{1}{N_2}\sum_{j=12}^{N_2}(X[k,j] - \mu_2(k))^2$$

The following 32 genes are detected when $T_c = \mathbf{1.7}$.

The dendrogram of complete-linkage based on the most 32 discriminative genes for distinguishing HCV from HBV patients is given below.

The following 31 genes associated with the accession numbers from Genbank by using the threshold of Fisher ratio $T_c = 0.95$ to distinguish a patient with vascular invasion from the one with non-vascular invasion are detected.

The dendrogram of complete-linkage based on the most 31 discriminative genes. for distinguishing patients with vascular invasion from those without vascular invasion is given below.

Figure 2. Dendrogram of 15 VI and 29 Non-VI patients with 31 Genes.

Index	Feature#	Accession#
↑ 7574	7559	AI133162
↑ 3268	3129	M12654
4820	4697	X01098
3175	3222	M13149
13844	13769	K02922
230	239	M21692
4863	4966	L32179
1298	1355	AL532086
14121	14116	AF152562
8878	8751	BG573805
5377	5388	BC008983
3237	3160	D29832
10683	10690	BI834172
13351	13326	AL119276
10739	10634	BG567504
2711	2750	BF663177
10605	10456	BF795929
3129	3268	X63652
5488	5589	AL531502
14581	14592	X03069
13460	13529	AW609791
1673	1604	AK026409
5578	5499	0
7362	7459	BG685150
6566	6695	AF135157
4729	4788	AF123050
5997	6016	M27487
11104	11205	J04080
1387	1266	AW270961
560	533	AF350254
4736	4781	D00096
3351	3358	BC004143
3143	3254	J00129
3761	3884	X13334
↑ 1706	1571	U44799
↑ 16210	16083	AJ002304
↑ 6378	6259	M13560
↓ 180	289	M24173
↓ 820	897	BG621010

Table 1: **Up(Down)-Regulated Genes:** $log_2(Normalized\ Ratio) \ \geq \ 2 \ \ (\leq -2)$

.

Index	Feature#	Accession#
7197	7312	BG259957
9443	9434	CAC51145
2918	2855	BI520001
11189	11120	AB008549
11087	11222	BC006496
13796	13817	U35376
13433	13556	AF126404
8495	8510	AJ012159
10965	11032	AAF36120
9546	9643	X52125
9587	9602	M13232
11214	11095	AK027210
10509	10552	AL575644
580	513	AL133645
10052	10073	AB011542
1001	1028	AI678859
587	506	AF386492
4164	4105	NM_000423
5434	5331	AB050785
113	44	Y16961
8140	8241	AK026068
11017	10980	BC002771
4885	4944	AK021818
10162	10275	BG830088
2738	2723	BI094014
8522	8483	AB012174
14269	14280	BC002456
16496	16421	Y00083
7353	7468	AF070641
679	726	CAA38920
5489	5588	D88152
10506	10555	AL565681

Table 2: **32 most discriminative genes for distinguishing HCV from HBV patients**

.

Index	Feature#	Accession#
13252	13113	AY009108
8096	7973	AB020678
9924	9889	BC002536
6579	6682	Y14747
4824	4693	AB037886
2644	2505	AA446039
5562	5515	BE560878
896	821	AAA40477
15192	15229	AL532220
16481	16436	AF024636
7068	7129	BC000987
6379	6258	AJ276162
8904	9037	NM_001830
7236	7273	BC007005
2796	2665	AF020089
16382	16535	AU138067
4302	4279	AF051151
8266	8115	M98262
8476	8529	J02625
9405	9472	AB026730
13659	13642	BI760179
2169	2044	S76773
570	523	BC003000
15099	15010	BG491617
2820	2953	Z30425
10574	10487	BG757994
6675	6586	BC007665
5017	5124	L15309
8780	8849	AL529007
4495	4398	BG831940
4332	4249	BG765209

Table 3: **31 most discriminative genes for distinguishing VI from NVI patients**

.

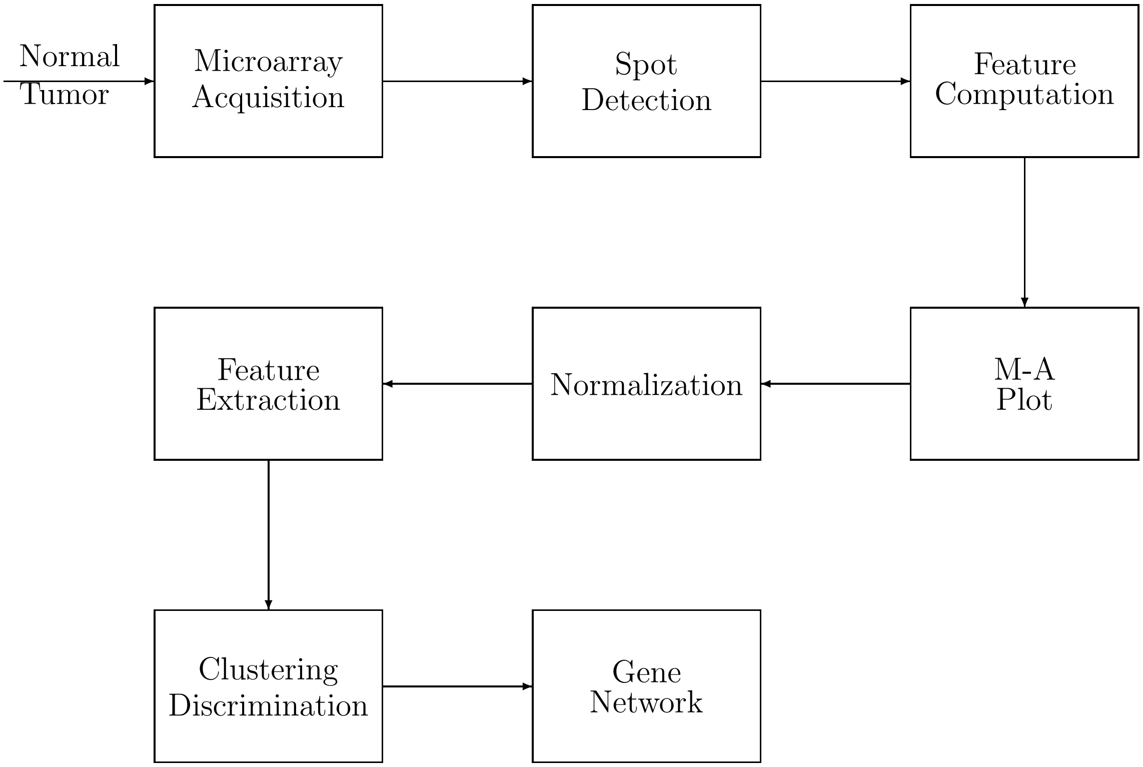

Figure 2. A Paradigm of Microarray Image Pattern Analysis.