

Mining Correlations of Human Gene Expression from Digital Gene Expression Profiles

JORNG-TZONG HORNG^{1,2,*}, HSIEN-DA HUANG², KUO-YEN TSENG²
TSUNG-SHAN TSOU³, BAW-JHIUNE LIU⁴ AND CHENG-YAN KAO⁵

¹*Department of Life Science*

²*Department of Computer Science and Information Engineering*

³*Institute of Statistics*

National Central University

Chungli, 320 Taiwan

**E-mail: horng@db.csie.ncu.edu.tw*

⁴*Department of Computer Science and Engineering*

Yuan-Ze University

Chungli, 320 Taiwan

⁵*Department of Computer Science and Information Engineering*

National Taiwan University

Taipei, 106 Taiwan

The study addressed here aimed to analyze a large number of human genome transcripts from diverse tissues and to discover genes that with similar expression profiles in different human tissues. These genes may be of potential biological or pharmaceutical relevance. We propose an approach to discover the correlations of tissue gene expression by analyzing digital gene expression profiles of different human tissues. A simple statistical test was used to correlate genes having similar expression profiles. We used the information of tissue gene expression to discover the correlations of expressed genes. The correlations of gene expression revealed that such genes were specifically expressed in particular tissues with similar expression profiles and could be used to identify the relationships of the genes that be co-regulated, involved in the same biochemical pathway and signal transduction process.

Keywords: gene expression, data mining, EST, SAGE, UniGene

1. INTRODUCTION

A large number of expressed sequences from diverse tissues are deposited in the form of ESTs in dbEST [1]. The presence in a tissue-specific library of an EST attributed to a certain gene implies that the given gene is expressed in that specific tissue (or set of tissues). This reasoning can be not only used to calculate general expression profiles, but also to identify genes that are specifically expressed in some particular tissues or organs.

The sequences in GenBank including ESTs are partitioned into a non-redundant set of gene-oriented clusters and are stored in UniGene database [2, 3]. Each UniGene cluster stands for a unique gene, has numerous sequences from different EST libraries and also contains much information such as the tissue types of the gene expressed, organism

Received November 1, 2002; accepted June 5, 2003.

Communicated by Chuen-Tsai Sun.

protein similarities modeling, and the LocusLink identifier. These EST sequences have also been used for gene mapping projects and large-scale gene expression analysis [4, 5]. UniGene database now contains thirteen organisms. For *Homo sapiens*, there are nearly 3 millions ESTs which are partitioned into 96,105 UniGene clusters (UniGene build #147).

The Cancer Genome Anatomy Project (CGAP) [6] data of the National Cancer Institute has thousands of expressed sequences, both known and novel, in the form of ESTs. These ESTs, derived from diverse normal and tumor cDNA libraries, offer valuable information for analyzing tissue gene expression. One of the uses of cDNA libraries is to identify genes whose expression differs between the issue sources of the libraries. Such genes may be of potential biological or pharmaceutical relevance.

Many techniques are now developed to identify differentially expressed genes. Thus, this type of data is becoming more widely available. There are a growing number of cDNA library databases available both commercially and in the public domain. Pertinent researches use the frequency of a gene in a cDNA library as a measure of its tissue-specific expression. One approach is Serial Analysis of Gene Expression (SAGE) [7, 8]. The approach relies on high-throughput sequencing of 10-bp gene-specific sequence tags to enumerate the expression of individual genes in a cell. A different approach uses EST counts to infer the relative level of expression of a gene. Both methods, having their own advantages and limitations, can identify both known and novel genes differentially expressed between different tissues.

Generally speaking, for unbiased randomly selected sequencing techniques, two randomly selected genes usually have different expression ratios in different libraries. If two genes have similar expression ratios in many different libraries, this may reveal some useful information. Those two genes may be implicated in similar pathways or coding for different subunits of multi-component enzyme complexes.

In this study, we applied Chi-square test to examine whether the group of genes had regular expression profiles in numerous tissues or not. The statistically correlated genes could potentially be co-regulated or involved in the same biochemical pathway and signal transduction process.

2. APPROACH

The proposed approach used in this study is illustrated in Fig. 1. Two kinds of genome-wide gene expression profiles were retrieved from UniGene and SAGE databases. Well-classified cDNA library catalogs were obtained from CGAP. Each SAGE tag could map to corresponding UniGene clusters using tag-to-gene mapping information. The Chi-square test was then applied to mine the correlations of genes. Finally, the mined statistically significant correlated genes were stored in a database. Furthermore, human genes could also be related to homologous mouse genes using human-mouse homology mapping. In addition, we designed a public web site where users can see the results we have mined.

We have mainly applied our mining approach to *Homo sapiens*. Two kinds of genome-wide gene expression profiles were obtained, respectively, from UniGene and SAGE databases. Furthermore, we obtained mouse gene expression profiles from UniGene database as additional information.

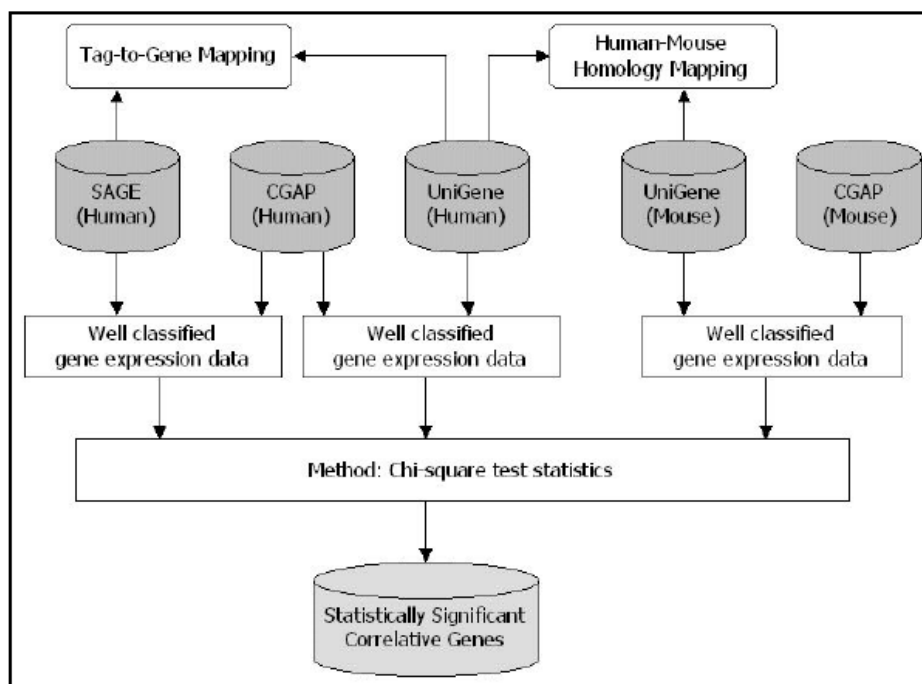


Fig. 1. The framework of the proposed approach.

In the UniGene human data, nearly three millions of ESTs belong to 96,105 gene clusters. UniGene database also collects EST libraries from dbEST. In the Cancer Genome Anatomy Project (CGAP) database, the tissue types of EST libraries are well classified. The CGAP uses 51 organs and five histologies to classify EST libraries from dbEST. By combining UniGene EST libraries and CGAP well-classified EST libraries using library names, 6,617 libraries could be classified according to corresponding tissue types. Those 6,617 EST libraries were assigned to 96,105 UniGene clusters in the UniGene database. 2,941,326 ESTs obtained from 6,617 EST libraries were found. We further classified 6,617 UniGene EST libraries into 107 tissue types.

ESTs belonging to a gene cluster maybe may exist in different tissue types. We used the UniGene database to compare the number of times ESTs from different libraries were assigned to a particular UniGene cluster. Fig. 2 shows the distribution of expressed genes and tissues. For example, we find that 14,140 UniGene clusters are expressed in only one tissue. That is, many genes are only expressed in one tissue even though we find that each gene is, on average, expressed in 7.86 tissues. Few genes are expressed in more than thirty tissues. These genes are perhaps house-keeping genes because they are expressed in many tissues. We also investigated the distribution of tissues and corresponding ESTs belonging to them. Fig. 3 shows the distribution of number of tissue size. For example, twenty tissues have less than one thousand ESTs.

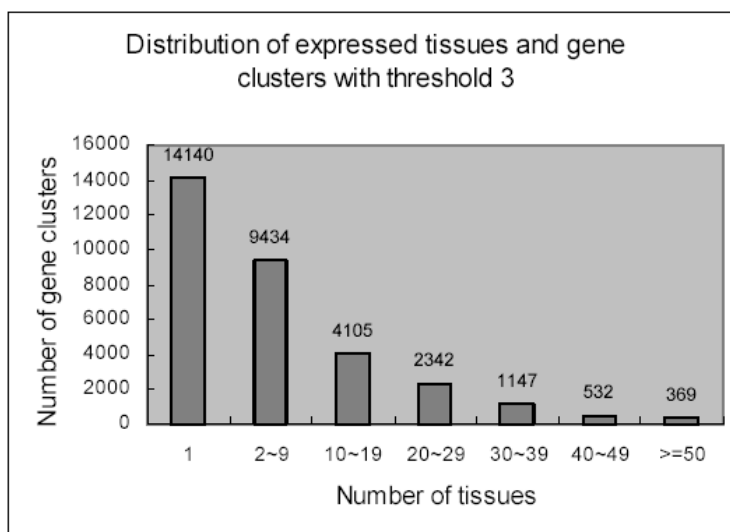


Fig. 2. Distribution of expressed genes and tissues which have at least three ESTs.

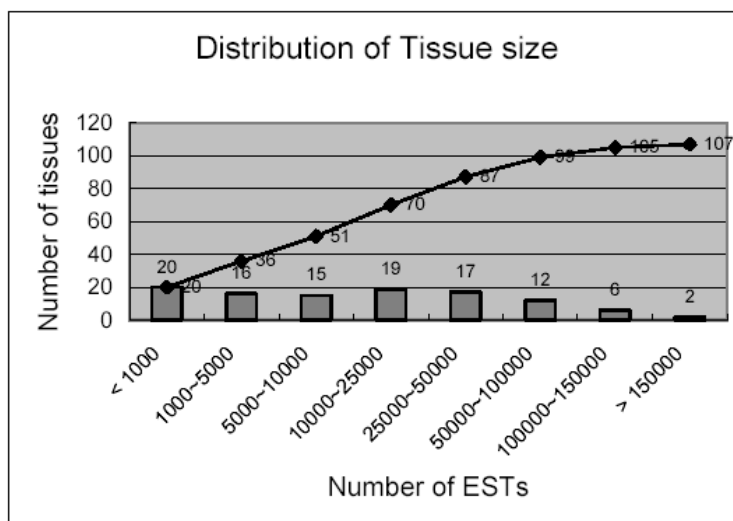


Fig. 3. Distribution of tissue size.

Besides the UniGene, human gene expression data were also retrieved from the SAGE database. As mentioned above, SAGE uses a short nucleotide sequence called a tag to represent the transcription products of a gene. Each SAGE library contains thousands of tags from various genes. We could use tag-to-gene mapping to map each tag to corresponding UniGene clusters. Unfortunately, not all the tags mapped to only one UniGene cluster. Some tags are mapped to more than two UniGene clusters. This is a disadvantage of SAGE. Table 1 gives some statistics concerning tag-to-gene mapping.

There are 274,985 tag-cluster pairs obtained from all the UniGene clusters and 219,538 tags are unique. This means that those 219,538 tags appear in only one UniGene cluster. If a tag from some SAGE library is the same as one of those 219,538 unique tags, then this tag must map to a specific corresponding UniGene cluster. Other tags which map to more than one UniGene cluster may cause some confusion. We do not know from which UniGene cluster is the tag comes. In Table 1, it can be seen that about 80% of the tags can be uniquely mapped to only one specific UniGene cluster. Less than 20% of tags mapped to more than one UniGene cluster. In this study, we filtered those tags which map to ambiguous UniGene clusters.

Table 1. The percent distribution of tage-to-gene mapping.

UniGene clusters (274985 entries)	Percentage of tags	Amount of tags
One clusters	80.22	176,113
Two clusters	15.89	34,885
Three clusters	3.09	6,784
Four clusters	0.61	1,339
More than four clusters	0.19	417
Total	100	219,538

After using tag-to-gene mapping to map all the tags to corresponding UniGene clusters, 2,290,846 times tags obtained from 108 human SAGE libraries were partitioned into 28 tissue types. The processed gene expression profiles were the same as the processed UniGene data, and those expression profiles could also be used for mine correlations.

Fig. 4 shows the distribution of the number of expressed genes of tissues for SAGE and UniGene. We can see that although there are only 28 tissues with SAGE data, most of the tissues contain more than 6,000 distinct genes. But in the UniGene data, many tissues have less than 4,000 genes. There are two reasons for this difference. First, the sample sizes between SAGE and UniGene data are very different. Most of the sample sizes of SAGE tissues are larger than those of UniGene tissues. Second, the SAGE method may detect more low expression level genes than the EST method.

Besides two kinds of human gene expression data, we also used mouse data in our study. We used 2,231,805 mouse ESTs from 658 mouse EST libraries belonging to 83,553 mouse UniGene clusters, and those ESTs could be classified into 53 different tissue types using CGAP well-classified EST library information. The correlation mining method we used in this study can also be used to mine correlated genes in such data. But in our study, the mouse data was mainly used for human-mouse homologous mapping to provide additional information.

In this study, we used the Chi-square test to mine the correlations of genes. The Chi-square test (X^2) is extensively applied for testing independence and correlation. It is based on comparing observed frequencies with the corresponding expected frequencies. The closer the observed frequencies are to expected frequencies, the greater is the weight of evidence in favor of independence. If the Chi-square value higher than a certain

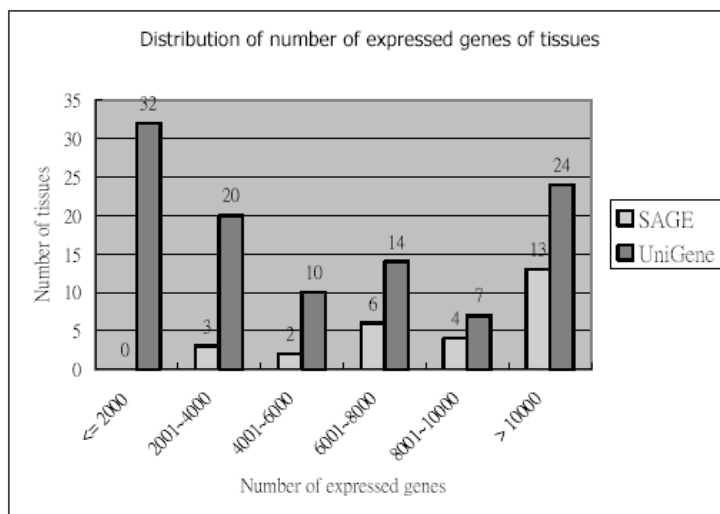


Fig. 4. Distribution of the number of expressed genes of tissues for SAGE and UniGene.

threshold, we say that it is statistically significant. If a group of genes with similar expression profiles in some different tissues, the Chi-square value will smaller than a certain threshold value. In other words, it means that no gene within this group is specifically expressed in some tissues. Those genes are statistically correlated.

The Chi-square test considers the expression levels of a gene in tissues. We mapped gene expression data to a contingency table. One gene could be expressed in various tissues with different expression levels, respectively. Fig. 5 presents two examples, where the Chi-square test was used to detect whether a group of genes had similar expression profiles. G1, G2, and G3 are a group of genes which expressed in tissue T1, T2, and T3, respectively. Each gene may have different expression counts in different tissues. In Fig. 5 (A), we can see that gene G2 is more highly expressed in tissue T1 than the other two, gene G1 is more highly expressed in tissue T2, and gene G3 is more highly expressed in tissues T1 and T3, respectively. The Chi-square value of this sample is 484.014. The hypothesis is that all the genes G1, G2, and G3, for example, are independent on their expressions in tissues. If rejecting the hypothesis, the chi-square value exceeds the value of 9.4877 in the example with 4 degrees of freedom and 95% confidence. Note that those genes are expressed dependently in the considered tissues. We can say that those genes are statistically independent. For this reason, we are interested in the group genes with test value is smaller than defined threshold. They are statistically correlated. An example with similar expression ratios is shown in Fig. 5 (B). Genes G1, G2, and G3 are independently expressed in three tissues T1, T2, and T3, so the Chi-square value is very small. (In both Fig. 5 (A) and (B), the significance threshold is 9.4877 with 4 degrees of freedom.)

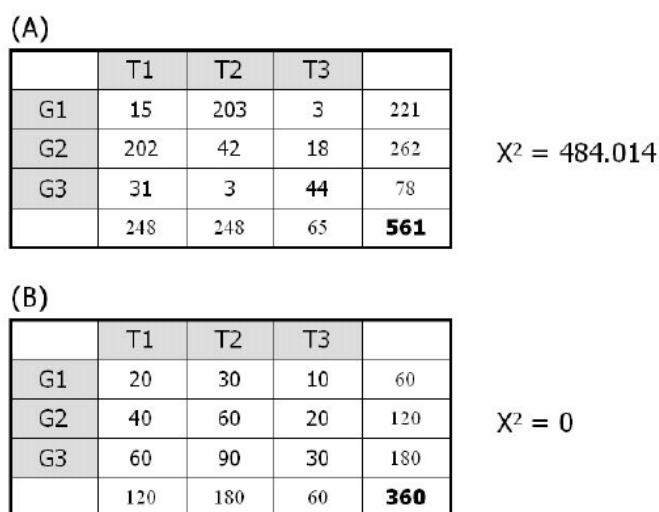


Fig. 5. Illustrative examples where the Chi-square test was used.

3. RESULTS

The Chi-square test could be applied to both kinds of human gene expression profiles (UniGene and SAGE) and also to the mouse data as homologous references. We will use SAGE data as an example. There are 28 kinds of tissues selected as experimental data. In this sample, we select eight tissues as experiment data. The eight selected experiment tissues are brain_cancer, brain_normal, colon_cancer, mammary gland_cancer, mammary gland_normal, ovary_cancer, pancreas_cancer, and prostate_cancer. The sampling sizes of tags in these experiment tissues are all at least 100,000.

Group of genes whose Chi-square values exceed significance thresholds are said the correlations of gene expressions. A correlation could have two, three, or more genes. The genes within a correlation are statistically correlated may have similar expression profiles in some experimental tissue types.

We selected some genes related to tumor, cancer, and transcription factors, and used SAGE data as the experimental gene sets. There were 225 “transcription factor” related genes and 159 “tumor or cancer” related genes obtained in the set. The gene expression information of the data set given in Table 2 was used to discover the correlations of similarly expressed genes. In this instance, all the gene data and experimental tissues were normalized before testing. We also used some detailed parameters for our program, such as minimum tag counts, to filter low abundant genes.

There were 751 correlations mined by our program from the above gene set. Partial results are given in Table 3 with Chi-square values. For instance, in correlation 1, the UniGene cluster Hs.13063 of “transcription elongation regulator 1 (CA150)” is correlated with the UniGene cluster Hs.78869 of “transcription elongation factor A (SII), 1” with a Chi-square value of 0.9507. (In this example, the significance threshold is 14.0617 with 7 degrees of freedom). The two UniGene clusters could be expressed simultaneously because they were statistically correlated in their gene expression in the

Table 2. UniGene clusters related to “tumor or cancer” and “transcription factor” (partially).

Description	Hs.#	Br_c	Br_n	Co_c	Mg_c	Mg_n	Ov_c	Pa_c	Pr_c
“tumor” or “cancer” related genes (25 of 159 genes)									
Tumor necrosis factor receptor superfamily, member 1A	159	36	3	8	12	21	18	9	15
Tumor necrosis factor (ligand) superfamily, member 5 (hyper-IgM syndrome)	652	2			1			1	
Tumor-associated calcium signal transducer 1	692	30	2	98	72	16	73	44	40
Wilms tumor 1	1145	2					8		
Mutated in colorectal cancers	1345	3		2					2
Tumor necrosis factor (ligand) superfamily, member 9	1524	23	3	5			1	2	1
Tumor protein p53 (Li-Fraumeni syndrome)	1846	26	4	6	7	8	16	2	17
Tumor protein D52	2384	2	4	19	3	4	2	2	20
Tumor necrosis factor receptor superfamily, member 17	2556	6	8	2	5	2		1	
Tumor suppressing subtransferable candidate 1	4992	2	1		3	1		1	
Bladder cancer associated protein	5300	46	36	19	16	5	11	16	17
Suppression of tumorigenicity 7	5814	2			1	4			2
Deleted in bladder cancer chromosome region candidate 1	6090	9	9			1			
Breast cancer anti-estrogen resistance 3	6564	16	5	2	11	7	9	8	6
Lung cancer candidate	8186	35	8	5	15	4	6	8	24
Deleted in liver cancer 1	8700	19	4	7		1	4	2	2
Tumor endothelial marker 8	8966	80	7	6	10	6	15	14	10
Tumor up-regulated CARD-containing antagonist of caspase nine	10031	12		1	5	1	2	2	3
Phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	10712	18	3	5	17	5	6	5	4
Tumor endothelial marker 6	12210	37	6	3	11	1	7	11	5
Downregulated in ovarian cancer 1	15432	13	5	2	3	2			4
Candidate tumor suppressor protein	16608	2			1		4		2
Tumor protein D52-like 1	16611	29	9	10	77	9	8	4	44
Tumor endothelial marker 5 precursor	17270	12	1	1	4	4	4	4	8
Cylindromatosis (turban tumor syndrome)	18827	14	1	4	3	9	1	1	2
“transcription factor” related genes (25 of 225 genes)									
GA binding protein transcription factor, alpha subunit (60kD)	78	1	2					2	1
Sterol regulatory element binding transcription factor 1	166	53	12	17	29	8	9	8	39
POU domain, class 4, transcription factor 2	266						1		
General transcription factor IIC, polypeptide 1 (alpha subunit, 220kD)	331	6		3	1	2	2		2
Transcription factor 15 (basic helix-loop-helix)	437						1		
Activating transcription factor 3	460	10	3	5	4	8	7	4	6
ISL1 transcription factor, LIM/homeodomain, (islet-1)	505	3						1	
Nuclear transcription factor Y, alpha	797	1	2						2
POU domain, class 2, transcription factor 2	1101	1		1	1		1	1	1
E2F transcription factor 3	1189	10	2	4	9	2	11	1	
Heat shock transcription factor 1	1499	7	1	3	9	2		3	3
Caudal type homeo box transcription factor 1	1545	1		13		1			1
nterferon-stimulated transcription factor 3, gamma (48kD)	1706	31	7	12	7	8	19	11	4
POU domain, class 3, transcription factor 1	1837	2	1						
Sp1 transcription factor	2021	7		2	2	4	3	1	
E2F transcription factor 5, p130-binding	2331	90	7	24	74	19	30	43	66
Transcription factor-like 1	2430	22	9	15	14	3	9	7	13
POU domain, class 6, transcription factor 1	2815	7	1	5		1	1		1
POU domain, class 5, transcription factor 1	2860	3		2			2	1	1
Transcription factor AP-4 (activating enhancer binding protein 4)	3005	5		12	7		7	1	8
Nuclear transcription factor, X-box binding 1	3187	13	7	4	19	5	2	3	5
Transcription repressor p66 component of the MeCP1 complex	4779	8	1	4	2	3	4	2	4
Transcription termination factor-like protein	5009	11	2	2	5		3	2	7
Activating transcription factor 6	5813	9	2	10	1	2	6	3	2
MYC-associated zinc finger protein (purine-binding transcription factor)	7647	195	44	72	125	16	55	29	139

(Abbrev. Br_c: brain_cancer, Br_n: brain_normal, Co_c: colon_cancer, Mg_c: mammary_gland_cancer, Mg_n: mammary_gland_normal, Ov_c: ovary_cancer, Pa_c: pancreas_cancer, Pr_c: prostate_cancer)

Table 3. The gene expressed correlations discovered.

No	Gene expression correlations	Chi-square Values
1	Hs.13063 transcription elongation regulator 1 (CA150) and Hs.78869 transcription elongation factor A (SII), 1	0.9507
2	Hs.68257 general transcription factor IIF, polypeptide 1 (74kD subunit) and Hs.165743 tumor suppressing subtransferable candidate 4	1.3516
3	Hs.75113 general transcription factor IIIA and Hs.89781 upstream binding transcription factor, RNA polymerase I	1.8783
4	Hs.158196 transcriptional adaptor 3 (ADA3, yeast homolog)-like (PCAF histone acetylase complex) and Hs.165743 tumor suppressing subtransferable candidate 4	2.4114
5	Hs.184693 transcription elongation factor B (SIII), polypeptide 1 (15kD, elongin C) and Hs.241493 natural killer-tumor recognition sequence	2.427
6	Hs.273219 breast cancer anti-estrogen resistance 1 and Hs.278898 tumor necrosis factor alpha-inducible cellular protein containing leucine zipper domains; Huntingtin interacting protein L; transcription factor IIIA-interacting protein	2.6249
7	Hs.89781 upstream binding transcription factor, RNA polymerase I and Hs.93649 upstream transcription factor 2, c-fos interacting	2.6308
8	Hs.14963 chromatin-specific transcription elongation factor, 140 kDa subunit and Hs.154718 tumor protein D52-like 2	2.6618
9	Hs.78788 leucine-zipper-like transcriptional regulator, 1 and Hs.273219 breast cancer anti-estrogen resistance 1	2.9123
10	Hs.239720 CCR4-NOT transcription complex, subunit 2 and Hs.252587 pituitary tumor-transforming 1	3.0761
11	Hs.13063 transcription elongation regulator 1 (CA150) and Hs.165743 tumor suppressing subtransferable candidate 4	3.2184
12	Hs.181243 activating transcription factor 4 (tax-responsive enhancer element B67) and Hs.241493 natural killer-tumor recognition sequence	3.2247
13	Hs.8966 tumor endothelial marker 8 and Hs.323949 kangai 1 (suppression of tumorigenicity 6, prostate; CD82 antigen (R2 leukocyte antigen, antigen detected by monoclonal and antibody IA4))	3.3656
14	Hs.227630 RE1-silencing transcription factor and Hs.252587 pituitary tumor-transforming 1	3.392
15	Hs.119222 suppression of tumorigenicity 13 (colon carcinoma) (Hsp70 interacting protein) and Hs.165743 tumor suppressing subtransferable candidate 4	3.7781

considered tissues. Therefore, we guess that these statistical correlations may reveal some useful information about regulation, biochemical pathways, and the signal transduction process. Another example of correlated genes is shown as follows:

Hs.25 ATP synthase, H⁺ transporting, mitochondrial F1 complex, beta polypeptide and
Hs.108319 thyroid hormone receptor-associated protein, 150 kDa subunit

The Chi-square value is 2.6553 (**statistically significant**) {Significance level is 14.0671}

“Hs.25” is the gene identifier in the UniGene database, and “ATP synthase, H⁺-transporting, mitochondrial F1 complex, beta polypeptide” is the description of the gene. In this example, the genes “Hs.25” and “Hs.108319” have the Chi-square value of 2.6553, which is smaller than the significance threshold, means that these two genes have similar expression profiles in the considered tissues. They are statistically correlated.

Table 4 shows the detailed statistics of the genes expressed in these experimental tissues. One can see the expression counts and expression levels of the genes in each tissue. The normalized results are also listed. The sample size of each experimental tissue is normalized to 100,000 counts.

Table 4. The expression profiles of genes in each experimental tissue.

Tissue name & Histology (Sample size)	Expression counts (Expression level)		Normalization (Counts per 100,000 times)	
	Hs.25	Hs.108319	Hs.25	Hs.108319
brain_cancer (568507)	218 (0.0383%)	108 (0.0190%)	38.35	19.00
brain_normal (128664)	42 (0.0326%)	20 (0.0155%)	32.64	15.54
colon_cancer (165450)	83 (0.0502%)	54 (0.0326%)	50.17	32.64
mammary_gland_cancer (310142)	62 (0.0200%)	52 (0.0168%)	19.99	16.77
mammary_gland_normal (120912)	25 (0.0207%)	15 (0.0124%)	20.68	12.41
ovary_cancer (142391)	71 (0.0499%)	39 (0.0274%)	49.86	27.39
pancreas_cancer (118994)	45 (0.0378%)	22 (0.0185%)	37.82	18.49
prostate_cancer (241899)	119 (0.0492%)	69 (0.0285%)	49.19	28.52

4. DISCUSSION

In this study, we found correlations of gene expressions in tissues with different histologies in human. We will probably discover those collaborative genes statistically if the genome-wide gene expression products in the majority of tissues can be really reflected.

We did not estimate whether a gene was differentially expressed in one or more tissues. We focused on those genes that were expressed concurrently or regularly expressed in many different tissues. This means that we are interested in such gene correlations, which might be co-regulated by the same inducer, involved in the same biochemical pathway or signal transduction process, or coded for different subunits of multi-component enzyme complexes.

Though we mixed the libraries according their tissue types to enlarge the sample size of the experimental data, some expression profiles of tissues are not enough to really reflect truly cellular expression because of the small sample size. In this study, we only selected tissues with more than 100,000 counts as experimental data in the example described above. Most of the genes expressed in these selected tissues could probably be sequenced, and the expression level of the genes would be close to that of the real cells.

Highly expressed genes are always sequenced using tag-sampling approaches, and low abundant ones may easily be lost unless the sampling is large enough, especially in ESTs. The main defect of this study is that a lot of the expression profiles of tissues were incomplete. So we only applied our methods to believable expression profiles.

The more tag-samplings, the more complete the expression profiles. If the expression profiles of most tissues can more completely reflect true cellular expression in the future, it will be to anticipate that more correlations of genes will be mined using the method we have proposed.

We did not examine gene expression within a specific cell-type. The principle aim of our study was to obtain overall correlations, for example, genes whose expression was correlated across many different tissues. In the future, the correlations we mined may be combined with other molecular data, such as protein interactions, to provide more useful information, and the mined correlations will be verified by biologists to identify more interesting biological relationships.

ACKNOWLEDGEMENT

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 89-2213-E-008-061. In addition, we would like to thank Professors Ueng-Cheng Yang at National Yang-Ming Univ. and Chi-Gong Tong at National Central University for their helpful suggestions.

REFERENCES

1. M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev, "dbEST—database for "expressed sequence tags"," *Nat Genet*, Vol. 4, 1993, pp. 332-333.
2. G. D. Schuler, "Pieces of the puzzle: expressed sequence tags and the catalog of human genes," *Journal of Molecular Medicine*, Vol. 75, 1997, pp. 694-698.
3. G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannikulchai, A. Chu, C. Clee, S. Cowles, P. J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, T. J. Hudson, and et al., "A gene map of the human genome," *Science*, Vol. 274, 1996, pp. 540-546.
4. H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M. C. Hermus, R. van Asperen, K. Boon, P. A. Voute, S. Heisterkamp, A. van Kampen, and R. Versteeg, "The human transcriptome map: clustering of highly expressed genes in chromosomal domains," *Science*, Vol. 291, 2001, pp. 1289-1292.
5. D. Zhuo, W. D. Zhao, F. A. Wright, H. Y. Yang, J. P. Wang, R. Sears, T. Baer, D. H. Kwon, D. Gordon, S. Gibbs, D. Dai, Q. Yang, J. Spitzner, R. Krahe, D. Stredney, A. Stutz, and B. Yuan, "Assembly, annotation, and integration of UNIGENE clusters into the human genome draft," *Genome Res*, Vol. 11, 2001, pp. 904-918.
6. J. L. Hess, "The cancer genome anatomy project: power tools for cancer biologists," *Cancer Invest*, Vol. 21, 2003, pp. 325-326.
7. V. E. Velculescu, S. L. Madden, L. Zhang, A. E. Lash, J. Yu, C. Rago, A. Lal, C. J.

- Wang, G. A. Beaudry, K. M. Ciriello, B. P. Cook, M. R. Dufault, A. T. Ferguson, Y. Gao, T. C. He, H. Hermeking, S. K. Hiraldo, P. M. Hwang, M. A. Lopez, H. F. Luderer, B. Mathews, J. M. Petroziello, K. Polyak, L. Zawel, K. W. Kinzler, and et al., "Analysis of human transcriptomes," *Nat Genet*, Vol. 23, 1999, pp. 387-388.
8. V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, "Serial analysis of gene expression," *Science*, Vol. 270, 1995, pp. 484-487.



Jorng-Tzong Horng (洪炯宗) was born in Nantou, Taiwan, on April 10, 1960. He received the Ph.D. degree in Computer Science and Information Engineering from National Taiwan University, Taipei, in April 1993. In 1993, he joined the Department of Computer Science and Information Engineering, National Central University, Chungli, Taiwan, where he became Professor in 2002. His current research interests include database systems, data mining, genetic algorithms, and bioinformatics.



Hsien-Da Huang (黃憲達) was born in Taoyuan, Taiwan, in 1975. He received his B.S. degree in 1997 in Computer Science and Information Engineering in National Central University, Taiwan. He started his graduate studies on Bioinformatics in 1999, and received his Ph.D. degree in Institute of Computer Science and Information Engineering in National Central University, Taiwan. His current research interests are bioinformatics, database systems, and data mining.



Kuo-Yen Tseng (曾國炎) was born in Hsinchu, Taiwan, on April 18, 1975. He received master degree in Computer Science and Information Engineering from National Central University, Chungli, in 2002. After graduation he entered into D-Link Corporation, a famous network device manufacturing company, serving in system research and develop department as an engineer from 2002 till now. He is interested in database systems and bioinformatics.



Tsung-Shan Tsou (鄒宗山) is currently associate Professor of Biostatistics of the Graduate Institute of Statistics, National Central University, Taiwan. He received his B.S. degree in Mathematics from National Taiwan University and got the M.S. degree in Statistics from National Central University, Taiwan, in 1983 and 1987, respectively. He then spent 5 years in the Department of Biostatistics, School of Hygiene and Public Health, the Johns Hopkins University, Baltimore, U.S.A., and received the Ph.D. degree in 1992. His current research interests include robust statistical inferences methodology and the development of statistical tools for data mining technologies for bioinformatics.



Baw-Jhiune Liu (劉寶鈞) is a Professor in the Department of Computer Science and Information Engineering at Yuan Ze University in Taiwan since 1999. He received his B.S. and M.S. degrees in Electrical Engineering from National Cheng Kung University, Taiwan, in 1967 and 1969 respectively, and his Ph.D. degree in Electrical Engineering from National Taiwan University, Taiwan, in 1979. He worked for Telecommunication Labs in Chungli, Taiwan from 1970 to 1973. He was Associate Professor in the Department of Computer Science and Information Engineering of National Taiwan University, Taiwan, from 1979 to 1983. He was a Professor in the Department of Computer Science and Information Engineering of National Central University, Taiwan, from 1983 to 1999. His current research interests include the development of data mining technologies for bioinformatics and the data models for web group learning.



Cheng-Yan Kao (高成炎) was born in Taipei, Taiwan, 1948. He received the B.S. degree in mathematics from National Taiwan University, Taipei, Taiwan, in 1971, and the M.S. degree in computer science in 1976, the M.S. degree in statistics in 1978, and the Ph.D. degree in computer science in 1981, all from the University of Wisconsin-Madison. He has previously been with Ford Aerospace, the Unisys Corporation, and General Electric from 1980 to 1989 at the Johnson Space Center, NASA, Houston, TX. He has been a professor with the Department of Computer Science and Information Engineering, National Taiwan University (NTU) since 1990. He has published more than 30 technical papers in various journals and international conferences. His research interests include bioinformatics, evolutionary algorithms, computational molecular biology, optimization theory and software engineering. He has been the director of NTU bioinformatics research center and president of Bioinformatics Society Taiwan (BST) since 2001.