Cross-SAGE: SAGE Data Mining Tool Based on Set Theory

Cheng-Hong Yang, Tsung-Mu Shih, De-Leung Gu, Hsueh-Wei Chang and Li-Yeh Chuang

***Abstract*—SAGE (Serial Analysis of Gene Expression) is a technique for analyzing gene expression of a series of tags obtained from cDNA. This technology lets biologists analyze and observe the relative gene expressions of many samples *in silico*. To date, some visible analyzing platforms for SAGE were not provided in a biologically significance way for cross-analysis and -comparison, thus limiting its application. Therefore, we retrieved various SAGE databases of *Homo sapiens* from NCBI SAGEmap and proposed a powerful tool for cross-analyzing gene expression among different SAGE libraries of tissue sources. In this paper, we combine the mathematical set theory with a unique multi-group method to analyze SAGE data, and provide the function for gaining the corresponding information between tags and genes. Some up- or down-regulated tissue-specific markers which are or are not common to others could be identified computationally. This method is a viable and convenient way to analyze gene expression in complex comparison, and can obtain analysis results by biological significance.**

***Index Terms*—cDNA, gene expression, restriction enzyme, SAGE, Set theory.**

I. Introduction

SAGE (Serial Analysis of Gene Expression) is a gene expression data quantifying technique, which was proposed by V.E. Velculescu et al [1]. In short, biologists aim at cDNA (or mRNA) sequences of interest in biological samples, then obtain a series of short sequences (SAGE tag, tag, base pair sequence) and the counts of each short sequence in these short sequences by specific restriction enzymes (e.g. BsmFI, NlaIII, and Sau3A) [1]. In principle, such techniques can be used to verify and quantify the transcripts of biological samples, and construct the cognition of the distribution and regulation of transcription in normal or abnormal cell types. Then, biologists can gain information of each specific tag representing each specific gene, and analyze the gene expression (tag counts) [2].

Manuscript received December 31, 2007.

Cheng-Hong Yang is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan (e-mail: chyang@cc.kuas.edu.tw).

Tsung-Mu Shih is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan (e-mail: gmtsungmu@gmail.com).

De-Leung Gu is with the Faculty of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Taiwan (e-mail: ed0958719325@yahoo.com.tw).

Hsueh-Wei Chang is with the Faculty of Biomedical Science and Environmental Biology, and Graduate Institute of Natural Products, Kaohsiung Medical University, Taiwan (e-mail: changhw@kmu.edu.tw).

Li-Yeh Chuang is with the Department of Chemical Engineering, I-Shou University, Taiwan (e-mail: chuang@isu.edu.tw).

The SAGE technique enables biologists to widely analyze gene expression in various biological samples in order to observe and compare comparatively high or low tag counts in normal and abnormal samples in. However, SAGE usually generates a huge amount of experimental data (including noisy and redundant data). It is necessary to extract, sift and arrange the useful information in the various biological samples (SAGE data). Hence finding a key tag (or tags) for further experimentation and identification is required.

As far as we know, most visible analysis platforms either lack in the analysis by biological significance or do not satisfy the requirement of an analysis function. For example, a web-based analysis platform of SAGE called SAGEmap is located in NCBI, which was proposed by A.E. Lash et al [3]. SAGEmap not only supplies various SAGE data of biological samples for downloading, but also provides an analysis function to obtain the tag counts for each tag in various samples and cell types. But this analysis function is restricted to only two groups for comparing and analyzing, and the displayed result was long-winded and had a poor ranking. Recently, an analysis platform called TIGR MeV was proposed by H.Y. Wang et al [4]. TIGR MeV clusters the SAGE data (Self-Organizing Map and hierarchical clustering) by using statistical methods (PoissonS and PoissonHC) to obtain a cluster of tags with similar statistics, but the methods of analysis may be ambiguous on biological significance. In addition, ACTG (by P.A.F. Galante et al) [5] and TAGmapper (by P. Bala et al) [6] only supply a function for gaining the corresponding information between tags and genes.

In light of the above cases, we provide a function for obtaining the corresponding information between tags and genes, and a function for analyzing the SAGE data by combining the mathematical set theory with a unique multi-group method, in order to construct a gene expression cross-mining tool for extracting and mining the information in large-scale SAGE data in various biological samples. Using the visualized Multi-Group Analysis method, biologists can obtain the tag counts of each tag in various samples more conveniently. Tags which are comparatively high or low are easy to find, providing a set of key tags of curative genes or pathogenic genes.

II. System Design

In this paper, we developed a cross-analysis platform for SAGE data. Biologists can pick up samples of interest freely, and analyze the SAGE data by using the Multi-Group Analysis method to obtain the tag counts of each tag in

various samples. In order for biologists to analyze the data more conveniently, we display the results in tabular and graphic form, and provide a function for gaining information of the specific tag to specific gene. The details of the source data, the system processes and structure, and the Multi-Group Analysis are described in detail below.

A. *Source data*

In this tool, we focus on the SAGE data in *Homo sapiens*, and take 327 samples (libraries) of SAGE data in various types of various samples from NCBI SAGEmap [7] for Multi-Group Analysis. Samples included brain, kidney, breast, ovary and colon, amongst others. In the Tag to Gene function, we retrieve the corresponding information of tags to genes in UniGene, which used the restriction enzymes NlaIII and Sau3A.

B. *System processes and structure*

In order to facilitate analysis of the SAGE data and to retrieve the results on various computer platforms, we constructed this tool based on the JAVA language, and separated it into a data unit, analysis function unit and appearance unit by using the MVC (Model-View-Controller) architecture [8]. As shown in Fig. 1, we divided our system into five units:

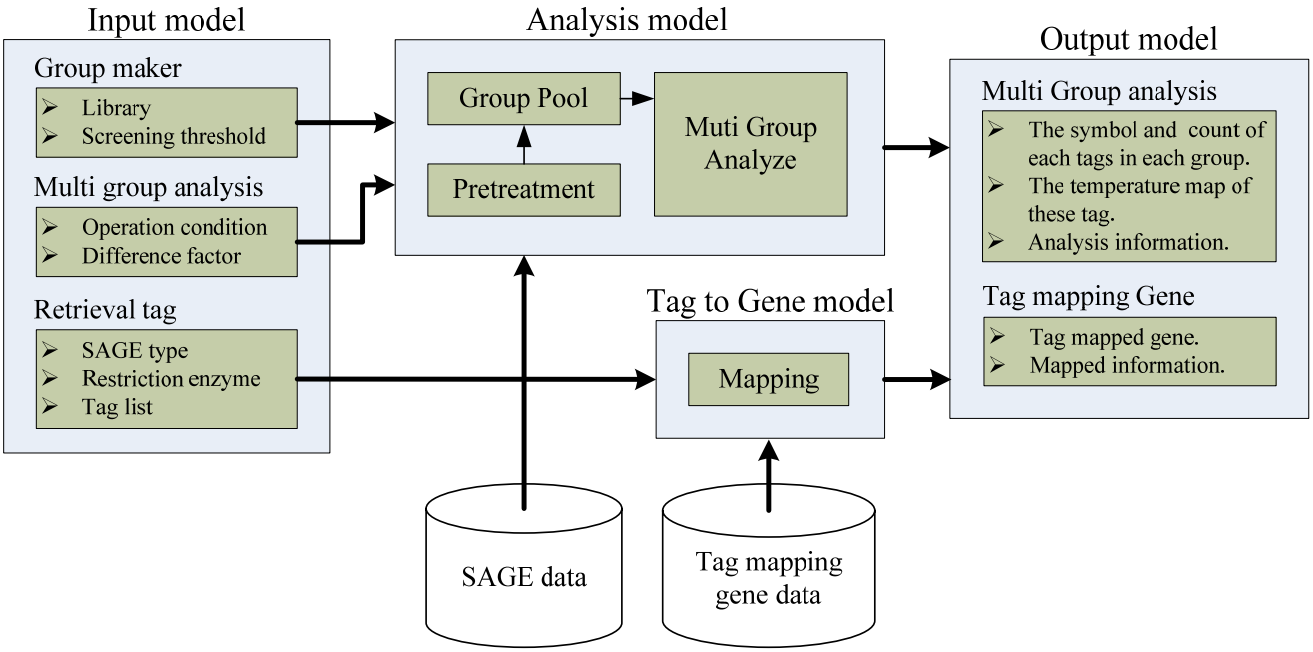

Fig. 1. The system processes and structure.

(1) Input module

Before analysis with Multi-Group Analysis, users must first create a group by choosing one or more libraries of interest and set the screening threshold (*tpm*, *tag per million*) for the SAGE data, then load two or more groups of interest and set the *operation condition* and *difference factor* between these groups for analysis. While in Tag to Gene, users need to submit a list of tags, and then choose the type of SAGE tag (10 bp or 17 bp) and the restriction enzyme (NlaIII or Sau3A).

(2) Analysis module

In this module, pretreatment of the groups of interest is performed, in which tag count of each tag in each group is transformed into *tpm* (*tag per million*) with the same expression level, followed by storing this of transformation. Users can extract the information of the transformation by using the Multi-Group Analysis function.

(3) Tag to Gene module

This module finds the corresponding information between a tag and a gene from the Tag mapping gene data, according to a submitted tag list, the type of SAGE tag and the restriction enzyme.

(4) Output module

With this tool biologists can obtain two kinds of result information. The first one is the tabular and graphic form of the results from the Multi-Group Analysis function. It contains *tpm* information of each group. The results can be sorted according to the *tpm* in a selected group or the difference of *tpm* of the same tag between two selected groups. The other is a list of corresponding information between tags and genes from the Tag to Gene function, which includes information about the cluster ID, UniGene cluster title, mapping score, gene symbol, and so on.

(5) Data set

The data set can be divided into two parts of SAGE data and Tag mapping gene data. They are 10 bp (normal SAGE, e.g. AGCCTTGCTA) or 17 bp (LongSAGE [9], e.g. CGTTCCTATGACTAGCC) for the SAGE data. SAGE data includes a list of non-repeat tags and the count of each tag in 327 libraries by using the restriction enzymes NlaIII or Sau3A. Tag mapping gene data includes a list of tags and the corresponding cluster ID, UniGene cluster title, mapping score and gene symbol by using the restriction enzymes NlaIII or Sau3A.

III. Multi-Group Analysis

Set theory is a practical and important theory in

mathematics. With each set being composed of a number of elements, we can construct the relation between several sets, [10], [11].

In this paper, we extract a list of relevant tags by combining the set theory and a unique multi-group method for analysis of the SAGE data. Before analysis, we assume that there exist: 1) n groups (G_1, G_2, …, G_n) as shown in Fig. 2., which include large-scale of non-repeat SAGE tags and the count of each tag in each group. 2) A list of non-repeat conditions as shown in equation (1), $1 \leqq i \leqq n$, $1 \leqq j \leqq n$, $i \neq j$, $1 \leqq k \leqq$ number of conditions, op_k represents the operation (more than, equal to or less than), and $factor_k$ represents the difference multiple between G_i and G_j. Then, we start to extract the information as shown in Fig. 3.

$$G_i \; op_k \; (factor_k \cdot G_j) \qquad (1)$$

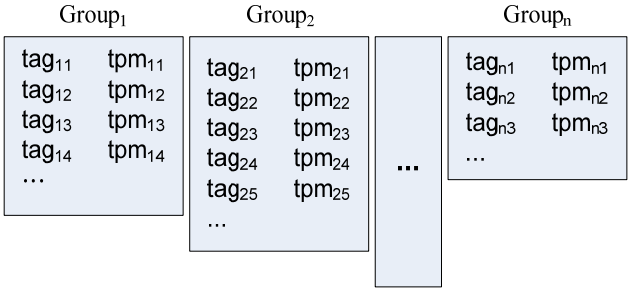

Fig. 2. n groups and the data in each group.

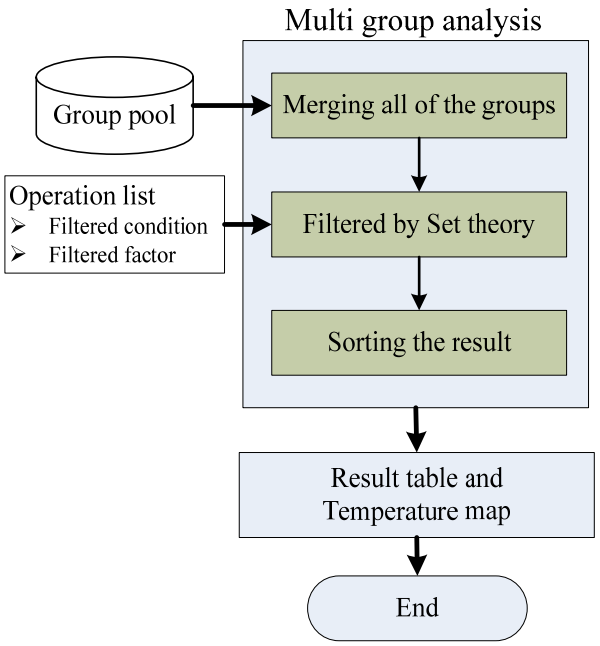

Fig. 3. The Multi-Group Analysis process.

A. *Merging of all groups*

First, we need to load all of the groups into the group pool. These group data will then be merged into a new form: the multi-group form. tpm_{kj} will be set to zero, if tag_{ij} in G_i does not exist in G_k. The content of the merged data is shown in Fig. 4.

B. *Filtering by Set theory*

A filtering process is implemented to extract significant tags and abandon trifling tags by using equation (2) in the multi-group form.

$$\left\{ tag_p, 1 \leq p \leq m : tag_p \middle| \; G_i \; op_k \; (factor_k \cdot G_j) \right\} \qquad (2)$$

	G_1	G_2	G_3	...	G_n
tag_1	tpm_{11}	tpm_{21}	tpm_{31}	...	tpm_{n1}
tag_2	tpm_{12}	tpm_{22}	tpm_{32}	...	tpm_{n2}
tag_3	tpm_{13}	tpm_{23}	tpm_{33}	...	tpm_{n3}
...	...	...	...	...	...
tag_m	tpm_{1m}	tpm_{2m}	tpm_{3m}	...	tpm_{nm}

Fig. 4. Multi-group form after merging of group data.

C. *Sorting the results*

This tool provides two ways for ranking the results in the multi-group form. It can rank all of the tags from small to big according to the *tpm* of each tag in the appointed group in the multi-group form, or it ranks all of the tags from small to big according to the *tpm difference* of each tag between two appointed groups in the multi-group form. The result can be displayed in table form or temperature difference.

IV. Results and Discussion

In this paper, we constructed a gene expression cross-mining tool for extracting information from large-scale SAGE data in various biological samples. By using this tool, we can choose one or more libraries of interest, and set a screening threshold to filter SAGE data to create a group by using Group Maker (as shown in Fig. 5.). Then, two or more groups of interest are loaded, and the *operation condition* and *difference factor* between these groups are set to find the significant tags and abandon the insignificant ones (Fig. 6.). Biologists can submit a tag list to obtain the corresponding information between tags and genes by using the Tag to Gene function.

In order to test the Multi-Group Analysis function, we created three groups to be analyzed as an example. The three groups were *normal_colon* (enzyme-NlaIII, Normal colonic epithelium, epithelium, normal, colon, SAGE, CGAP, non-normalized, SAGE library method, bulk), *primary tumor_colon* (enzyme-NlaIII, Colon, primary tumor) and *tumor_stomach* (enzyme-NlaIII, tumor, stomach, adenocarcinoma, CGAP, non-normalized, SAGE library method, bulk). These groups consisted of a total of 99,772 tags with 29,272 kinds of non-repeat tags for *normal_colon*, a total of 57,686 tags with 23,001 kinds of non-repeat tags for *primary tumor_colon*, and a total of 66,032 tags with 21,967 kinds of non-repeat tags for *tumor_stomach*. Then we set the operation conditions and difference factors: 1) *normal_colon* was more than *primary tumor_colon* by a multiple of 10. 2) *primary tumor_colon* was equal to *tumor_stomach*. We obtained the 66 non-repeat tags and tag counts of each tag in *normal_colon*, *primary tumor_colon* and *tumor_stomach*. The results are partly shown in Fig. 7 and Fig. 8.

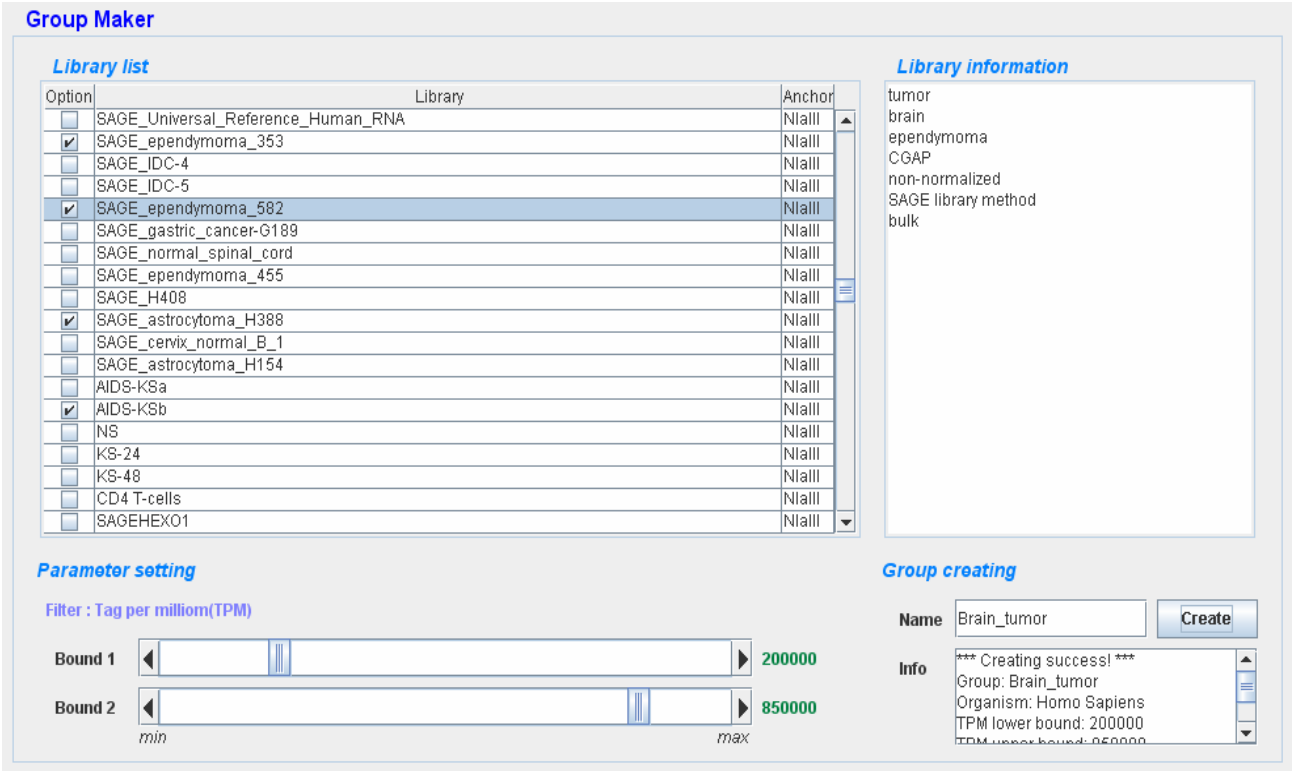

Fig. 5. The Group Maker interface.

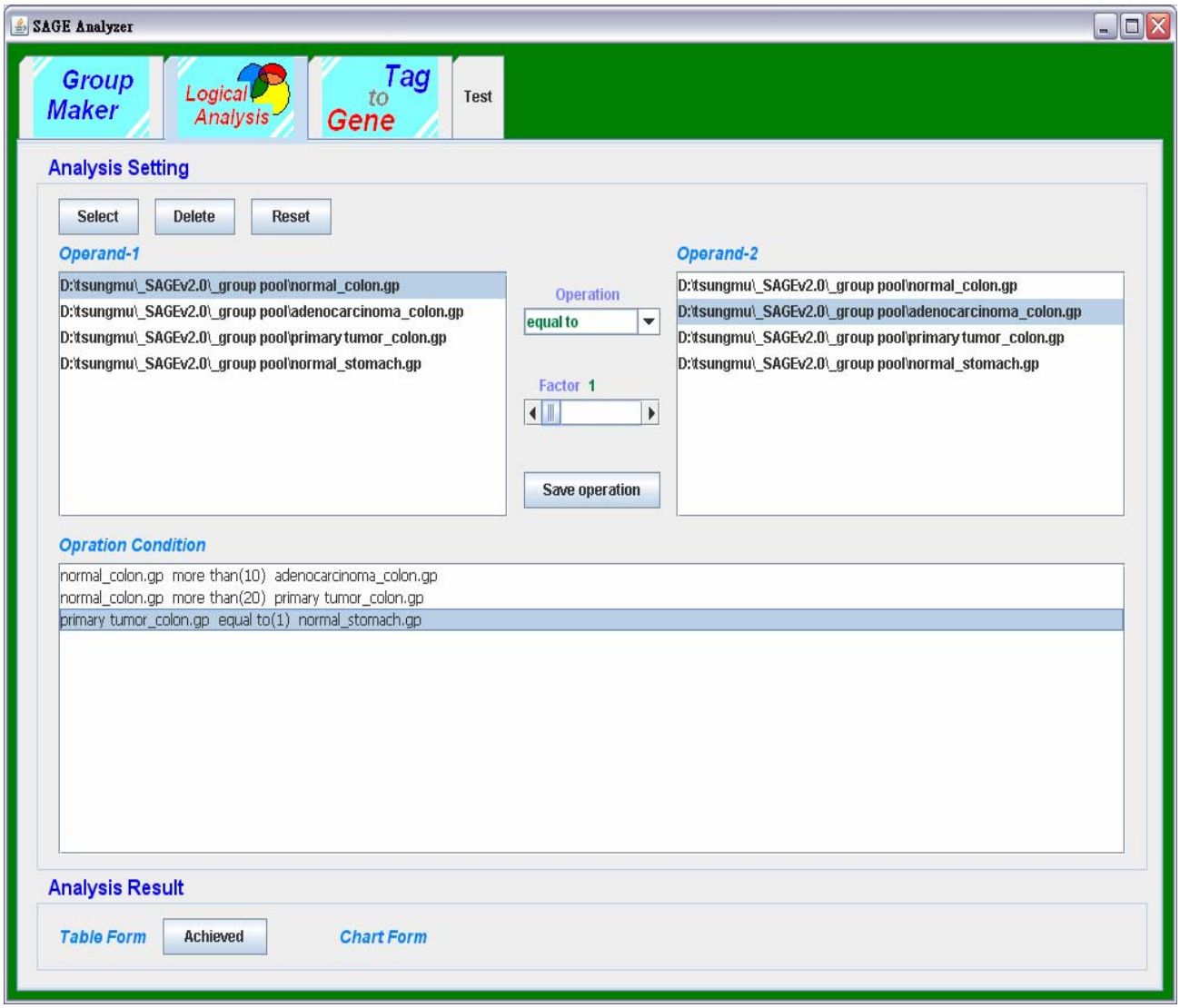

Fig. 6. The Multi-Group Analysis interface.

Table. 1. Comparison of SAGEmap, TIGR MeV, ACTG and TAGmapper.

	SAGEmap [3]	TIGR MeV [4]	ACTG [5]	TAGmapper [6]	Cross-SAGE
Type of program	Web	Download	Web	Web	Download
Multi sample analyzer	Only two	Yes	No	No	Yes
Tag to Gene	Yes	No	Yes	Yes	Yes
Visualization of results	No	Yes	No	No	Yes

The brief and neatly ranked results show the tags and the tag count of each tag in various groups, and helps biologists understand which tag is abnormal. For example, the tag count of GCCCAGGCTA and ACATTGGGTG was correspondingly low in *primary tumor_colon* and *tumor_stomach*, as shown in Fig. 7 and Fig. 8. To decide whether actual research and verification is needed, we retrieve the corresponding information between tags and genes in these 66 tags. In the mapped results, there are 65 tags with the mapped genes in UniGene cluster, and only one tag without a corresponding gene. A part of the results is shown in Fig. 9, which contains the corresponding UniGene cluster ID, UniGene cluster title, mapping score (reliability) and gene symbol for each tag. In this paper, we used two functions for SAGE data analysis, and made a comparison with other systems as shown in Table. 1.

Result Table

Option	Tag	normal_colon	primary tumor_colon	tumor_stomach
	GCCTCCCAGG	22	1	1
	GGCTGCCTGC	24	2	2
	TCAGAGCGCT	26	1	1
	TTTCTCGTCG	26	2	2
	CCAACACCAG	28	1	0
	GCCAGACACC	28	2	1
	GCCACGTGGA	35	1	0
	ATGACGCTCA	40	4	0
	CCCGCCTCTT	40	1	0
	GTGTTGGGGG	40	3	0
	CTGGGCCTCT	44	3	3
	GCAAGAAAGT	48	1	1
	GCCCAGGGCC	48	1	0
	GCAGCTCCTG	60	4	4
	GGCACCGTGC	66	4	4
	TCAGCTGCAA	72	3	0
	TGAGTGACAG	77	7	2
	ATGCGGGAGA	78	6	5
	GCCGACCAGG	93	9	4
	ACACAGCAAG	118	5	0
	ATACTCCACT	142	3	0
	TAAATTGCAA	162	3	0
	GGAACTGTGA	174	2	2
	ATTGGAGTGC	221	19	1
	TGCTCCTACC	253	22	3
	ACATTGGGTG	712	33	2
	GCCCAGGTCA	968	22	3

Fig. 7. Part of the results in the three groups in table form.

Result Table

Option	Tag	normal_colon	primary tumor_colon	tumor_stomach
	GCCTCCCAGG			
	GGCTGCCTGC			
	TCAGAGCGCT			
	TTTCTCGTCG			
	CCAACACCAG			
	GCCAGACACC			
	GCCACGTGGA			
	ATGACGCTCA			
	CCCGCCTCTT			
	GTGTTGGGGG			
	CTGGGCCTCT			
	GCAAGAAAGT			
	GCCCAGGGCC			
	GCAGCTCCTG			
	GGCACCGTGC			
	TCAGCTGCAA			
	TGAGTGACAG			
	ATGCGGGAGA			
	GCCGACCAGG			
	ACACAGCAAG			
	ATACTCCACT			
	TAAATTGCAA			
	GGAACTGTGA			
	ATTGGAGTGC			
	TGCTCCTACC			
	ACATTGGGTG			
	GCCCAGGTCA			

Fig. 8. Part of the results in three groups in temperature difference.

Tag mapped Gene

Tag	UniGene cluster...	UniGene cluster title	mapping score	Gene symbol
GCTGGCCTTG	67928	E74-like factor 3 (ets do...	3004026	ELF3
GCTGGCCTTG	661581	Transcribed locus, weak...	2001005	
GGAACTGTGA	38972	Tetraspanin 1	4013625	TSPAN1
GGAACTGTGA	572587	Transcribed locus	1001002	
GGACTAAATG	466814	Carcinoembryonic antig...	3001508	CEACAM5
GGCACCGTGC	390599	Hypothetical gene supp...	3003523	LOC440335
GGCCCTGCAG	423756	Sirtuin (silent mating typ...	4025562	SIRT6
GGCCCTGCAG	502914	Dipeptidyl-peptidase 3	1000193	DPP3
GGCTGCCTGC	569809	Rho GTPase activating ...	3020052	ARHGAP27
GGCTGCCTGC	165859	Anthrax toxin receptor 1	2001003	ANTXR1
GTATTGGTGA	122927	CAP-GLY domain contai...	1000017	CLIP4
GTATTGGTGA	380135	Fatty acid binding protei...	1000013	FABP1
GTATTTGCAA	197922	Calcium/calmodulin-de...	4031604	CAMK2N1
GTATTTGCAA	591269	Calcium/calmodulin-de...	1000003	CAMK4
GTCAACAGTA	463421	ATP-binding cassette, s...	2004009	ABCC3
GTGACAGAAT	516217	UDP-glucose pyrophos...	4070714	UGP2
GTGACAGAAT	129673	Eukaryotic translation ini...	2000503	EIF4A1
GTGATGAGCT	634941	Immunoglobulin heavy v...	1000740	IGHV1-69
GTGATGAGCT	510635	Immunoglobulin heavy c...	1000376	IGHG1
GTGGCCAGAG	1420	Fibroblast growth factor ...	4035569	FGFR3
GTGGCCAGAG	591833	Scavenger receptor clas...	2001751	SCARA5
GTGGTGCGCG	592837	Transcribed locus	1000003	
GTGTTGGGGG	55016	EPS8-like 2	5018038	EPS8L2

Fig. 9. Part of the results in tag to gene form.

V. Conclusion

In this paper, we introduce a convenient and relevant analysis tool for cross-mining gene expression. Brief and neatly ranked results are obtained by using a Multi-Group Analysis function. Biologists can obtain a list of tags and the tag count of each tag in various groups in an obvious way and easily identify abnormal tags. Subsequently, the mined tags are accepted for actual research and verification to identify curative or pathogenic tags. Biologists also gain the corresponding information between tags and genes by using the convenient Tag to Gene function. This gene expression mining tool and the form in which the results are presented allows biologists to conveniently handle and analyze gene expression in various samples from a biologically feasible point of view.

References

[1] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler, “Serial Analysis of Gene Expression”, Science, 270:pp. 484-487, Oct. 1995.

[2] S.M. Wang, “Understanding SAGE data”, Trends in Genetics, Vol. 23, Issue 1:pp. 42-50, Jan. 2007.

[3] A.E. Lash, C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg and G.J. Riggins, S.F. Altschul, “SAGEmap : a public gene expression resource”, Genome Research, Vol. 10, Issue 7, pp. 1051-1060, Jul. 2000. Available: http://www.ncbi.nlm.nih.gov/SAGE/

[4] H.Y. Wang, H. Zheng, and F. Azuaje, “Poisson-Based Self-Organizing Feature Maps and Hierarchical Clustering for Serial Analysis of Gene Expression Data”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, Issue 2, pp. 163-175, Apr. 2007. Available: http://www.tm4.org/mev.html

[5] P.A.F. Galante, J. Trimarchi, C.L. Cepko, S.J. de Souza, L. Ohno-Machado and Winston P. Kuo, “Automatic correspondence of tags and genes (ACTG): a tool for the analysis of SAGE, MPSS and SBS data”, Bioinformatics, Applications note, Vol. 23 no. 7, pp. 903–905, Feb. 3, 2007. Available: http://retina.med.harvard.edu/ACTG/

[6] P. Bala, R.W. Georgantas III, D. Sudhir, M. Suresh, K. Shanker, B.M. Vrushabendra, C.I. Civin and A. Pandey, “TAGmapper: A web-based

tool for mapping SAGE tags", ScienceDirect, Gene, Vol. 364, pp. 123-129, Dec 30. 2005. Available: http://tagmapper.ibioinformatics.org/

[7] A.E. Lash, C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg and G.J. Riggins, S.F. Altschul, "SAGE data and tag mapping gene data from SAGEmap in NCBI", ftp://ftp.ncbi.nlm.nih.gov/pub/sage/.

[8] S. Burbeck, "Applications Programming in Smalltalk-80(TM): How to use Model-View-Controller (MVC)", The UIUC Smalltalk Archive, Mar 4. 1997. Available: http://st-www.cs.uiuc.edu/users/smarch/st-docs/mvc.html

[9] S. Saha, A.B. Sparks, C. Rago, V. Akmaev, C.J. Wang, B. Vogelstein, K.W. Kinzler and V.E. Velculescu, "Using the transcriptome to annotate the genome", Nature Biotechnology, Vol. 20, Issue 5, pp. 508-512, May. 2002.

[10] R.P. Grimaldi, "Discrete and Combinatorial Mathematics: An Applied Introduction (4th Edition)", Addison Wesley Publishing Company, Oct. 1998.

[11] M. Foreman, A. Kanamori and M. Magidor, "Handbook of Set Theory", http://www.tau.ac.il/~rinot/host.html.