Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

Extract-SAGE: An integrated platform for cross-analysis and GA-based selection of SAGE data

Cheng-Hong Yang[1], Tsung-Mu Shih[1], Yu-Chen Hung[2], Hsueh-Wei Chang[2,3,4,*], Li-Yeh Chuang[5]

[1]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan; [2]Graduate Institute of Natural Products, College of Pharmacy, Kaohsiung Medical University, Taiwan; [3]Center of Excellence for Environmental Medicine, Kaohsiung Medical University, Taiwan; [4]Faculty of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Taiwan; [5]Department of Chemical Engineering, I-Shou University, Taiwan. Hsueh-Wei Chang Email: changhw@kmu.edu.tw; * Corresponding authors

received January 19, 2009; accepted February 06, 2009; published February 27, 2009

Abstract:
Serial analysis of gene expression (SAGE) is a powerful quantification technique for gene expression data. The huge amount of tag data in SAGE libraries of samples is difficult to analyze with current SAGE analysis tools. Data is often not provided in a biologically significant way for cross-analysis and -comparison, thus limiting its application. Hence, an integrated software platform that can perform such a complex task is required. Here, we implement set theory for cross-analyzing gene expression data among different SAGE libraries of tissue sources; up- or down-regulated tissue-specific tags can be identified computationally. Extract-SAGE employs a genetic algorithm (GA) to reduce the number of genes among the SAGE libraries. Its representative tag mining will facilitate the discovery of the candidate genes with discriminating gene expression.

Availability: This software and user manual are freely available at ftp://sage@bio.kuas.edu.tw/Extract-SAGE.zip

Keywords: SAGE; genetic algorithm; set theory; software

Background:
Serial analysis of gene expression (SAGE) is a technique that allows global profiling of gene expression in a genome without a priori knowledge **[1].** The SAGE technique enables biologists to identify a series of short sequences, as well as the count of each sequence (SAGE tag) for the gene expression profile of cell or tissue types. Each short sequence is collected in a SAGE library, and the count of each short sequence represents the gene expression of its corresponding genes. Recently, many public gene expression profile platforms have been developed for use in SAGE analysis. However, most of these platforms are restricted to only two groups of paired comparison and analysis, and the displayed results are often long-winded and show poor ranking **[2, 3].** Therefore, it is necessary to extract, filter and arrange the useful information a way applicable to profile gene expressions, especially when it comes to multiple SAGE libraries containing myriad biological samples. In this study, we construct a cross-analysis method with visualized output for SAGE data analysis, along with retrieval of the corresponding information between SAGE tags and genes. A genetic algorithm (GA) is introduced to facilitate the analysis and accuracy of the SAGE data available to biologists, thus avoiding manual browsing and comparison of the original SAGE data.

Methodology:
Implementation:
Extract-SAGE is programmed in the JAVA language **[4]** and compatible to many computer platforms. We analyzed 327 samples of Homo sapiens SAGE data in various types of samples from NCBI SAGEmap **[5],** i.e. as brain, kidney, breast, ovary, and colon data, amongst others. For tag to gene data, restriction enzymes NlaIII and Sau3A generated the SAGEmap **[5]**. A filtering process of gene expression data was implemented to extract significant tags and abandon trifling tags by incorporating set theory **[6].** A GA was used to implement the feature selection process, and the K-nearest neighbor (KNN) method was used to evaluate the classification accuracy **[7].**

Software description:
Figure 1 shows three functions provided by Extract-SAGE, i.e. 1) cross-analysis, 2) tag to gene, and 3) reducing-analysis (using GA). The "cross-analysis" function provides significant genes extracted by setting some operation conditions and difference factors between samples or sample groups of interest. Two output results, a tabular and graphic form, are provided. Both of them contain tag expression (tag per million, tpm) information of each group, and can be sorted based on the expression in the selected group or the expression difference between two selected groups. The graphic visualization of the results in gradient colors for the tag count in various samples is convenient for selecting gene candidates of interest. Tags with high or low expression (tpm) are easy to identify, and a set of key tags of curative or pathogenic genes is also provided. Users can submit a tag sequence with the "tag to gene" function to retrieve the corresponding information between tags and genes.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(7): 291-292 (2009)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

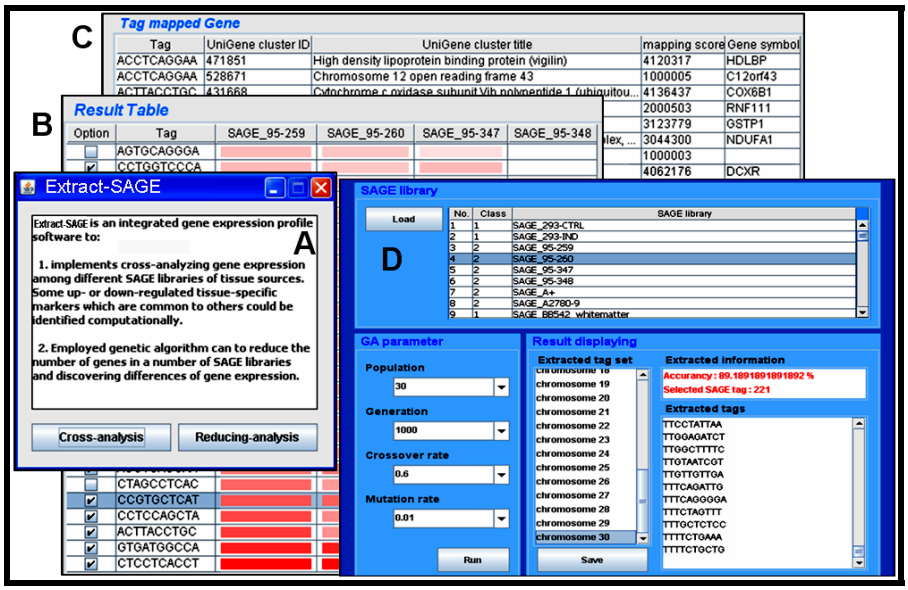

Figure 1: Screenshot of Extract-SAGE. (A) The main window. Demonstration of (B) cross-analysis result, (C) tag to gene results, and (D) extract result using GA.

Relevant genes in huge output genes can be extracted with the "reducing-analysis" function. After inputting huge sample data in a defined format, the GA function provides a class labeling selection, e.g. cases and controls, for each sample, and the representative tags are output with accurate evaluation. Setting a higher population and a higher number of generations (GA parameters) results in higher performance (higher accuracy and fewer genes).

Concluding remarks:

Extract-SAGE constitutes a novel, effective and accurate SAGE analysis platform for comparison of multiple libraries. Common or tissue- and cancer-specific biomarkers can easily be mined *in silico*.

Acknowledgements:

This work was partly supported by the National Science Council in Taiwan under grant NSC97-2622-E-151-008-CC2, NSC96-2221-E-214-050-MY3, NSC96-2622-E-151-019-CC3, and KMU-EM-97-1.1b.

References:

[1] V. E. Velculescu *et al.*, *Science*, 270, 484 (1995) [PMID: 7570003]

[2] P. Liang *et al.*, *Nucleic Acids Res.*, 99, 11547 (2002) [PMID: 12195021]

[3] P. Divina *et al.*, *Nucleic Acids Res.*, 32, D482 (2004) [PMID: 14681462]

[4] http://www.sun.com/java/

[5] A. E. Lash *et al.*, Genome Research 10, 1051 (2000) [PMID: 10899154]

[6] R. P. Grimaldi, "Discrete and Combinatorial Mathematics: An Applied Introduction (IV ed.)", Addison Wesley Publishing Company, (1998)

[7] M. L. Raymer *et al.*, *IEEE Trans. Evolutionary Computation* 4, 164 (2000)

Edited by P. Kangueane

Citation: Yang *et al.*, Bioinformation 3(7): 291-292 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(7): 291-292 (2009)