NONNEGATIVE LEAST SQUARE – A NEW LOOK INTO SAGE DATA

Erliang Zeng

Department of Computer Science
Center for Computational Science
University of Miami
Miami, FL 33146, USA
Email: zeng@cs.miami.edu

Mitsunori Ogihara

Department of Computer Science
Center for Computational Science
University of Miami
Miami, FL 33146, USA
Email: ogihara@cs.miami.edu

Serial Analysis of Gene Expression (SAGE) is a technology for quantifying gene expression using sequencing of short stretches (tags) of DNA that are produced by reverse transcription and enzymatic restriction. A major issue in SAGE data analysis is ambiguity of tags, i.e., single tags matching multiple genes and single genes matching multiple tags. The ambiguity produces groups of interrelated quantitative constraints among tag counts and gene expression values. We propose to solve the web of relations between tags and genes using nonnegative least square (NNLS) method. In this paper we present a fast algorithm to do this task. The effectiveness of the method is confirmed by examining a published data that involves SAGE and a method called GLGI. The method is then applied to a SAGE data for a human neurodegenerative disease. The experimental results show that more reliable gene expression can be inferred from SAGE tags using our method, suggesting that our method is powerful for exploring gene expression patterns and identifying candidate genes from SAGE data that potentially contribute to the susceptibility of human complex disease.

1. INTRODUCTION

Gene expression profiling is widely used for exploring genome-wide gene activity patterns [1–3]. Microarrays and Serial Analysis of Gene Expression (SAGE) are two major techniques used for surveying gene expression profile. Microarrays measure gene expression levels by using probe-target hybridization [2, 3], where a probe is a short stretch of DNA, often derived from the sequence of a gene, and a target is a cDNA sample that is expected to represent the complete gene activity in a cell. The SAGE, in contrast, obtains gene expression by counting thousands of polyadenylated transcripts by sequencing concatemers of short sequence tags (10bp or 17bp long) derived from biological samples [4]. Both microarrays and SAGE have been used in biological research for many years for measuring the expression of a large number, if not all, of the genes in a given sample. There are some advantages of using SAGE over the other [5, 6]. First, the quantification in SAGE is based on the actual RNA sequences expressed in the sample and thus, unlike microarrays, one does not need to know beforehand the sequence to be measured. This enables one to discover unknown genes. Second, SAGE is able to detect small changes in expression levels, making it more sensitive comparing to microarray [7]. Finally, SAGE can detect over- and under-expressed transcripts equally well. This means it has less biases than microarray [8].

Despite these advantages, one big challenge exists in SAGE data analysis — to estimate gene expression values from SAGE tag counts. Due to technical limitations, the current SAGE technique can detect transcript segments (tags) of 10bp or 17bp long only. Even 17bp tags are not long enough to uniquely represents a gene. That is, for many a tag more than one candidate gene exists whose cDNA corresponds to it, and for many a gene more than one tag is identified that matches it. One crucial step in analyzing SAGE data is production of the so-called Tag-to-Gene assignments, which is to annotate tags with original genes they come from. This is carried out by comparing the tags to a database of virtual tags extracted from known transcript sequences. When single tag matches multiple genes or single gene matches multiple tags, the problem of ambiguity occurs. Most current research only focuses on those tags that can be uniquely mapped to genes and ignores many tags with ambiguous mapping. Although this strategy has been standardly used, much information is undoubtedly lost by its stringent restriction. We in this paper answer how to remove this restriction and make good use of all the tags.

The reference database plays a key role in the Tag-to-Gene assignment step. Because of high degree of redundancies among transcript sequences, it is difficult to use the sequences in transcript sequence databases directly for SAGE tag annotation. The UniGene project (`http://www.ncbi.nlm.nih.gov/UniGene`) is an experimental system for automatically partitioning GenBank transcript sequences (e.g., proteins, well-characterized mRNA/cDNA sequences and ESTs) into a non-redundant set of gene-oriented clusters. Each UniGene cluster fundamentally contains a set of transcript sequences which appear to come from the same transcription locus, and therefore potentially represents a unique transcript. The SAGEmap data repository is constructed so as to make a SAGE tag mapped to genes using UniGene cluster identifiers [9]. The construction process of the tag to UniGene cluster assignments (tag-UniGene assignments) itself is an automated process consisting of multiple computational steps. The result of this process is a "full" tag to gene mapping called "SAGEmap full" which includes the whole extracted virtual SAGE tags. Also, "SAGEmap reliable" is constructed by using SAGE tags extracted from high-quality sequences in "SAGEmap full". Although "SAGEmap reliable" provides more reliable information for SAGE annotation, the scope of transcriptomes that "SAGEmap reliable" covers is much less than that of "SAGEmap full." SAGEmap provides an automatic link between gene names and SAGE transcript levels, accounting for alternative transcriptions and many potential errors. SAGEmap is powerful, but there are additional ways of processing and presenting this valuable data. SAGE Genie, a set of tools for processing SAGE data, is then developed [10]. The foremost of these tools is the SAGE Anatomic Viewer, which allows nearly any gene's transcript level to be easily viewed in normal and malignant tissues.

All this tremendous amount of development effort notwithstanding, SAGE data is still noisy. The noise in SAGE data largely comes from two sources: sequencing errors and Tag-to-Gene mapping ambiguity. As to the former, it is conjectured that the unmapped tags could largely result from an accumulation of sequencing errors [6]. Considering that many steps are involved in SAGE tag collection, and in particular the errors introduced by single-pass DNA sequencing, many SAGE tags are expected to contain base errors and thus cannot be reliably mapped to their known transcripts. This problem is particularly serious for SAGE tags with lower copy numbers. Therefore, it was suggested to eliminate unmapped tags from further analysis. However, we argue that even for tags that are mapped to genes, sequencing errors still exist. This issue unfortunately has not received much attention [6]. As to the latter, the problem is that some genes mapped to a tag may not be the true gene origins for that tag because the length of SAGE tags is limited and transcript sequences that appear in SAGE reference databases are highly heterogeneous.

Attempts have been made to resolve these issues. "Long SAGE" attempts to reduce the noise caused by redundancies by extending the capability of original SAGE by sequencing extra 7 base pairs, which allows a high percentage of long SAGE tags to be mapped directly to genomic sequence data [11–13]. Other attempts using computational and experimental approaches have been made for solving this problem [14–16]. Ge *et al.* used microarray expression data from different tissue types to define contexts of gene expression and to predict the original transcript contributing a ambiguous tag [14]. Chen *et al.* identified the correct genes for SAGE tags by extending the SAGE tags into $3'$ complementary DNAs (cDNAs) using of the GLGI technique (generation of longer cDNA fragments from SAGE tags for gene identification) [15]. Griffitha *et al.* performed global coexpression analysis by assessing and integrating publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data [16].

In this paper, we propose to use NonNegative Least Squares (NNLS) methods to analyze SAGE data. Our goal is to predict gene expression levels from tag counts using Tag-to-Gene assignments information as much as possible and address noise problems in the meantime. This is obtained using the information of interrelated quantitative constraints among tag counts and gene expression values.

2. METHODS

In this section, we first introduce the basic concepts behind our NNLS method and its link to SAGE data analysis. Then, we present an improved algorithm to solve the NNLS problem.

2.1. NNLS problem and its link to SAGE data analysis

The NNLS problem refers to the problem of finding, given an $n \times m$ nonnegative matrix $A = (a_{ij})$ and a nonnegative n-dimensional vector $b = (b_i)$, a nonnegative m-dimensional vector $x = (x_i)$ that minimizes the functional $f(x) = \frac{1}{2}\|Ax - b\|^2$, *i.e.*,

$$\begin{aligned}\min_x f(x) &= \frac{1}{2}\|Ax - b\|^2 \\ &= \frac{1}{2}\sum_{i=1}^{n}(a_{i1}x_1 + \cdots + a_{im}x_m - b_i)^2, \\ &\text{subject to } x \geq 0.\end{aligned}$$

The problem of estimating gene expression levels from the tag counts of SAGE data can be formulated as a NNLS problem. Suppose there are n tags that appear in the data and there are m genes that at least one of the tags matches. Let A be a binary (i.e., 0/1) $n \times m$ matrix so that for all rows i and for all columns j the (i, j) entry of the matrix, a_{ij}, is 1 if and only if the i-th tag mapped to the j-th gene, and 0 otherwise. Let b be an n-dimensional positive vector such that for all i the i-th entry of b, b_i, is the count of tag i in the SAGE data. Let x be the m-dimensional vector of unknowns such that for all j the j-th entry of x, x_j, represents the expression of the j-th gene. Our problem is then to estimate vector x in terms of the NNLS problem.

The $n \times m$ nonnegative matrix A can be viewed as the connectivity matrix of the bipartite graph G over n tags and m genes, where tag i and gene j are connected by an edge if tag i matches gene j. Normally, A is of high dimension with over ten thousand tags and genes. This makes the computation of a solution to the NNLS problem very time-consuming. One way to reduce the dimensionality is to divide the bipartite graph G into connected components and then solve the NNLS problem on each individual connected components represented, since solutions for a connected component do not interfere with those for another component. As illustrated in Figure 1, the original matrix A is a 8×10 matrix. After graph partition, matrix A is divide into two matrices with dimensionality 4×7 and 4×3, respectively.

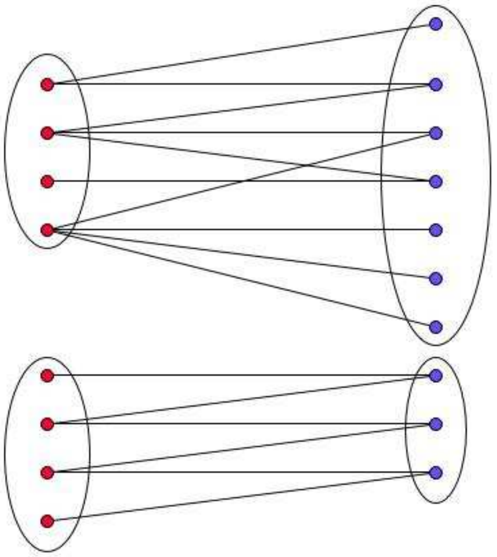


Fig. 1. An example showing the bipartite graph over a set of tags and a set of genes. The vertices on the left-hand side represent a set of tags and the vertices on the right-hand side represent a set of genes.

The advantages of formulating the assignments of ambiguous SAGE tags to genes as an NNLS problem are two-fold. First, the new formulation enables us to perform more comprehensive transcriptome analysis since ambiguous tags are used in the analysis in a meaningful manner. Second, with this new formulation it is possible to improve the accuracy and reliability in SAGE data analysis. Because the new formulation takes a full account of the inventory of tages that are expressed and their mapping to genes. The expression level of a gene is estimated based not just on unique sequence tags, but on all sequence tags it links to. The contribution of tags that have sequence errors to the prediction of gene expression level is thus diminished.

2.2. Algorithm for solving NNLS problem

The standard method (also known as the first method) for solving NNLS is proposed by Lawson *et al.* [17]. The method uses an iterative procedure that repeatedly identifies the dimension along which the gradient of $\|Ax-b\|^2$ is the smallest. This greedy method is fast but does not necessary produce close-to-optimal solutions. To overcome this issue the Project Gradient Method is devised [18], where instead of a single dimension, a projection along which the gradient is the smallest. A variant of this approach has been recently proposed [19], where the amount of changes along the projected dimension exhibits behavior similar to the Newton method. The full implementation of this unfortunately requires the use of non-sparse $m \times m$ scaling matrix that is updated at every iteration. Since m is in the order of tens of thousands in our case, the full-fledged version of the algorithm is impractical. We thus modify the algorithm so that a fixed scaling matrix (e.g., the identity scaling) is used.

In the following for an n-dimensional vector x, $\nabla f(x)$ denotes the m-dimensional vector $A(Ax-b)$ and $P(x)$ denotes the vector constructed from x by setting all negative entries to 0. We set the initial value of x to the n-dimensional 0-vector. Then repeat the following loop until a forced convergence condition is reached (the improvement in $\|Ax-b\|^2$ is smaller than a pre-determined threshold).

(1) Set I to the set of all indices between 1 and n such that the i-th entry of x is 0 and the i-th entry of $\nabla f(x)$ is positive. Set J to the set of all indices between 1 and n that are not in I.

(2) Decompose x as $y+z$, where y is the x with all the entries at positions in I set to 0 and z is the x with all the entries at positions in J set to 0.

(3) Repeat Compute two quantities α and β:

  (a) Set $\bar{A}$ to the matrix constructed from A by setting 0 all the columns and rows whose indices are in I. For an n-dimensional vector u, let $g(u) = \frac{1}{2}\|\bar{A}u - b\|^2$.

  (b) Compute the smallest nonnegative integer μ that satisfies

$$g(y) - g(P(y - (1/2)^{\mu-1})$$
$$\geq \frac{1}{4}(\bar{A} - b)^T \bar{A}^T (y - P(y - (1/2)^{\mu-1} g(y)))$$

  (c) Set β to $(1/2)^{\mu-1})$ and v to $P(y - (1/2)^{\mu-1} g(y))$.

  (d) Set α to the argmin of $g((1-\alpha)y + \alpha v$. If $\alpha > 1$, set α to 1.

(4) Set w to $P(y - \beta \nabla f(y))$.

(5) Set $\tilde{y}$ to $\alpha(w - y)$.

(6) Set x to $\tilde{y} + z$.

3. EXPERIMENTAL RESULTS

3.1. Data sets

The virtual SAGE tags with assigned UniGene clusters were downloaded from the SAGEmap database (`ftp://ftp.ncbi.nlm.nih.gov/pub/sage/mappin-gs`). The NNLS method was performed on two sets of data: the human CD34+ hematopoietic cell SAGE library (`http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE2346`) [20] and the SAGE libraries of hippocampus samples in Alzheimer's disease (`http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSM154136`) [21]. For each SAGE library, a bipartite graph was generated using the experimental tags of the library and the Tag-to-Gene assignment indicated by "SAGEmap full".

3.2. Algorithm performance

Many real world networks are scale-free, that is, the degree distribution follows a power-law. One property of scale-free network is that the clustering coefficient is high — the network tends to be highly locally-connected. This is also true for the network formed by the tags and UniGenes. For example, the SAGE libraries of hippocampus samples in Alzheimer's disease produce a very large connected component (20K+ × 40K+) and this connected component slows down the calculation. It is thus necessary to break down such huge component into several smaller connected components using graph partitioning algorithm. In such graph partitioning, the input graph is divided into components by removal of edges. In our case the removal of an edge corresponds to the elimination of the information contained in the link between a tag and its matching UniGene. Breaking down the input graph into components might thus be harmful. We tested the NNLS algorithm performance on different partitions generated by METIS graph partition packages [22] for the aforementioned large connected component. The algorithm performance is measured by the mean squared error (MSE) of the tag counts. As shown in Figure 2, the algorithm performance does not change much when the number of partitions increases (bottom panel), while the running time is dramatically reduced when the number of partitions is 500 and above (top panel). This suggests that, whenever necessary, it is suitable to execute partition before employing NNLS calculation.

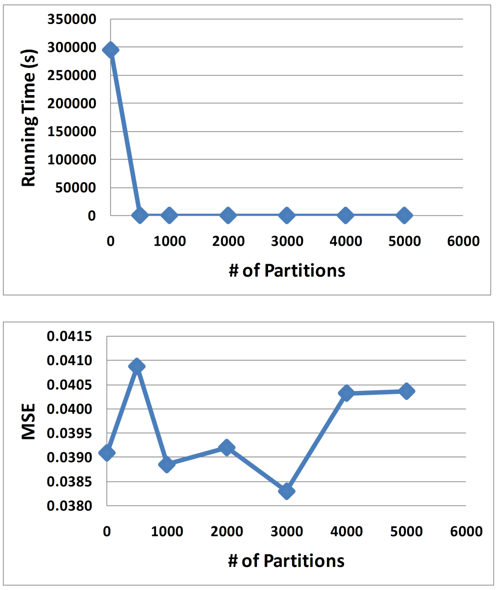


Fig. 2. NNLS performance on graph decomposition.

3.3. Evaluation using SAGE data of hematopoietic cells

3.3.1. *Evaluation of NNLS method*

The accuracy of NNLS method was evaluated using SAGE data from CD34+ hematopoietic cells. Zhou *et al.* analyzed the pattern of gene expression in human primary CD34+ stem/progenitor cells using SAGE approach [20]. Among 21,546 tags that matched known expressed sequences, 34% matched more than one UniGenes. Zhou *et al.* employed a method called GLGI to convert the tags that have multiple matching genes into 3′-ESTs and then used these longer sequences to search in databases their corresponding genes [15]. This is a wet-lab strategy to annotate tags that have multiple gene matches.

We employed our NNLS method on CD34+ SAGE data and compared the results against the results published in the paper by Zhou *et al.* [20]. Because the SAGE tags with multiple matches tend to be the ones with more copies, we compared the results for tags with more than 50 copies that have multiple matches. There are 98 tags fall into this category. After the GLGI annotation, many tags in this category are found to come from housekeeping genes and from genes with unknown functions, including 40 ribosomal proteins, 2 hypothetical genes, and 23 ESTs. We focused on comparing our results with those genes that have specific functions, which consist of 19 tags as shown in Table 1. For each tag, the UniGenes it matches are ranked in the decreased order of their estimated gene expression values, and are then compared against the annotation by the GLGI method. The top annotations of thirteen tags by the NNLS method are confirmed by the GLGI method. When the top two UniGene clusters are included, five more annotations are confirmed, which represent 95% total of the tags that are compared (Table 1).

3.3.2. *Annotations of function unknown tags*

Note that all the tags in Table 1 have more than one matching UniGene clusters. Each of those UniGene clusters in turn has multiple matching tags, thus form a complicated tag-gene network. Table 1 shows that the top UniGene cluster assigned to each ambiguous tag is likely to be the true gene origin. Considering that there are many tags mapped to ESTs without known function when the work of Zhou *et al.* was published, it is possible to assign UniGene clusters to those tags using our NNLS method and the new version of “SAGEmap full” Tag-to-Gene assignments, as UniGene builds are constantly updated. Table 2 shows some examples of these new cluster asignments. These UniGene clusters are involved in basic cellular activity such as ribosomal proteins. The observations are consistent with the knowledge that many housekeeping genes are highly expressed in CD34+ cells. In addition, some tags are assigned with specified functions with high expression level, indicating their potential important role for hematopoiesis.

Table 1. Evaluation of the NNLS method.

SAGE tag	Count	# of UniGene[a]	# of Tag[b]	UniGene[c]	UniGene[d]	Gene Symbol[e]	Rank[f]
TGTGTTGAGA	1711	6	8	Hs.181165	Hs.644639	EEF1A1	1
GTGAAACCCT	239	155	2	Hs.184376	Hs.539304	SNAP23	0
CCAGAGAACT	165	3	15	Hs.6975	Hs.642877	MALAT1	2
TAGGTTGTCT	148	3	8	Hs.279860	Hs.374596	TPT1	1
ATTGTTTATG	127	8	2	Hs.181163	Hs.181163	HMGN2	1
GCTCCCCTTT	126	2	6	Hs.1817	Hs.458272	MPO	2
TGTAATCAAT	117	3	5	Hs.249495	Hs.546261	HNRNPA1	2
TCACAAGCAA	109	3	5	Hs.32916	Hs.505735	NACA	1
GGGCATCTCT	102	2	2	Hs.76807	Hs.520048	HLA-DRA	1
CTCATAGCAG	84	2	8	Hs.279860	Hs.374596	TPT1	1
AAAAGAAACT	76	2	9	Hs.172182	Hs.387804	PABPC1	1
GCTTTATTTG	75	5	4	Hs.288061	Hs.520640	ACTB	1
GCCTTCCAAT	74	2	4	Hs.76053	Hs.279806	DDX5	1
TACCATCAAT	72	4	1	Hs.169476	Hs.544577	GAPDH	1
GCATTTAAAT	64	5	2	Hs.275959	Hs.421608	EEF1B2	1
GTCTGGGGCT	63	5	2	Hs.75725	Hs.517168	TAGLN2	2
TCTGCTAAAG	58	5	2	Hs.274472	Hs.593339	HMGB1	1
GTTCCCTGGC	55	4	2	Hs.177415	Hs.387208	FAU	1
CCTAGCTGGA	50	7	6	Hs.182937	Hs.356331	PPIA	2

[a] Number of UniGenes matching SAGE tag in column 1.

[b] Number of SAGE tags matching UniGene as shown in column 7.

[c] UniGene ID in the paper of Zhou *et al.* [20] verified by GLGI method.

[d] UniGene ID extracted from SAGEmap build #218 and ranked as top candidate by NNLS method.

[e] Gene Symbol corresponding to UniGene as shown in column 5 or column 6. Because the UniGene database is updated constantly, although the identifiers in column 5 and column 6 are different, they refer to as the same gene symbol.

[f] The rank of gene as shown in column 7. The ranking is obtained based on the decreased order of gene expression values estimated by NNLS method for all the UniGenes the corresponding tag matches. 0: the gene has no ranking; 1: the gene ranked top 1; 2: the gene ranked second.

3.4. Results of SAGE data of human brain samples

The results in Section 3.3.2 demonstrate the effectiveness of NNLS method for SAGE data analysis. The NNLS method was then applied to another SAGE library — SAGE library of human Alzheimer's disease.

*APOE*4 allele is a major risk factor for late-onset Alzheimer's disease (AD). However, the mechanism of action of *APOE* in AD remains unclear. Xu *et al.* studied the effects of APOE alleles on gene expression in AD by analyzing SAGE data obtained from human hippocampus samples with AD patients with *APOE*3/3, *APOE*3/4, and *APOE*4/4 alleles and samples from a control [21]. We re-analyzed this SAGE data set using our NNLS method with the goal of finding new genes that contribute to the susceptibility of Alzheimer's disease.

3.5. Differentially expressed genes in AD

Xu *et al.* used Chi-square Test and Fisher's Exact Test to examine the difference in tag counts between two samples [21]. We performed the same statistical analysis but on the expression values obtained from NNLS method. Table 3 summarizes the results from the comparison of two studies. Xu *et al.* found that gene expression patterns in the hippocampus tissues of *APOE*3/4 and *APOE*4/4 AD patients differ substantially from those of *APOE*3/3 AD patients. *APOE*3/4 and *APOE*4/4 allele expressions may activate similar genes or gene pools with associated functions. Our results show similar expression patterns and confirm the conclusions in the study of Xu *et al.*[21]. Actually, for each sample, almost one half of the significant UniGenes found by our NNLS method overlap with the discoveries by Xu *et al.* [21].

Table 2. Annotations for tags with unknown function.

SAGE tag/UniGene	Count	Symbol & Description
CCTGTAATCC	516	Hypothetical protein
Hs.433701	57	RPL37A Ribosomal protein L37a
Hs.527193	26	RPS23 Ribosomal protein S23
Hs.489190	13	SLC25A13 Solute carrier family 25, member 13 (citrin)
GTGAAACCCC	471	EST
Hs.477789	11	ATP1B3 ATPase, Na+/K+ transporting, beta 3 polypeptide
Hs.194236	8	LEP Leptin
Hs.713420	7	LOC100130236 CDNA FLJ46301 fis, clone TESTI4036012
CCACTGCACT	331	EST
Hs.683922	46	C8orf54 CDNA FLJ35455 fis, clone SMINT2004547
Hs.709365	5	RIOK3 RIO kinase 3 (yeast)
Hs.289123	5	DCTN2 Dynactin 2 (p50)
GCCTCAGTTC	256	EST
Hs.631498	102	Transcribed locus, strongly similar to XP_001082381.1
Hs.631499	44	Transcribed locus, strongly similar to NP_536851.1
Hs.586920	29	Transcribed locus, strongly similar to NP_001090497.1
AATGGATGAA	253	EST
Hs.680613	253	CDNA FLJ30447 fis, clone BRACE2009270
TGATTTCACT	138	EST
Hs.451549	46	FLJ44313 FLJ44313 protein
Hs.129283	46	CDNA clone IMAGE:5265638
Hs.703561	46	LOC100131532 Transcribed locus NP_536849.1
GCAAGCCAAC	132	EST
Hs.703684	121	LOC100131532 Transcribed locus, weakly similar to NP_536852.1
Hs.704111	11	FAM78A Transcribed locus NP_536852.1
AGGTCAGGAG	122	EST
Hs.489190	13	SLC25A13 Solute carrier family 25, member 13 (citrin)
Hs.352768	10	PSMB1 Proteasome (prosome, macropain) subunit, beta type, 1
Hs.620557	4	ANK2 Ankyrin 2, neuronal
ACCCTTGGCC	122	EST
Hs.631495	49	Transcribed locus, strongly similar to NP_536843.1
AAGGTGGAGG	118	EST
Hs.585012	115	WTIP Transcribed locus, strongly similar to XP_001724292.1
Hs.699463	1	CEBPA CCAAT/enhancer binding protein, alpha
Hs.337766	1	RPL18A Ribosomal protein L18a
TTGGCCAGGC	109	EST
Hs.370504	9	RPS15A Ribosomal protein S15a
Hs.651923	8	CNN2 Calponin 2
Hs.130293	7	CROP Cisplatin resistance-associated overexpressed protein
CCTGTAGTCC	100	EST
Hs.368960	5	NGLY1 N-glycanase 1
Hs.325978	4	NUMA1 Nuclear mitotic apparatus protein 1
Hs.631616	4	LOC147727 Hypothetical LOC147727
ACTTTTTCAA	76	EST
Hs.659985	69.99	XYLB Xylulokinase homolog (H. influenzae)
Hs.557644	5	Transcribed locus
Hs.383050	1	Transcribed locus, weakly similar to NP_001077.2 arylacetamide deacetylase

Table 3. Comparison of differentially ($p < 0.05$) expressed SAGE tags and UniGenes[*].

SAGE library	Sig. tags(%)	Sig. UGs	Uniq. UGs[a]	Sig. UGs	Comm. UGs[b]	Uniq. Comm. UGs[c]
		(Up/Down)	(Up/Down)	(Up/Down,%)	(Up/Down)	(Up/Down)
E33AD vs Ctl	156(0.41)	276(114/162)	100(48/52)	43(21/22,0.38)	24(9/15)	16(3/13)
E34AD vs Ctl	906(2.32)	1594(714/880)	523(97/426)	306(115/191,2.86)	189(43/146)	109(11/98)
E44AD vs Ctl	625(1.71)	1270(847/423)	320(108/212)	215(98/117,2.10)	88(25/63)	41(3/38)
E34AD vs E33AD	918(2.26)	1673(848/825)	499(131/368)	329(144/185,3.08)	189(53/136)	99(16/83)
E44AD vs E33AD	771(2.01)	1469(934/535)	428(172/256)	218(120/98,2.12)	94(31/63)	45(9/36)
E44AD vs E34AD	476(1.24)	761(533/208)	287(188/99)	114(59/55,1.11)	62(35/27)	34(17/17)

[*] Columns 2 to 4 shows results from Xu *et al.* [21], including significant tags and the percentage of all discovered tags, significant UniGenes that significant tags match, and unique significant UniGenes. Column 5 shows the significant UniGenes and the percentage of all discovered UniGenes obtained by our analysis. Columns 6 and 7 shows the common UniGenes shared by two methods. The numbers X/Y in the parentheses show the breakdown between the up-regulated and down-regulated tags.

[a] The *Unique UniGene* refers to the one used in the paper of Xu *et al.* [21], *i.e.*, the tags that map to only one UniGene cluster.

[b] These are the UniGenes shared by Column 4 and Column 5.

[c] The *Unique UniGene* refers to a UniGene all of whose tags match the UniGene only.

Table 4. Significant genes uniquely identified by NNLS method known to be susceptibility candidates to AD

UniGene	Symbol	SAGE library	Expression[a]
Hs.551642	ANXA8	E34AD vs Ctl	Down
Hs.551642	ANXA8	E34AD vs E33AD	Down
Hs.551642	ANXA8	E44AD vs Ctl	Down
Hs.551642	ANXA8	E44AD vs E33AD	Down
Hs.546241	C4A	E34AD vs Ctl	Up
Hs.546241	C4A	E34AD vs E33AD	Up
Hs.546241	C4A	E44AD vs E34AD	Down
Hs.522555	APOD	E34AD vs Ctl	Up
Hs.522555	APOD	E34AD vs E33AD	Up
Hs.391561	FABP4	E44AD vs Ctl	Up
Hs.391561	FABP4	E44AD vs E34AD	Up
Hs.414795	SERPINE1	E44AD vs Ctl	Up
Hs.414795	SERPINE1	E44AD vs E33AD	Up
Hs.511367	CYP19A1	E34AD vs Ctl	Up
Hs.514220	GRN	E44AD vs Ctl	Up
Hs.348387	GSTM4	E44AD vs Ctl	Down

[a] *Up* means "Up-regulated", *Down* means "Down-regulated".

Significant differences between the two studies are shown in Table 3. For example, the number of significant UniGenes from our analysis is much smaller than those discovered by Xu *et al.*[21]. Xu *et al.* asserted that a UniGene is *unique* if there is at least one significantly expressed tag that matches the gene and there are no other UniGenes that match such a tag. A drawback of this definition is that it does not take into account all the expressed tags that match the gene and thus may be prone to errors. In our analysis, we can avoid the use of this concept of *uniqueness* since the expressions of the tags and of the genes are estimated by taking into account matching relations among them. Because of this our results may be more reliable than those by Xu *et al.*. The same argument can be made from the results in Table 4, which shows the genes that are found to be candidates for susceptibility to AD only in our study. This gene set is obtained by comparing our significant genes to the genes listed in AlzGene database (`http://www.alzforum.org/res/com/gen/alzgene/`) for each SAGE library. The AlzGene database provides a comprehensive, unbiased and regularly updated collection of genetic association studies performed on Alzheimer's disease phenotypes. Some genes, such as $ANXA8$ and $C4A$, appear in multiple SAGE libraries. That seems to suggest their potential important role in the development of Alzheimer's disease.

The detailed Venn diagrams of the overlaps and differences between our results and Xu's results are shown in Figure 3. As shown in the figure, the differentially expressed genes shared between $APOE3/4$ and $APOE4/4$ AD patients are much more than those shared between $APOE3/4$ and $APOE3/3$ AD patients, and between $APOE4/4$ and $APOE3/3$ AD patients.

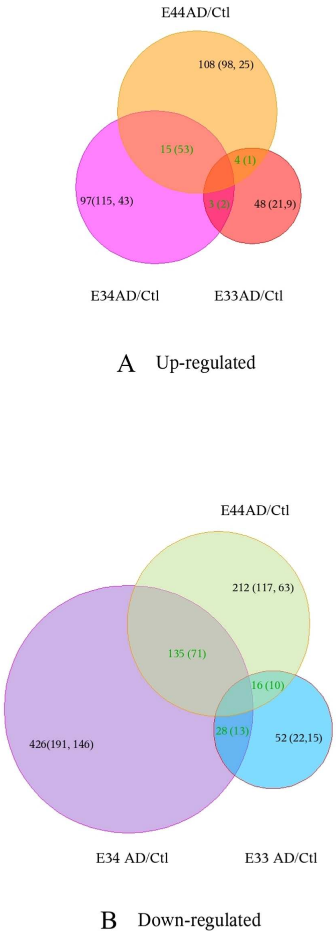
With respect to several gene functional categories, the expression profiles that our analysis identifies are similar to those discovered by Xu *et al.* (data not shown). For example, $APOE4$ AD alleles activate multiple tumor suppressors, tumor inducers and negative regulator of cell growth or repressors that may lead to increased cell arrest, senescent and apoptosis. In contrast, expression is decreased for large clusters of genes associated with synaptic plasticity, synaptic vesicle trafficking (metabolism) and axonal/neuronal outgrowth. In addition, reduction of neurotransmitter receptors and Ca^{2+} homeostasis, disruption of multiple signal transduction pathways, and loss of cell protection and notably mitochondrial oxidative phosphorylation/energy metabolism are associated with $APOE3/4$ and $APOE4/4$ AD alleles.

4. Discussion and Conclusions

SAGE data analysis based on NNLS method have been performed on two sets of data: a human CD34+ hematopoietic cell SAGE library and a human hippocampus cell SAGE library. Previous research in this area focused on investigating genes that can be uniquely matched SAGE tags, and thus ignores the tags that have multiple matches. This limits the range of genes for which expression patterns are analyzed. To break the barrier, one must accurately estimate the gene expression levels based on all the tags counts. Although it is possible to obtain more reliable annotate SAGE tags using "SAGEmap reliable" Tag-to-Gene assignments, the analysis scope is largely decreased. Previous study showed that 38.1% of the virtual tags in the "SAGEmap full" contain more than one UniGene cluster, even in "SAGEmap reliable" subdatabase, 12.7% contain more than one UniGene cluster [14]. Thus there is a pressing need to develop new computational method for SAGE data analysis that fully uses the tag counts and interrelated quantitative constraints among tags and genes.

The research described in this paper addresses the above challenges. To the best of our knowledge, there have been few attempts along our line. Our significant contributions are:

(1) A nonnegative least squares method has been introduced to SAGE data analysis. The expression levels of genes have been estimated from the counts of all tags presented in a cell. This NNLS method has been evaluated with another wetlab SAGE tag annotation method, GLGI.
(2) A new algorithm to solve NNLS problem has been proposed. The new algorithm uses a fixed scaling matrix so as to alleviate compute intensity. Also the algorithm uses graph partitioning.
(3) The newly proposed NNLS algorithm enables systematic, consistent, and comprehensive analysis of SAGE data. The strategy has been employed to identify new candidate genes for susceptibility to Alzheimer's desease.



A Up-regulated

B Down-regulated

Fig. 3. Comparison of regulation of gene expression by *APOE* genotype. $N(n_1, n_2)$ means the number of unique significant genes discovered by Xu *et al.* is N, the number of significant genes discovered by us is n_1, and the number common significant genes shared by two studies is n_2. The diagram only show the number of genes in each area. (A) Venn diagram of the UniGenes that are discovered in $APOE3/3$ AD, $APOE3/4$ AD and $APOE4/4$ AD samples and are upregulated from the control. (B) Venn diagram of the UniGenes that are discovered in $APOE3/3$ AD, $APOE3/4$ AD and $APOE4/4$ AD samples and are down-regulated from the control.

Transcriptomics studies using next-generation sequencing technologies have received intensive attention recently. One promising application is to quantify gene expression and alternative splicing using new sequencing technologies. In spite of rapidly evolving, these new technologies have relatively high error rate comparing to conventional sequencing technology. One challenge in next-generation sequencing data analysis is to align sequence reads to referance genome. Like SAGE data, the sequence reads and genes have ambiguous mapping. Thus the NNLS method described in this paper provides a potential solution for analyzing next-generation sequencing data.

References

1. DeRisi, J., Iyer, V., Brown, P.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278**(5338) (1997) 680–686
2. Zhang, M.: Large-scale gene expression data analysis: a new challenge to computational biologists. Genome Research **9**(12) (1999) 681–688
3. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell **9**(12) (1998) 3273–3297
4. Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W.: Serial analysis of gene expression. Science **270**(5235) (October 1995) 484–487
5. Evans, S.J., Datson, N.A., Kabbaj, M., Thompson, R.C., Vreugdenhil, E., Kloet, E.R.D., Watson, S.J., Akil, H.: Evaluation of affymetrix gene chip sensitivity in rat hippocampal tissue using SAGE analysis. European Journal of Neuroscience **16**(3) (2002) 409–413
6. van Ruissen Fred, Jan, R., Gerben, S., Lida, A., Danny, Z., Marcel, K., Frank, B.: Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. BMC Genomics **6**(1) (2005) 91
7. Scott, H.S., Chrast, R.: Global transcript expression profiling by serial analysis of gene expression (SAGE). Genetic Engineering: Principles and Methods. **23** (2001) 210–219
8. Iyer, V., Struhl, K.: Absolute mrna levels and transcriptional initiation rates in saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America **93**(11) (1996) 5208–5212
9. Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J., Altschul, S.F.: SAGEmap: a public gene expression resource. Genome Res **10** (2000) 1051–1060
10. Boon, K., Osrio, E.C., Greenhut, S.F., Schaefer, C.F., Shoemaker, J., Polyak, K., Morin, P.J., Buetow, K.H., Strausberg, R.L., de Souza, S.J., Riggins, G.J.: An anatomy of normal and malignant gene expression. Proceedings of the National Academy of Sciences of the United States of America **99**(17) (2002) 11287–11292
11. Li, Y.J., Xu, P., Qin, X., Schmechel, D.E., Hulette, C.M., Haines, J.L., Pericak-Vance, M.A., Gilbert, J.R.: A comparative analysis of the information content in long and short SAGE libraries. BMC Bioinformatics **7** (2006) 504–514
12. Gowda, M., Jantasuriyarat, C., Dean, R.A., Wang, G.L.: Robust-LongSAGE (RL-SAGE): A Substantially Improved LongSAGE Method for Gene Discovery and Transcriptome Analysis. Plant Physiol. **134**(3) (2004) 890–897
13. Hanriot, L., Keime, C., Gay, N., Faure, C., Dossat, C., Wincker, P., Scote-Blachon, C., Peyron, C., Gandrillon, O.: A combination of LongSAGE with Solexa sequencing is well suited to explore the depth and the complexity of transcriptome. BMC Genomics **9**(1) (2008) 418
14. Ge, X., Wang, S.M.: Identifying nonspecific SAGE tags by context of gene expression. Methods in Molecular Biology **387** (2007) 199–204
15. Chen, J., Lee, S., Zhou, G., Wang, S.M.: High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3' complementary DNAs. Molecular and Cellular Neuroscience **33**(3) (2002) 252–261
16. Griffitha, O.L., Pleasancea, E.D., Fultonb, D.L., Oveisia, M., Esterc, M., Siddiquia, A.S., Jones, S.J.: Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. Genomics **86** (2005) 476–488
17. Lawson, C.L., Hanson., R.J. In: Solving Least Squares Problems, Prentice-Hall (1974)
18. Bierlaire, M., Toint, P.L., Tuyttens., D.: On iterative algorithms for linear least squares problems with bound constraints. In: Linear Algebra and Its Applications. Volume 413. (1991) 111–143
19. Kim, D., Sra, S., , Dhillone, I.S.: A new projected quasi-newton approach for the nonnegative least square problem. In: Technical Report TR-06-54., Department of Computer Science, The University of Texas at Austin (2006)
20. Zhou, G., Chen, J., Lee, S., Clark, T., Rowley, J.D., Wang, S.M.: The pattern of gene expression in human CD34+ stem/progenitor cells. Proceedings of the National Academy of Sciences of the United States of America **98**(24) (2001) 13966–13971
21. Xu, P.T., Li, Y.J., Qin, X.J., Kroner, C., Green-Odlum, A., Xu, H., Wang, T.Y., Schmechel, D.E., Hulette, C.M., Ervin, J., Hauser, M., Haines, J., Pericak-Vance, M.A., Gilbert, J.R.: A SAGE study of apolipoprotein E3/3, E3/4 and E4/4 allele-specific gene expression in hippocampus in Alzheimer disease. Molecular and Cellular Neuroscience **36**(3) (2007) 313–331

22. Abou-Rjeili, A., Karypis, G.: Multilevel algorithms for partitioning power-law graphs. Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International (2006) 10 pp.+