

***MmSAGEClass* — software manual**

An online database for the functional classification of mouse SAGE tags[§]

Katarzyna Rejniak¹ Adrienne Frostholm² Julie Besco²
Magdalena Popesco² Andrej Rotter^{1,2}

¹Mathematical Biosciences Institute

²Department of Pharmacology

The Ohio State University, Columbus, OH 43210

Abstract

Genomic information is growing rapidly and it is becoming increasingly important not only to identify and quantify expressed genes, but also to classify them in a systematic way. We present here a WWW-based tool for the functional classification of murine genes obtained by the SAGE (Serial Analysis of Gene Expression) technology. *MmSAGEClass* (Mus musculus SAGE Classification) assigns each SAGE tag to a specific functional category. Results obtained by using *MmSAGEClass* give an insight into the biological functions of genes expressed in the tissue from which the SAGE library has been constructed.

The online database is accessible at: <http://mbi.osu.edu/~rejniak/MmSAGEClass.html>

Table of contents

1. Introduction	(2)
2. SAGE technique	(2)
3. <i>MmSAGEClass</i> online database	(2)
3.1 Class description	(3)
3.2 Input data format	(3)
3.3 <i>MmSAGEClass</i> algorithm	(4)
3.4 Expected results and data post-processing	(4)
4. Database architecture	(4)
4.1 Primary data	(4)
4.2 Secondary data	(6)
4.3 User query	(7)
4.4 Online interface	(7)
5. Step-by-step guide to online database	(8)
5.1 Data format and size	(8)
5.2 Submitting data	(8)
5.3 Analyzing data	(9)
5.4 Presenting computed results	(10)
6. SAGE tags classification	(11)
6.1 Analysis of the whole library	(12)
6.2 Analysis of the libraries of high and low abundance tags	(14)
7. Summary	(14)

[§] This work was supported by the NSF under Agreement No. 0112050.

1. Introduction

Genomic information is growing rapidly and it is becoming increasingly important not only to identify and quantify expressed genes, but also to classify them in a systematic way. Serial Analysis of Gene Expression (SAGE) is a technique used to detect and quantify both known genes, and novel genes for which no sequence information exists, thus providing comprehensive and global gene expression profiles. SAGE tag frequency is directly proportional to the number of mRNAs originally present in the tissue and, therefore, is a measure of tag abundance (Velculescu et al., 1995). Several computational tools for the analysis of tag abundance in SAGE libraries are available, including *SAGE300*, *SAGE2000* (Velculescu et al., 1995), *SAGEmap* (Lash et al., 2000), *eSAGE* (Margulies and Innis, 2000), and *USAGE* (van Kampen et al., 2000). However, to our knowledge, the fully automated process of tag classification according to biological functions of the corresponding genes is not widely available. Our online tool, *MmSAGEClass* (Mus musculus SAGE Classification), provides a solution to this problem. *MmSAGEClass* was designed to work with SAGE libraries of short (10 base pairs) nucleotide sequence tags and their associated abundances and assigns each tag to a specific ontology class defined by the Gene Ontology Consortium (Ashburner et al., 2000).

The rest of the paper is organized as follows. The SAGE technique is described in section 2. The *MmSAGEClass* online tool is presented in section 3. The architecture of the *MmSAGEClass* is described in section 4. Section 5 contains step-by-step instruction on how to use the online database. Detailed description of a specific application of *MmSAGEClass* is described in section 6.

2. SAGE technique

Serial Analysis of Gene Expression (SAGE) is a technique which allows for a global analysis of gene expression. SAGE has the potential to identify the full set of expressed genes and to provide a digital indicator of message abundance. In contrast to microarray technology, SAGE does not require an 'a priori' selection of gene targets, and therefore is more appropriate for both, detection and quantification of known genes and of novel genes for which no sequence information exists.

The basic principle of SAGE is the isolation of a short cDNA sequence (SAGE tag) from a specific and invariable position within a given mRNA. This 14-15-bp tag is located immediately adjacent to the 3'proximal Nla III restriction site. Its sequence varies according to the particular mRNA from which it was derived. The frequency with which any particular tag is detected is directly proportional to the number of mRNAs originally present in the cell or tissue being studied. Therefore, the number of identical tags detected is a measure of the abundance of the corresponding mRNAs in the original tissue. The fact that only 14-bp tags are being sequenced, rather than the entire cDNA, makes SAGE useful for relatively rapid screening of global gene expression. This is made possible by the ligation of the 14-bp tags end to end to form linear strings of 20-30 tags, which are then subcloned and sequenced. The final product is a SAGE library of 10-bp tags and their abundances.

3. *MmSAGEClass* online database

MmSAGEClass is a secondary database with primary information collected from the following three sources accessible online: National Center for Biotechnology Information (NCBI) SAGEmap (<http://www.ncbi.nlm.nih.gov>), Cancer Genome Anatomy Project (CGAP) Genes (<http://cgap.nci.nih.gov>) and Gene Ontology (GO) Consortium (<http://www.geneontology.org>). The NCBI SAGEmap database relates specific SAGE tags to their corresponding UniGene numbers. The CGAP Genes database matches the UniGene numbers with gene ontology codes provided by the Gene Ontology Consortium. Finally, the GO codes are used to assign data into three basic ontologies: Biological Process, Molecular Function and Cellular Component, and their next level gene ontology classes. The list of all classes employed in the *MmSAGEClass* database, together with the actual number of known genes falling in each class, is listed below. All data used by *MmSAGEClass* are updated periodically, following update schedules of the three primary databases (usually on a monthly basis).

3.1 Class description

The functional classification of the mouse genes obtained by the *MmSAGEClass* is based on the ontology classes defined by the Gene Ontology Consortium (<http://www.geneontology.org>). The whole classification tree contains a few hundred subclasses in up to 7 different levels; the number of genes in each subclass rises from 1 to more than 2000. To make our presentation more clear, we have reduced the number of classes under consideration by restricting our analysis to the first level subclasses only. These led us to consider about 50 classes in three main categories: Biological Processes, Cellular Component and Molecular Function. The whole list of classes together with the current number of genes in each class is presented below:

Ontology		Class name	Total number of genes
Cellular component	1	Cell	2853
	2	cellular component unknown	1
	3	extracellular	273
	4	immunoglobulin complex	0
	5	obsolete	4
	6	unlocalized	71
	7	virion	2
Biological process	8	behavior	7
	9	biological process unknown	3
	10	cellular process	1686
	11	development	347
	12	obsolete	94
	13	physiological processes	2860
	14	Viral life cycle	0
Molecular function	15	anticoagulant activity	0
	16	antifreeze activity	0
	17	antioxidant activity	7
	18	apoptosis regulator activity	34
	19	binding activity	2241
	20	cell adhesion molecule activity	22
	21	chaperone activity	63
	22	chaperone regulator activity	0
	23	cytoskeletal regulator activity	0
	24	defense/immunity protein activity	39
	25	enzyme activity	1496
	26	enzyme regulator activity	124
	27	ice nucleation activity	0
	28	Lysine activity	2
	29	molecular function unknown	17
	30	motor activity	35
	31	nutrient reservoir activity	0
	32	obsolete	258
	33	protein stabilization activity	0
	34	protein tagging activity	1
	35	regulator of establishment of competence for transformation	0
	36	signal transducer activity	901
	37	structural molecule activity	190
	38	surfactant activity	3
	39	Toxin activity	14
	40	transcription regulator activity	338
	41	translation regulator activity	47
	42	transporter activity	472
	43	triplet codon-amino acid adaptor activity	0

3.2 Input data format

The input data submitted to *MmSAGEClass* is a typical SAGE library. An acceptable data format will contain only two columns – one for the SAGE 10-pb tags and the other for their numerical counts. The columns should be separated by tabulation or space signs. The submitted data **cannot** contain any header line. Users can input their data via the Internet either directly to the text window or by specifying the path to the text file containing the SAGE library under consideration. *MmSAGEClass* has been tested successfully with SAGE libraries containing up to 24,000 distinct tags.

3.3 *MmSAGEClass* algorithm

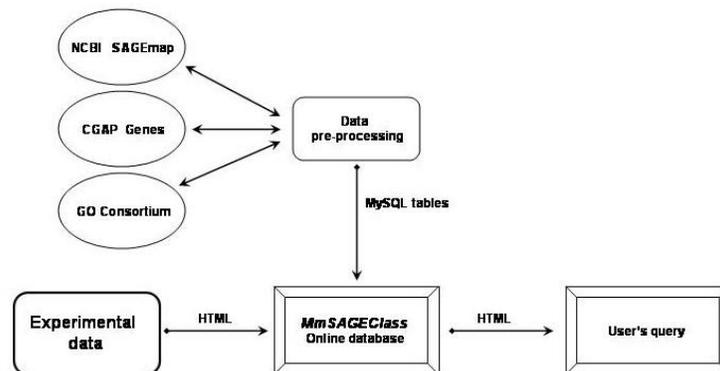
Each tag from the user's library submitted online is matched to the UniGene number obtained from the SAGEmap database. Only those tags that possess a unique UniGene number are used in further analysis. Tags that are not matched to a UniGene number, or that possess more than one UniGene number, are discarded. Subsequently, each remaining tag is assigned to one or more GO number provided by the CGAP resources for mouse tissues. Finally, tags are placed into gene ontology classes defined by the GO Consortium. Tag counts contribute to the class frequency table.

3.4 Expected results and data post-processing

MmSAGEClass returns a frequency table of the number of SAGE tags from the submitted library that fall into the separate gene ontology classes. This table is presented on the computer screen and can be also sent in the space-delineated text format to the email address provided by the user. The return data can be directly input into appropriate software, such as Microsoft Excel, to draw bar or pie charts. A detailed example of functional classification of SAGE tags is presented in section 6.

4. Database architecture

MmSAGEClass is a secondary database with primary information collected from online resources of the National Center for Biotechnology Information (NCBI) SAGEmap (<http://www.ncbi.nlm.nih.gov>), Cancer Genome Anatomy Project (CGAP) Genes (<http://cgap.nci.nih.gov>) and Gene Ontology (GO) Consortium (<http://www.geneontology.org>). All primary data are preprocessed and stored as a MySQL database. User's SAGE data are submitted via the Internet and user's SQL queries for gene classification are executed using HTTP and CGI scripts. Results are displayed on the computer screen and can be also sent to the email address provided by the user. A schematic representation of the *MmSAGEClass* architecture and data flow is shown below



4.1 Primary data

National Center for Biotechnology Information (NCBI) SAGEmap is a public gene expression data repository (Lash et al., 2000) accessible online at: <http://www.ncbi.nlm.nih.gov/sage>. It provides free and open access to raw SAGE data, precomputed tag extractions, and several tools for data analysis. In our database we use the cumulated data deposited in a compressed file: *SAGEmap_tag_ug-rel.zip* at the ftp site: <ftp://ftp.ncbi.nih.gov/pub/sage/map/Mm/NlaIII/>. This file contains mouse SAGE data in the following format:

- (1) 10 base SAGE tag;
- (2) abbreviation of the organism name;
- (3) UniGene cluster number;
- (4) commonly used gene alias (or '-' if gene alias is not available) and UniGene cluster name;
- (5) tag-to-gene mapping reliability score.

We used the NCBI SAGEmap data to relate mouse SAGE tags with the corresponding UniGene cluster numbers (1) → (3).

Cancer Genome Anatomy Project (CGAP) was created to determine the gene expression profiles of normal, precancer, and cancer cells (<http://cgap.nci.nih.gov>). The CGAP Genes gathers information on specific genes and collections of genes. In our database we use cumulated data deposited in the ASCII format text

file: *Mm_GeneData.dat* at the www side: <http://cgap.nci.nih.gov/Info/CGAPDownload>. This file contains mouse data in the following format:

- (1) >>UniGene Cluster ID;
- (2) UNIGENE: Organism designator.UniGene Cluster ID;
- (3) SYMBOL: Gene symbol;
- (4) TITLE: Title for the cluster;
- (5) LOCUSID: LocusLink identifier associated with at least one sequence in this cluster;
- (6) CYTOBAND: Cytogenic location;
- (7) OMIM: OMIM number;
- (8) SEQUENCE: Representative GenBank Sequence ID;
- (9) BIOCARTA: BioCarta pathways in which this gene participates;
- (10) KEGG: Kegg pathways in which this gene participates;
- (11) GO: Gene Ontology categories with which this gene is associated;
- (12) MOTIF: Protein similarities based on shared motif content;
- (13) HOMOLOG: Homologous genes related to this gene;
- (14) ALIAS: Alternate symbols for this gene;
- (15) MGC_CLONE: Full-length MGC clones;
- (16) SNP: Single Nucleotide Polymorphisms;
- (17) SV_CLONE: Sequence Verified Clones available from the IMAGE Consortium.

For more information see <ftp://ftp1.nci.nih.gov/pub/CGAP/README>. We used the CGAP data to relate UniGene cluster numbers with gene ontology codes (1)→(11).

Gene Ontology (GO) Consortium was established to provide dynamic, controlled vocabularies for the description of the molecular function, biological process and cellular component of gene products (Ashburner et al., 2000). The GO database is accessible online at: <http://www.geneontology.org>. All relations between gene products and gene ontologies are of the one-to-many type, as a single gene product can have one or more molecular functions, can be used in one or more biological processes and may be associated with one or more cellular components. In our tool we use data deposited in the compressed file: *go_<YYYYMM>-termdb-tables.tar.gz* available online: <http://www.godatabase.org/dev/database/archive>. For instructions on installation of those data within the MySQL system refer to the online README file: <http://www.godatabase.org/dev/database/archive/2003-07-01/README>. During this installation several MySQL tables are created, which can be used to relate GO codes to three basic ontologies: biological process, molecular function and cellular component, and their next level gene ontology classes.

The UniGene number is associated with a non-redundant set of gene-oriented clusters, each of which may contain several sequences that represent a unique gene (<http://www.ncbi.nih.gov/UniGene>). Therefore, it is possible, that several distinct SAGE tags correspond to the same UniGene number. On the other hand, same SAGE tags match more than one UniGene number in the NCBI database. They will be disregarded in our further analysis because this ambiguity indicates that there is no common agreement upon which gene is represented by such a tag. Moreover, some SAGE tags are not associated with any UniGene number and therefore can be considered as potential candidates for unknown genes. The presence of such tags will contribute to the Unknown category. A few examples of SAGE tags falling into each of these cases are presented below

	SAGE tag	UniGene number	Description
Non-unique SAGE tags	AAAAAAAAAAC	28564	protein tyrosine phosphatase-like
	AAAAAAAAAAC	39709	Mus musculus adult male thymus cDNA
	AAAAAAAAAAC	193062	demilune cell and parotid protein
	AAAAAAAAAAC	200776	guanine nucleotide binding protein, beta 1
	AAAAAAAAAAG	856	transmembrane 4 superfamily member 1
	AAAAAAAAAAG	22086	ring finger protein 14
Unique SAGE tags	AAAAAAAAACC	5032	retinal degeneration, slow (retinitis pigmentosa 7)
	AAAAAAAAACG	13052	kidney androgen regulated protein
	AAAAAAACCA	55060	EST
Multiple SAGE tags	ATGGCCTTAC	71	protein kinase, DNA activated, catalytic polypeptide
	CAATTAAGGA	71	protein kinase, DNA activated, catalytic polypeptide
	CAGTTAAGGA	71	protein kinase, DNA activated, catalytic polypeptide
	TTGCCCATTG	71	protein kinase, DNA activated, catalytic polypeptide
Unknown SAGE tags	CATCCCCAAA	None	unknown / novel
	GCTGCCCTCC	None	unknown / novel
	TACTCCCCAAA	None	unknown / novel

Similarly, some UniGene numbers fall into several gene ontologies, whereas other does not match any GO class. Multiple assignments of UniGene numbers to distinct GO ontology classes follow from the fact, that a gene product can have one or more molecular functions, be used in one or more biological processes and may be associated with one or more cellular components. We take this fact into account by placing the corresponding SAGE tag counts into several matching positions in the frequency table. These UniGene numbers, which do not match any GO ontology are considered unclassified and the corresponding tag counts are put into the Unclassified category. A few examples of GO classes that match unique or multiple UniGene numbers are presented below:

UniGene number	GO class
5032	Cellular component: cell
5032	Biological process: physiological process
5032	Molecular function: signal transducer activity
13052	None
55060	None
71	Cellular component: cell
71	Cellular component: unlocalized
71	Biological process: physiological process
71	Molecular function: enzyme activity

To summarize the procedure of functional classification of SAGE tags, we consider a SAGE library with tags from previous examples, and the corresponding frequency table for GO classes. Notice, that we present only GO classes with non-zero entries.

SAGE tag	Counts	GO class	Frequency
AAAAAAAAAAC	40	Cellular component: cell	30+15+10+5+4= 64
AAAAAAAAAAG	35	Cellular component: unlocalized	15+10+5+4= 34
AAAAAAAAACC	30	Biological process: physiological process	30+15+10+5+4= 64
AAAAAAAAACG	25	Molecular function: enzyme activity	15+10+5+4= 34
AAAAAAAAACCA	20	Molecular function: signal transducer activity	30= 30
ATGGCCTTAC	15	Unclassified	25+20= 45
CAATTAAGGA	10	Unknown	40+35+3+2+1= 81
CAGTTAAGGA	5		
TTGCCATTG	4		
CATCCCCAAA	3		
GCTGCCCTCC	2		
TACTCCCCAAA	1		

The two initial SAGE tags, AAAAAAAAAAC and AAAAAAAAAAG, have non-unique UniGene numbers, so they will not contribute to the final result of gene classification and their counts are included in the Unknown category. Each of the next three tags, AAAAAAAAAACC, AAAAAAAAAACG and AAAAAAAAAACCA, matches a unique UniGene number; the first is assigned to three GO classes, whereas the following two are not associated with any GO number and are put into the Unclassified category. The next four tags, ATGGCCTTAC, CAATTAAGGA, CAGTTAAGGA, and TTGCCATTG, are matched to the same UniGene number that, in turn, is put into four different GO classes. The last three tags, CATCCCCAAA, GCTGCCCTCC and TACTCCCCAAA, have no UniGene number associated with them and are put into the Unknown category.

4.2 Secondary data

The procedure of gene classification described in the previous section could be divided into three steps: first the SAGE tags from the user's library are related to the corresponding UniGene numbers using the NCBI SAGEmap resources; then UniGene numbers are matched with gene ontology codes using the CGAP Genes database; and finally GO resources are used to create the frequency table by putting counts from user's library into corresponding ontology classes. To reduce complexity of the classifying SQL query and, more importantly, the time of SQL query execution, the primary data are preprocessed and stored as a MySQL database. After preprocessing the MySQL database contains two tables: *SAGEmapUnique* and *UnigeneClass*, both in the normal forms. Each tag from the user's SAGE library is related to the unique UniGene number (stored in table *SAGEmapUnique*) that in turn is matched with the corresponding ontology classes (stored in table *UnigeneClass*). The preprocessing of the primary data is a complex and time consuming process, but it needs to be performed only once after each update of the primary data (usually once a month).

The MySQL table *SAGEmapUnique* contains those SAGE tags from NCBI SAGEmap resources which are matched uniquely to the UniGene number. The *SAGEmapUnique* table consists of four fields: Tag, Unigene, Name, Description. The primary key in this table is constituted by the Tag field. Structure of the *SAGEmapUnique* table is shown below:

<i>SAGEmapUnique</i> table:				
Field	Type	Null	Key	Default
Tag	Varchar(1)	Yes	Pri	Null
Unigene	Int(11)	Yes		Null
Name	Varchar(15)	Yes		Null
Description	Varchar(250)	Yes		Null

This table has been created by searching the whole file *SAGEmap_tag_ug-rel.zip* and disregarding all repeated entries that match different UniGene numbers.

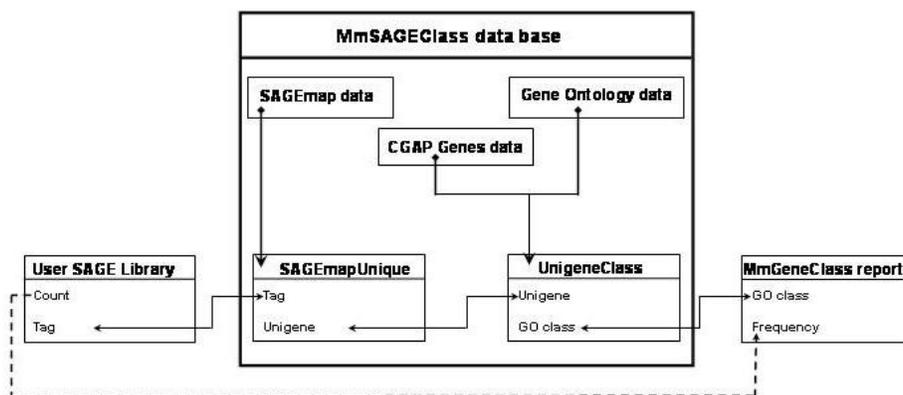
The MySQL table *UnigeneClass* contains pairs of data: UniGene number and the associated GO class. Both fields together constitute the primary key, due to the fact, that the UniGene number can correspond to more than one functional GO class and similarly, every GO class can be associated with more than one UniGene number. Structure of the *UnigeneClass* table is shown below:

<i>UnigeneClass</i> table:				
Field	Type	Null	Key	Default
Unigene	Int(11)	Yes	Pri	0
Class	Int(15)	Yes	Pri	0

This table has been constructed in three steps. First the whole file *Mm_GeneData.dat* from NGCP Genes is searched and pairs of data: UniGene number and GO ontology code are put into a temporal table *MmGeneData*. Next, three basic ontologies and their next level ontology classes are extracted by searching two tables: *term* and *graph_path* from MySQL package *go_<YYYYMM>-termdb-tables.tar.gz*. These data are inserted into a temporal table *OntologyTree*. Finally, the *MmGeneData* table is searched for all pairs (unigene,code) for which the GO code match an ontology class in the *OntologyTree* table (code, class). All pairs of data (unigene,class) are then put into *UnigeneClass* table.

4.3 User query

After preprocessing, the SQL query is completed by matching each tag from the user's library with an identical tag from the *SAGEmapUnique* table and then the corresponding UniGene number with the GO class from the *UnigeneClass* table. The matched tags contribute their counts to the ontology frequency table. The unmatched tags are placed in the Unknown category. The tags which match a unique UniGene number but do not match any GO class are put into the Unclassified category. Note, that user's data are not stored on the server. They are read one-by-one and corresponding frequency counts are updates. A flowchart showing how *MmSAGEClass* manages SAGE data is presented below:



4.4 Online interface

Online interface of the *MmSAGEClass* system is constructed using Perl and DBL scripts which call the SQL queries to the MySQL database. All results are presented on the user's monitor using HTTP and CGI scripts. The step-by-step instructions on how to use the *MmSAGEClass* system are available online and are described below in details.

5. Step-by-step guide to online database

MmSAGEClass database is accessible online at: <http://mbi.osu.edu/~rejniak/MmSAGEClass.html>.

After loading this web page the main window will appear:

MmSAGEClass - Mus musculus SAGE tag Classification		mbi 
<p>Home</p> <p>Submit data</p> <p>Description of</p> <ul style="list-style-type: none"> • Database • SAGE • Classification <p>Comments and contact info</p>	<p><i>MmSAGEClass</i></p> <p>An online database for the functional classification of mouse genes derived from SAGE tags.</p> <hr/> <p><small>An online database <i>MmSAGEClass</i> performs an analysis of SAGE tags and their counts submitted online by the user and presents classification of biological functions of mouse genes associated with the submitted tags.</small></p> <p><small>A process of placing each SAGE tag into a functional class is completed in three steps - first by matching each tag with its unique Unigene number, then by relating each Unigene number to one or more Gene Ontology number and finally by contributing tags counts to ontology classes corresponding to each GO number. Data required to complete these tasks have been obtained by matching records from the following three databases accessible on line:</small></p> <ul style="list-style-type: none"> • NCBI SAGEmap -- http://www.ncbi.nlm.nih.gov/SAGE • CGAP Genes -- http://cgap.nci.nih.gov/Genes • Gene Ontology Consortium -- http://www.geneontology.org <p><small><i>MmSAGEClass</i> operates currently on the data deposited on May 8th 2003.</small></p>	
Last modified on May 8th, 2003		

The user can choose the following options from an interactive menu in the left panel:

- **[Home]** – the main web page of the *MmSAGEClass* system
- **[Submit data]** – an online depository of user's data,
- Description of:
 - **[Database]** – detailed information about the *MmSAGEClass* system and a step-by-step description of the classification process,
 - **[Sage]** – a short description of the SAGE technique,
 - **[Classification]** – a list of all functional categories currently used in the process of SAGE tag classification,
- **[Comments and contact info]** – an online anonymous depository of user's comments and some information about the authors.

This menu is accessible at any time during use of *MmSAGEClass*.

5.1 Data format and size

An acceptable data format will contain only two columns – one for the SAGE 10-based tags and the other for their numerical counts. The columns should be separated by tabulation or space signs. The submitted data **cannot** contain any header line.

The online database *MmSAGEClass* has been tested successfully with SAGE libraries containing up to 24,000 tags, however the analyzing process in the case of large data files is quite slow. For instance, time needed to complete the analysis task for the library containing 5000 distinct tags takes about an hour. Therefore, we advise to divide large tag libraries into smaller parts, to analyze each file separately and finally combine the results manually in one common frequency table.

5.2 Submitting data

The user's data are not stored in the *MmSAGEClass* system and must be re-submitted every time the user wishes to use the *MmSAGEClass* online database. The option **[Submit data]** from the main menu, loads a web page with the online depository of user's data. There are two ways to submit the user's data to the *MmSAGEClass* system – either by pasting the data directly into the text window or by browsing for a path to the file containing SAGE data. After pressing the **[Submit data]** button, the following web page will appear:

Submit your data

An acceptable data format should contain two columns, for tag and counts respectively, separated by a space or tab sign. It may be reasonable to divide large SAGE libraries into smaller parts, analyse each part separately, and finally combine the results.
MmSAGEClass database is capable of analyzing about 80 tags per minute.

Paste your SAGE data directly into the window below:

ATAATACATA	474
GCTGCCCTCC	374
ATACTGACAT	234
GTGGCTCACA	186
ACCAATGAAC	104
AGCAGTCCCC	78
GCACAACTTG	76
AAAAAAAAAAA	73
AAAAATCATC	71
GCTTCGTCCA	54
ATGACTGATA	53
GTAAGCATAA	50
GCTTCTTAC	48
GAAAGCAGGAC	47
CAAACCTCCA	46

The user's data can be pasted directly into the text window. Once the SAGE data are placed in the window, the **[Send Data in the Window]** button should be pressed to submit them to the system. The submitted data **cannot** contain any header line.

The bottom part of this web page contains the browser window, which gives the user an opportunity to browse for a path to the text file containing SAGE data:

Paste your SAGE data directly into the window below:

or

Browse for the text file with your SAGE data:

C:\SAGE\P300E.txt

Once the path to the text file with the SAGE data is placed in the window, the **[Send Data in the Text File]** button should be pressed to submit them. The submitted data **cannot** contain any header line.

5.3 Analyzing data

After pressing the **[Send Data...]** button, *MmSAGEClass* system is checking first if the submitted data satisfy the required format described in section 3.2. If the data does not meet the required format, the system will stop the analysis and will report the inappropriate data format. If the data meet the required format, the system will proceed with the analysis and extract the functional classification. This process may last a while, as the system is searching trough the internal database with more than 500,000 entries. The current time of analysis of the submitted data depends on the size of the submitted library and of the speed on the online connection.

To indicate progress in analyzing data, the system displays information about the number of tags which have already been analyzed: “**k records have been analyzed (out of N)**”. The appearance of the corresponding window is shown below.

The screenshot shows the MmSAGEClass web interface. The title is "MmSAGEClass - Mus musculus SAGE tag Classification" with the mbi logo. The main content area displays "Your data satisfy the required format" and "Start analyzing data (total number: 16)". Below this, a list shows the progress of analysis from 1 to 16 records, each followed by "(out of 16)". At the bottom, there is a "Show Results" button. The left sidebar contains navigation links: Home, Submit data, Description of (Database, SAGE, Classification), and Comments and contact info. The footer indicates "Last modified on May 8th, 2003".

When the analysis is finished, the user is asked to press the button [Show Results] which will display the results in the new window.

5.4 Presenting computed results

After pressing the [Show Results] button, *MmSAGEClass* system displays results of the SAGE tags classification. This is shown in the form of a table displaying the names of all considered classes together with the numbers of genes associated with the SAGE tags which fall into the corresponding functional class. Notice, that each gene can be assigned to more than one functional class, so the total number of genes falling in all listed classes could be much greater than the total number of tags in the user's library. The table contains also the Unclassified and Unknown categories. The corresponding web page is shown below:

The screenshot shows the MmSAGEClass web interface displaying the results of functional classification. The title is "MmSAGEClass - Mus musculus SAGE tag Classification" with the mbi logo. The main content area displays "MmGene Class Results of the functional classification of mouse genes derived from the SAGE library submitted online". Below this, a table shows the results of the classification. The table has three columns: Ontology, Class, and Number of genes. The left sidebar contains navigation links: Home, Submit data, Description of (Database, SAGE, Classification), and Comments and contact info. The footer indicates "Last modified on May 8th, 2003".

	Ontology	Class	Number of genes
1	cellular_component	cell	474
2	cellular_component	cellular_component unknown	0
3	cellular_component	extracellular	0
4	cellular_component	immunoglobulin complex	0
5	cellular_component	obsolete	0
6	cellular_component	unlocalized	0
7	cellular_component	union	0
8	biological_process	behavior	0
9	biological_process	biological_process unknown	0
10	biological_process	cellular process	0
11	biological_process	development	0
12	biological_process	obsolete	0
13	biological_process	physiological processes	474
14	biological_process	viral life cycle	0
15	molecular_function	anticoagulant activity	0
16	molecular_function	antifreeze activity	0
17	molecular_function	antioxidant activity	0
18	molecular_function	apoptosis regulator activity	0

The computed frequency table can also be sent via email to the address provided by the user. In this case the user is asked to enter a proper email address in the window located below the frequency table on the same web page as shown below:

The screenshot shows the 'MmSAGEClass - Mus musculus SAGE tag Classification' web page. The page features a navigation menu on the left with links for 'Home', 'Submit data', 'Description of Database', 'SAGE', 'Classification', and 'Comments and contact info'. The main content area displays a table with 15 rows of classification data. Below the table, there is a form to provide an email address for receiving results, with a pre-filled email 'rejniak@mbl.osu.edu' and an 'Email Results' button. The footer indicates the page was last modified on May 8th, 2003.

Index	Category	Function	Count
31	molecular_function	nutrient reservoir activity	0
32	molecular_function	obsolete	474
33	molecular_function	protein stabilization activity	0
34	molecular_function	protein tagging activity	0
35	molecular_function	regulator of establishment of competence for transformation	0
36	molecular_function	signal transducer activity	0
37	molecular_function	structural molecule activity	54
38	molecular_function	surfactant activity	0
39	molecular_function	toxin activity	0
40	molecular_function	transcription regulator activity	0
41	molecular_function	translation regulator activity	0
42	molecular_function	transporter activity	474
43	molecular_function	triplet codon-amino acid adaptor activity	0
44	Unclassified		466
45	Unknown		1019

The final data can be directly input into a proper software, such as Microsoft Excel, to draw bar or pie charts. A detailed example of functional classification of SAGE tags is presented in section 6.

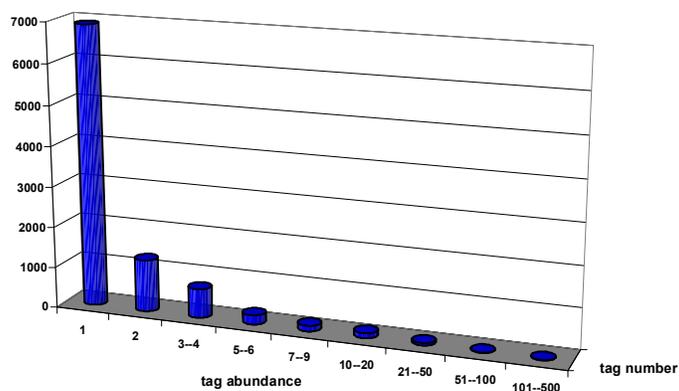
6. SAGE tags classification

MmSAGEClass system analyses the user's library containing tags and their counts and returns a frequency table of the number of SAGE tags that fall into the separate gene ontology classes. When interpreting the obtained results one must remember that the process of functional classification of SAGE tags depends strongly on the primary data deposited at the public databases. Those data show the current biological knowledge about the SAGE tags, mouse genes and their functions, and are constantly updated as new experimental results become available and are deposited by their authors. Therefore, it is important to notice that the functional classification of SAGE tags obtained by the *MmSAGEClass* system reflects only a current biological knowledge and is not an absolute final result. In fact, the interpretation of the submitted SAGE library can change in time when the updated versions of the primary data are available, because some previously unknown SAGE tags can be matched with some known genes or some previously unclassified genes can be assigned to one or more ontology classes. This in turn results in changes of the outcome of the *MmSAGEClass* system. However, the *MmSAGEClass* system enables one to obtain the updated results without repeating of biological experiments, simply by resubmitting the same SAGE library at different time periods. Such a strong dependence on the current biological knowledge requires also a proper interpretation of the frequency tables obtained from the *MmSAGEClass* system. Notice, that the GO classes are not uniformly represented by the number of genes falling into each category. Therefore we need to verify that the functional classification of tags from the SAGE library under consideration is not a consequence of the GO database structure, but it is in fact an indication of genes function. To do so, we will use the one-sided two-sample z-test for population proportions.

To illustrate a process of functional classification of SAGE data by the *MmSAGEClass* system, we will analyze a library constructed from the cerebellum of a 23 day-old C57BL/6 mouse. The data were generated in this laboratory and have been deposited in the NCBI SAGEmap repository, (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2415>). Because each SAGE library contains a relatively small subset of highly expressed tags and a large subset of low abundance tags, we will divide our analysis into three parts. First, we analyze the whole SAGE library, next we will extract two data subsets, one containing a hundred of the most abundant tags and a second subset containing all single-count tags.

6.1 Analysis of the whole library

The GSM2415 SAGE library contains a total of 19436 tags, 9472 of which are distinct and have abundance of 1 to 431 copies per SAGE tag. The distribution of all distinct tags with respect to their abundances is presented below:



Tag abundance	1	2	3-4	5-6	7-9	10-20	21-50	51-100	101-500
Number of tags	6913	1279	714	229	138	127	59	7	6
	72.98%	13.50%	7.54%	2.42%	1.46%	1.34%	0.62%	0.07%	0.06%

The majority of tags have a very low abundance: 72% of all distinct SAGE tags are present at one copy each, and 13.5% have count of 2. High abundance tags constitute less than 0.1% of the GSM2415 library contents (see groups of 50-100 or 100-500 copies per SAGE tag).

In the considered SAGE library 913 tags (9.6%) are unknown, i.e. does not match any UniGene number. Those tags may represent real novel tags or some known tags but not yet completely sequenced. Among all unknown tags in the GSM2415 library, the most abundant are the following:

Tag	Count	Tag	Count	Tag	Count
AGCAATCAA	55	AACGGCTAAA	28	GTGACCACGG	18
ATGACTGATA	55	TAGATATAGG	25	TTCGTCCTTT	10
CAAACCTCCA	44	CTGCGGCTTC	18	AGAGGTGTAG	9

In the SAGE library under consideration, 93% of the unknown tags (847 tags) are present only once. This high percentage of unknown single tags may be an indirect argument for the hypothesis that such tags do not represent real genes, but are a consequence of some errors.

Within the known tags, 3311 (38.7%) are non-uniquely assigned to a UniGene number, matching from 2 to 8 of them. This problem may be resolved in the future when additional biological experiments confirm precisely which gene is represented by a currently non-unique tag. Among all non-unique tags in the GSM2415 library, the most abundant are:

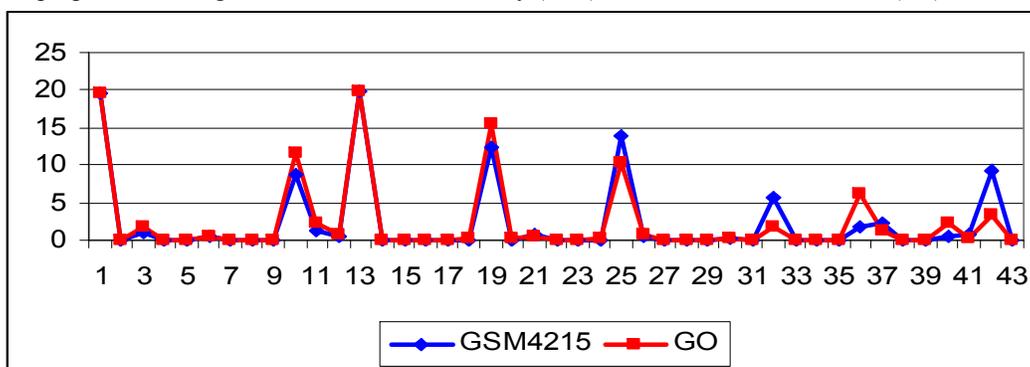
Tag	Count	UniGene	Description
GTGGCTCACA	234	Mm.21758	cytochrome P450, 2e1, ethanol inducible
GTGGCTCACA	234	Mm.38241	interleukin-1 receptor-associated kinase 1
GCTGCCCTCC	195	Mm.104368	hypothetical protein LOC231386
GCTGCCCTCC	195	Mm.196325	ribosomal protein L32
AAAAAAAAAA	104	Mm.16620	glutamic acid decarboxylase 1
AAAAAAAAAA	104	Mm.245395	phosphatase and tensin homolog
GCACAACCTG	69	Mm.18041	calmodulin 2
GCACAACCTG	69	Mm.35677	epidermal growth factor receptor pathway substrate 15
CCCTTCTTCT	68	Mm.89136	H3 histone, family 3A
CCCTTCTTCT	68	Mm.196110	hemoglobin alpha, adult chain 1

All non-uniquely matched SAGE tags are disregarded in the functional classification performed by the current version of the *MmSAGEClass* system, because the classification process depends on associated UniGene numbers.

The remaining 5248 tags match the UniGene number uniquely. Their counts rise from 1 to 431. 75% of them (3945 tags) are single tags. The list of the most frequent unique tags in the GSM2415 library and the corresponding UniGene numbers are presented below:

Tag	Count	UniGene	Description
ATAATACATA	431	Mm.200362	cytochrome b-245, beta polypeptide
ATACTGACAT	281	Mm.257643	ESTs, Weakly similar to 0806162E protein COIII
GCTTCGTCCA	125	Mm.2992	myelin basic protein
AGCAGTCCCC	79	Mm.142498	DNA segment, Chr 18, ERATO Doi 240, expressed
ACCAATGAAC	78	Mm.4962	Tumor differentially expressed I
CAGGCCACAC	52	Mm.103838	ATP synthase, H ⁺ transporting mitochondrial F1 complex
GTAAGCATAA	49	Mm.235	ubiquitin B
AGGACAAATA	41	Mm.258917	TAF1 RNA polymerase II
CAAACCTCTCA	37	Mm.35439	secreted acidic cysteine rich glycoprotein
AGGAGGACTT	30	Mm.14534	RIKEN cDNA 9130221H12 gene

All uniquely matched SAGE tags have been classified into 43 classes using the *MmSAGEClass* system. The proportion of tags falling into each of the considered categories is shown below. For comparison we present proportions of tags from the GSM2415 library (blue) and from the GO database (red):



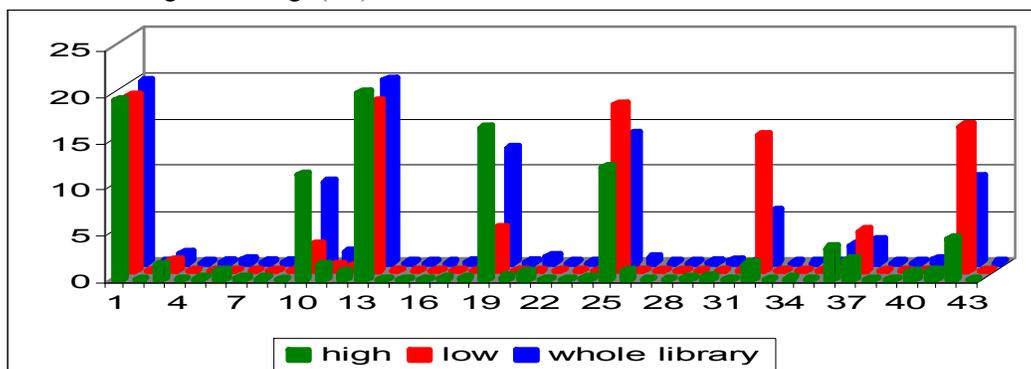
Notice, that in most cases proportions of tags from the SAGE library and from the GO database are very similar. To find out if this is an intrinsic property of SAGE data or simply a result of the GO database structure, we analyzed the differences in tags proportions for all considered categories, using the one-sided two-sample z-test. The over- and under-represented functional classes in the SAGE library, along with their corresponding p-values, are presented below:

		Class name	p-value	
Cellular component	3	extracellular	0.0000	Under-represented
	10	cellular process	0.0000	Under-represented
Biological process	11	development	0.0000	Under-represented
	18	apoptosis regulator activity	0.0149	Under-represented
Molecular function	19	binding activity	0.0000	Under-represented
	20	cell adhesion molecule activity	0.0455	Under-represented
	21	chaperone activity	0.0006	Over-represented
	25	enzyme activity	0.0000	Over-represented
	26	enzyme regulator activity	0.0006	Under-represented
	32	obsolete	0.0000	Over-represented
	36	signal transducer activity	0.0000	Under-represented
	37	structural molecule activity	0.0000	Over-represented
	40	transcription regulator activity	0.0000	Under-represented
	41	translation regulator activity	0.0001	Over-represented
	42	transporter activity	0.0000	Over-represented

For classes not shown in this table, the performed statistical test does not allow to conclude that the tags proportions from both populations are or are not the same.

6.2 Analysis of the libraries of high and low abundance tags

The GSM2415 SAGE library contains a relatively small subset of highly expressed tags (less than 5% of all tags) and a large subset of tags of small abundance (about 73% of tags with a single count). Therefore, we analyzed a subset containing a hundred of the most abundant tags and a subset of single tags, separately. The proportion of tags falling into each of the 43 functional categories is shown below. For comparison, we present proportions of tags from the GSM2415 library (blue), from the highly expressed tags (green), and from a subset of single count tags (red):



Notice that tag proportions of the most abundant tags are, in most cases, similar to the proportions of tags from the whole SAGE library, whereas the proportions of tags in the low abundance group do not follow the same pattern. The significantly different proportions of SAGE tags from both high and low subsets, together with the corresponding p-values are presented below. Last column indicates in which subset the corresponding category is better represented.

		Class name	High	Low	p-value	subset
Biological process	10	cellular process	347	92	0.0000	High
	11	development	46	21	0.0009	High
Molecular function	19	binding activity	499	147	0.0000	High
	25	enzyme activity	371	585	0.0000	Low
	32	obsolete	49	478	0.0000	Low
	37	structural molecule activity	64	141	0.0000	Low
	42	transporter activity	134	511	0.0000	Low

Notice, that in the first ontology, cellular component (classes 1-7), proportions of SAGE tags in all three libraries are similar. In the second ontology, biological process (classes 8-14), the highly abundant tags are over-represented, whereas in the last ontology, molecular function (15-43), the single count SAGE tags are over-represented.

7. Summary

We have described a WWW-based computational tool *MmSAGEClass* for the functional classification of murine genes obtained by the SAGE technique. *MmSAGEClass* system is a MySQL database that combines primary data available from three public repositories: National Center for Biotechnology Information (NCBI) SAGEmap, Cancer Genome Anatomy Project (CGAP) Genes and Gene Ontology (GO) Consortium and assigns each tag from the user's SAGE library to a specific functional category. Results obtained by using *MmSAGEClass* system can be used to gain an insight into the biological functions of genes expressed in tissue from which the SAGE library has been constructed.

References

1. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M, Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M., Sherlock,G. (2000), “Gene Ontology: tool for the unification of biology”, *Nature Genetics*, **25**, 25-29.
2. van Kampen,A.H.C., van Schaik,B.D.C., Pauws,E., Michiels,E.M.C., Ruijter,J.M., Caron,H.N., Versteeg,R., Heisterkamp,S.H., Leunissen,J.A.M., Baas,F., vander Mee,M. (2000), “USAGE: a web-based approach towards the analysis of SAGE data”, *Bioinformatics*, **16(10)**, 899-905.
3. Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggans,G.J., Altschul,S.F. (2000), “SAGEmap: a public gene expression resource”, *Genome Research*, **10**, 1051-1060.
4. Margulies,E.H. and Innis,J.W. (2000), “eSAGE: managing and analysing data generated with Serial Analysis of Gene Expression (SAGE)”, *Bioinformatics*, **16(7)**, 650-651.
5. Velculescu,V.E., Zhang,L., Vogelstein,B. And Kinzler,K.W. (1995), “Serial Analysis of Gene Expression”, *Science*, **270**, 484-487.