SAM
“Significance Analysis of Microarrays”
Users guide and technical document

Gil Chu [*] Balasubramanian Narasimhan [†] Robert Tibshirani [‡]
Virginia Tusher [§]

Contents

1 Important Announcement **3**

2 Summary of Changes **3**
2.1 Changes in SAM 1.21 3
2.2 Changes in SAM 1.20 3
2.3 Changes in SAM 1.15 3
2.4 Changes in SAM 1.13 4
2.5 Changes in SAM 1.12 4
2.6 Changes in SAM 1.10 4

3 Introduction **5**

4 Obtaining SAM **5**

5 System Requirements **5**

6 Installation **6**

[*] Department of Biochemistry, Stanford University, Stanford CA 94305. Email: `chu@cmgm.stanford.edu`.

[†] Department of Statistics and Department of Health Research & Policy, Stanford University, Stanford CA 94305. Email: `naras@stat.stanford.edu`.

[‡] Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford CA 94305. Email: `tibs@stat.stanford.edu`.

[§] Department of Biochemistry, Stanford University, Stanford CA 94305. Email: `goss@cmgm.stanford.edu`.

7 Uninstalling SAM 7

8 Documentation 7

9 Examples 8

10 Data Formats 8
10.1 Response Format . . . 9
10.2 Example Input Data file for an unpaired problem . . . 10
10.3 Block Permutations . . . 10
10.4 Normalization of experiments . . . 11

11 Handling Missing Data 11

12 Running SAM 11
12.1 Using data in Multiple Sheets . . . 15
12.2 Format of the Significant gene list . . . 16

13 Interpretation of SAM output 16

14 Technical details of the SAM procedure 17
14.1 Computation of s_0 . . . 20
14.2 Details of r_i and s_i for different response types. . . . 21
14.3 Details of Permutation Schemes . . . 22

15 Frequently Asked Questions 23
15.1 General Questions . . . 23
15.2 SAM Registration Questions . . . 23
15.3 Installation, Uninstallation Questions . . . 23
15.4 SAM Usage Questions . . . 25

List of Figures

1 Highlighting and invoking SAM . . . 12
2 The SAM Dialog Box . . . 13
3 The SAM Plot . . . 14
4 SAM results for 3 different datasets . . . 18

List of Tables

1 Response Formats 9

2 Example Dataset for an unpaired problem 10

3 Example Dataset for a blocked unpaired problem 10

4 SAM false positive results for 3 scenarios 19

1 Important Announcement

To foster communication between SAM users and make new announcements, a new Yahoo group has been established. See `http://groups.yahoo.com/group/sam-software`.

2 Summary of Changes

The following are changes since the initial release of SAM 1.0.

2.1 Changes in SAM 1.21

Two bugs were fixed.

- A bug relating to what SAM percieves as a large number of permutations was fixed. The default was very naive.

- A bug in adding the imputed data sheets for multiple sheets was fixed. See last para in section 12.1.

2.2 Changes in SAM 1.20

- SAM can now handle a large number of samples. Input data can span several sheets (contiguous or non-contiguous). An example file, named `twoclassbig.xls` included with the distribution. For more details on using multiple sheets, see 12.1.

- A bug in the calculation of FDR for paired data, with a fold change specified, was fixed.

Versions 1.16–1.19 were skipped.

2.3 Changes in SAM 1.15

Bugfix release. A bug that caused SAM to bomb during the calculation of $\hat{\pi}_0$ was fixed.

2.4 Changes in SAM 1.13

Bug fix release. A bug was fixed in the calculations for Censored Survival data. Everyone is advised to upgrade to this version.

2.5 Changes in SAM 1.12

This is mostly a bug fix release. Users of SAM 1.10 should immediately upgrade to this release. Uninstall the previous version and install the new one per instructions in section 7.

- Bug fix: An error in the calculation of the fold-change was fixed. The criterion for applying fold-change to significant genes was also corrected. We thank alert users for catching this.
- By popular request, a new column called **Fold Change** has been added to the significant genes list. This applies only to Two-class and Paired responses. Where the fold change cannot be calculated, it is flagged with an `NA` for "Not Applicable."

2.6 Changes in SAM 1.10

- Bug fixes: a serious bug in the imputation was fixed. The bug caused some data to be imputed with the value 65535. A symptom of this bug was that the plot would have a strange appearance due to the scaling.
- A new facility for block permutations has been added, to handle different experimental conditions such as array batches. See section 10.3.
- In cases where the total number of possible permutations is small, the full set of permutations is used rather than a random sampling.
- The "threshold" now is replaced by a "fold change" criterion, and now handles logged (base 2) and unlogged data appropriately. The fold change applies only to two-class or paired data.
- We have added a new output column to the significant gene list: the "q-value": for each gene, this is the lowest False Discovery Rate at which that gene is called significant. It is like the well-known p-value, but adapted to multiple-testing situations.
- The reported False Discovery Rates are now lower and more accurate than in Version 1.0. They are scaled by a factor $0 \leq \hat{\pi}_0 \leq 1$, that is now displayed on all output. See Section 14 and reference [3].
- Significant gene ids are now linked directly to the Stanford SOURCE web database. Several options for search are provided. Default is by gene name.

- For two-class and paired data, one must now specify whether data is in log-scale or not.
- Stricter checks on response variable values are now performed.
- Several efficiency issues have been addressed.
- The web version of SAM is no longer under development. Hence we have removed it from this manual. The old version still works for the time being, and the version 1.0 manual contains documents it.

Due to changes in the internals of SAM, results using SAM 1.10 will be close to, but not exactly those obtained with SAM 1.0.

We have also updated the FAQ with the latest information. See section 15.

3 Introduction

SAM (Significance Analysis of Microarrays) is a statistical technique for finding significant genes in a set of microarray experiments. It was proposed by Tusher, Tibshirani and Chu [4]. The software was written by Balasubramanian Narasimhan.

The input to SAM is gene expression measurements from a set of microarray experiments, as well as a response variable from each experiment. The response variable may be a grouping like *untreated, treated* [either unpaired or paired], a multiclass grouping (like breast cancer, lymphoma, colon cancer, . . .), a quantitative variable (like blood pressure) or a possibly censored survival time. SAM computes a statistic d_i for each gene i, measuring the strength of the relationship between gene expression and the response variable. It uses repeated permutations of the data to determine if the expression of any genes are significantly related to the response. The cutoff for significance is determined by a tuning parameter **delta**, chosen by the user based on the false positive rate. One can also choose a **fold change** parameter, to ensure that called genes change at least a pre-specified amount. See section 14.

4 Obtaining SAM

SAM is licensed software. Information on licensing of SAM can be obtained from Kirsten Leute (Email: `kirsten.leute@stanford.edu`, Phone: (650) 725-9407) at the Stanford University Office of Licensing (`http://otl.stanford.edu`).

5 System Requirements

SAM requires:

- The latest updates for your operating system available from `http://windowsupdate.microsoft.com`. To prevent any problems, access this and other Microsoft sites using **Internet Explorer** rather than Netscape. Clicking on the **Product Updates** link pops up a box that will automate the installation of the latest patches. Beware that several (time-consuming) reboots are usually needed and you might need administrative privileges to install the patches. It is generally a good idea to update your system for security reasons any way.

- The latest **Microsoft Java Virtual Machine**. This is freely available from the web-site `http://www.microsoft.com/java`. To prevent any problems, access this and other Microsoft sites using **Internet Explorer** rather than Netscape.Choose the correct version for your operating system and install it. Windows XP and Office XP users should especially do so, as Microsoft doesn't distribute Java with its products anymore.

- The **Microsoft Data Access Components**. This is usually available on all newer Windows machines by default, but older Windows NT installations might require you to install it. It is freely available from the web-site `http://www.microsoft.com/data`. To prevent any problems, access this and other Microsoft sites using **Internet Explorer** rather than Netscape.Choose the correct version for your operating system and install it.

- Microsoft Excel 97 or higher. We recommend that users install appropriate Microsoft Office service packs that are available from `http://officeupdate.microsoft.com`. Office 97 users are especially encouraged to do so; there are two service packs for Office 97. The Office 97 service packs are not easy to find; one often has to search the Microsoft website to access them.

Again, performance gets better with faster processors and more RAM. The size of the problem that one can handle is limited by the largest spreadsheet Excel can handle and the memory resources that are allocated to the Microsoft Java Virtual Machine.

6 Installation

If you received SAM on a CDROM, then inserting the CDROM into the drive will bring up the `Setup` program for installing the software. If for some reason that doesn't happen, you can access the CDROM by clicking on `My Computer` and double clicking on the CDROM drive. Then follow the steps below. Otherwise, if you downloaded SAM from the web, you need to use a program like `WinZip` (freely available from `http://www.winzip.com`) on the file `sam.zip` and extract the contents to a suitable folder. This folder will contain a `Setup` program that you need to run.

1. Double click on `Setup`. In some cases, `Setup` will complain that it needs to update your computer and reboot it before it can install SAM. It is safe to click `OK` to update your computer and run `Setup` once again after the reboot.

   Sometimes, the installation process might warn you that a version of a DLL that is being installed is older than one already on your computer. *Elect to keep the existing version.*

   SAM usually installs itself in `C:\Program Files\SAMVB`. Although users can change this directory at the time of installation although we recommend that only the drive letter be changed and not the name of the directory.

2. Fire up Excel and click on the `Tools` menu. Choose `Addins` and click on `Browse`. Select the directory where the setup process installed SAM (`C:\Program Files\SAMVB`) if you chose the defaults) and click on the `Addin` subdirectory. Double click on the `SAM` file. The SAM addin will be loaded and the box against the phrase `Significance Analysis for Microarrays` will be checked.

   Once you click `OK`, you should now see two buttons on your Excel toolbar named `SAM` and `SAM Plot Control`.

This completes the installation.

Windows XP, Office XP users Microsoft doesn't ship Java anymore with its operating system. You *must* install Java yourself as indicated in section 5. Otherwise, SAM will not work!

7 Uninstalling SAM

Before uninstalling SAM, one has to fire up Excel and click on the `Tools` menu. Choose `Addins` and uncheck the box against the phrase `Significance Analysis for Microarrays`.

Then use the `Control Panel` to uninstall the software. If you are asked if shared components should be kept and not discarded, elect to keep them as a conservative measure, unless you are really hard-pressed for space.

8 Documentation

This manual for SAM is available to authorized users from the SAM web-site. After SAM has been installed, the manual is also available as a PDF file in the subdirectory `doc` of the SAM installation directory.

If you don't already have a PDF reader installed, you can do so from the web-site `www.adobe.com`.

9 Examples

Some examples of the use of SAM are in the directory `C:\Program Files\SAMVB\Examples` in the default installation. These examples are meant to familiarize the users with the format in which SAM expects the data.

We briefly describe the examples below.

`twoclass` An example of two class, unpaired data.

`twoclassm` An example of two class, unpaired data, with missing data.

`twoclassb` An example of two class, unpaired data, with experimental blocks defined.

`oneclass` An example of oneclass data.

`multi` An example of multiclass response.

`paired` An example of paired data.

`censored` An example of censored survival data. Note the format of the labels in the first row!

`quantitative` An example of quantitative data.

Instructions on using SAM on these examples is discussed in section 12.

10 Data Formats

The data should be put in an Excel spreadsheet. The first row of the spreadsheet has information about the response measurement; all remaining rows have gene expression data, one row per gene. The columns represent the different experimental samples.

- The first line of the file contains the response measurements, one per column, starting at column 3. This is further described below in section 10.1.

- The remaining lines contain gene expression measurements one line per gene. We describe the format below.

  Column 1 This should contain the gene name. It is for the user's reference.

  Column 2 This should contain the gene ID, for the user's reference. Note that the gene ID column is the column that is linked to the SOURCE website by SAM. Hence a unique identifier (e.g. Clone ID, Accession number or Gene Name/Symbol) should be used in this column, if SOURCE web-site gene lookup is desired.

Remaining Columns These should contain the expression measurements as numbers. Missing expression measurements should be coded as `NA`. This is done easily in good editor or in Excel. In Excel, to change blank fields to `NA`, choose all columns, pull done the **Edit** menu, choose **Replace** and then `nothing (Blank)` with `NA`.

10.1 Response Format

Table 1 shows the formats of the response for various data types. A look at the example files is also informative.

Response type	**Coding**
Quantitative	Real number eg 27.4 or -45.34
Two class (unpaired)	Integer 1, 2
Multiclass	Integer 1, 2, 3, ...
Paired	Integer -1, 1, -2, 2, etc. eg - means Before treatment, + means after treatment -1 is paired with 1, -2 is paired with 2, etc.
Survival data	(Time, status) pair like (50,1) or (120,0) First number is survival time, second is status (1=died, 0=censored)
One class	Integer, every entry equal to 1

Table 1: Response Formats

A *quantitative* response is real-valued, such as blood pressure. *Two class (unpaired)* groups are two sets of measurements, in which the experiment units are all different in the two groups. For example control and treatment groups, with samples from different patients. With a *Multiclass* response there are more than two groups, each containing different experimental units. This is a generalization of the *unpaired* setup to more than 2 groups. *Paired* groups are two sets of measurements in which the same experimental unit is measured in each group. For example samples from the same patient, measured before and after a treatment. *Survival data* consists of a time until an event (such as death or relapse), possibly censored. In the *One class* problem we are testing whether the mean gene expression differs from zero. For example each measurement might be the log(red/green) ratio from two labelled samples hydridized to a cDNA chip, with green denoting before treatment and red, after treatment. Here the response measurement is redundant and is set equal to all 1s.

Sometimes, it is difficult to enter blocking information (see section 10.3) without confusing Excel. Excel thinks such entries are formulae. Therefore, SAM allows any response to be enclosed within quotes (not apostrophes!) and strips the quotes off before doing any computation.

10.2 Example Input Data file for an unpaired problem

The response variable is $1 = untreated$, $2 = treated$. The columns are gene name, gene id, followed by the expression values.

The first row contains the response values.

		1	1	2	2	1	1	2	2
GENE1	101	7.64	-0.50	-1.95	10.12	-10.77	-4.47	-7.65	7.58
GENE2	102	38.10	4.86	7.87	-13.59	-9.79	-13.46	-8.91	-5.07
GENE3	103	21.15	5.96	3.20	-4.74	-3.70	-12.35	-10.17	0.63
GENE4	104	187.21	-23.81	16.76	14.10	-99.76	-89.11	-10.92	5.52

Table 2: Example Dataset for an unpaired problem

Note that there are two blank cells at the beginning of line 1. The gene expression measurements can have an arbitrary number of decimal places.

10.3 Block Permutations

Responses labels can be specified to be in blocks by adding the suffix *blockN*, where N is an integer, to the response labels. Suppose for example that in the two-class data of section 2, samples 1,3,5,7 came from one batch of microarrays, and samples 2,4,6,8 came from another batch. We call these batches “blocks.” Then we might not want to mix up the batches in our permutations of the data, in order to control for the array differences. That is, we’d like to allow permutations of the samples within the set 1,3,5,7 and within the set 2,4,6,8, but not across the two sets. We indicate the blocks (batches) as follows:

		1block1	1block2	2block1	2block2	1block1	1block2	2block1	2block2
GENE1	101	7.64	-0.50	-1.95	10.12	-10.77	-4.47	-7.65	7.58
GENE2	102	38.10	4.86	7.87	-13.59	-9.79	-13.46	-8.91	-5.07
GENE3	103	21.15	5.96	3.20	-4.74	-3.70	-12.35	-10.17	0.63
GENE4	104	187.21	-23.81	16.76	14.10	-99.76	-89.11	-10.92	5.52

Table 3: Example Dataset for a blocked unpaired problem

For example, “1block1” means treatment 1, block (or batch) 1. “1block2” means treatment 1, block (or batch) 2. In this example, there are $4! = 24$ permutations within block 1, and $4! = 24$ permutations within block 2. Hence the total number of possible permutations is $24 \cdot 24 = 196$. If the block information is not indicated in line 1, all permutations of the 8 samples would be allowed. There are $8! = 40320$ such permutations.

Please note that block permutations cannot be specified with Paired response as there is an implicit blocking already in force.

10.4 Normalization of experiments

Different experimental platforms require different normalizations. Therefore, *the user is required to normalize the data from the different experiments (columns) before running SAM*.

SAM does not do any normalization!

For cDNA data, centering the columns of the expression matrix (that is, making the columns mean equal to zero) is often sufficient.

For oligonucleotide data, a stronger calibration may be necessary: for example, a linear normalization of the data for each experiment versus the row-wise average for all experiments.

11 Handling Missing Data

There are currently two options for imputing missing values in SAM.

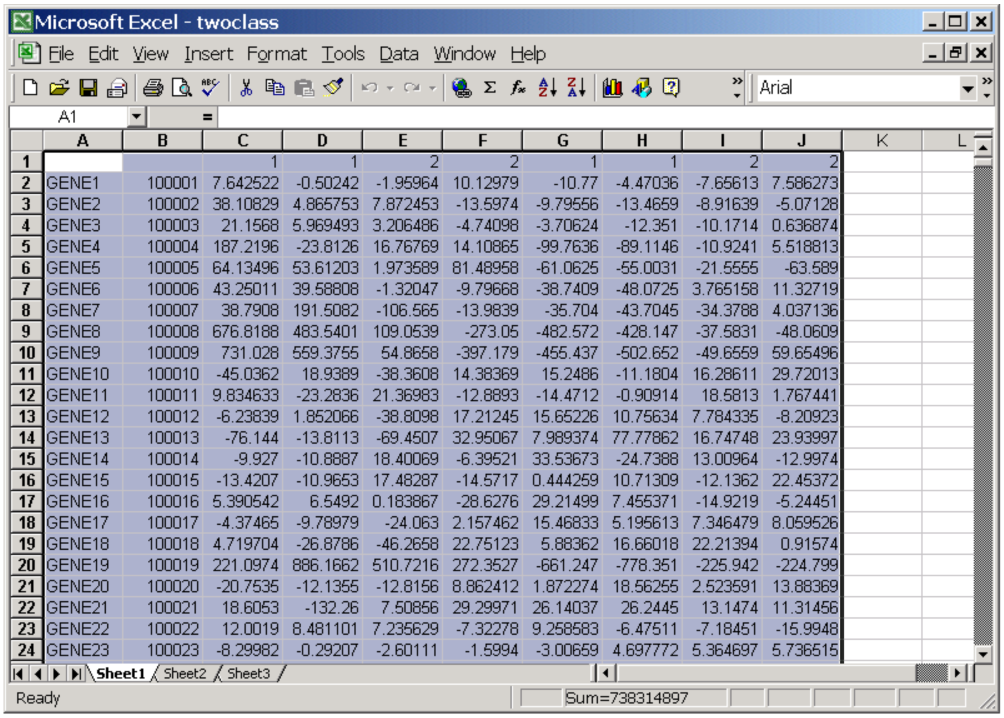
Row Average Each value is imputed with the average of non-missing values for that gene.

K-Nearest Neighbor In the other (default) option- missing values are imputed using a k-nearest neighbor average in gene space (default $k = 10$):

1. For each gene i having at least one missing value:
   (a) Let S_i be the samples for which gene i has no missing values
   (b) Identify all other genes G_i having no missing values for samples S_i
   (c) find the k nearest neighbors to gene i among genes G_i, using only samples S_i to compute the Euclidean distance
   (d) impute the missing values in gene i, using the averages of the non-missing from the k nearest neighbors for that sample.
2. If a gene still has missing values after the above steps, impute them with the average (non-missing) expression for that gene.

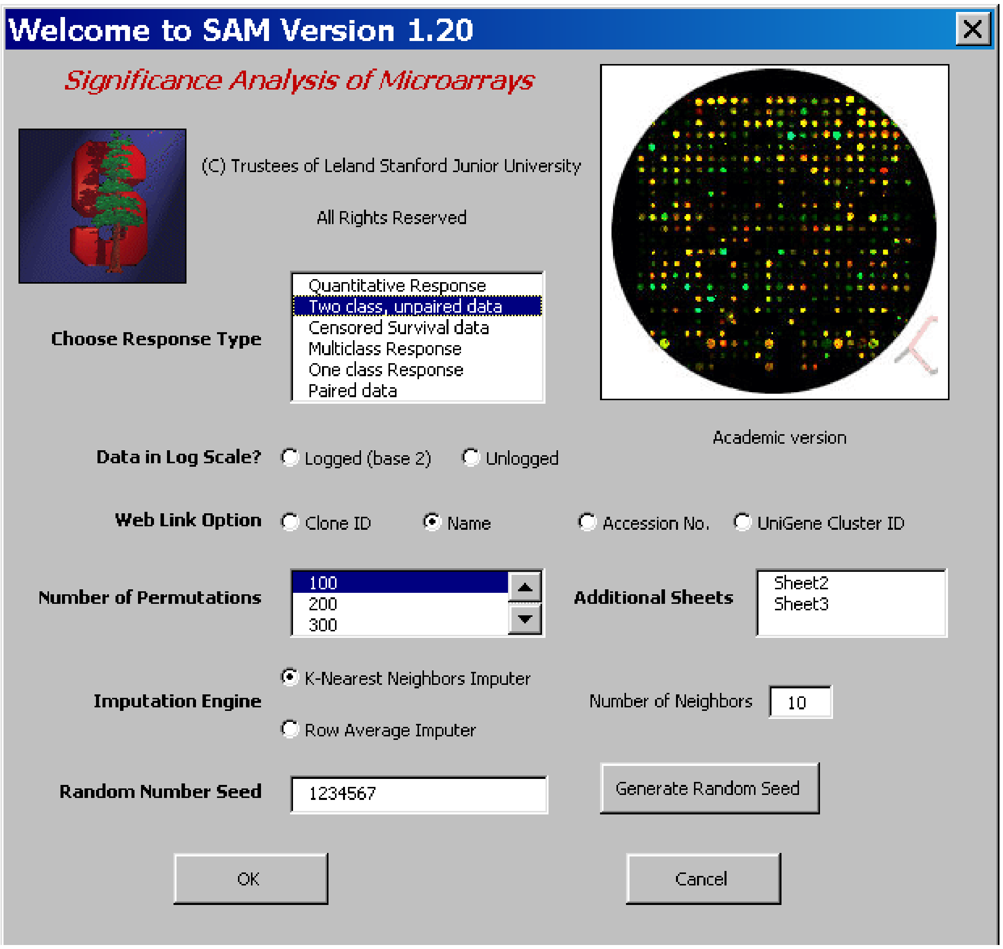
12 Running SAM

To begin, you highlight an area of the spreadsheet that represents the data by first clicking on the top-left corner and then shift-clicking on the bottom right corner of the rectangle. Then, click on the `SAM` button in the toolbar. See illustration in figure 1.


Microsoft Excel - twoclass
File Edit View Insert Format Tools Data Window Help
Arial
A1
=
Sheet1
Sheet2
Sheet3
Ready
Sum=738314897

	A	B	C	D	E	F	G	H	I	J	K	L
1			1	1	2	2	1	1	2	2		
2	GENE1	100001	7.642522	-0.50242	-1.95964	10.12979	-10.77	-4.47036	-7.65613	7.586273		
3	GENE2	100002	38.10829	4.865753	7.872453	-13.5974	-9.79556	-13.4659	-8.91639	-5.07128		
4	GENE3	100003	21.1568	5.969493	3.206486	-4.74098	-3.70624	-12.351	-10.1714	0.636874		
5	GENE4	100004	187.2196	-23.8126	16.76769	14.10865	-99.7636	-89.1146	-10.9241	5.518813		
6	GENE5	100005	64.13496	53.61203	1.973589	81.48958	-61.0625	-55.0031	-21.5555	-63.589		
7	GENE6	100006	43.25011	39.58808	-1.32047	-9.79668	-38.7409	-48.0725	3.765158	11.32719		
8	GENE7	100007	38.7908	191.5082	-106.565	-13.9839	-35.704	-43.7045	-34.3788	4.037136		
9	GENE8	100008	676.8188	483.5401	109.0539	-273.05	-482.572	-428.147	-37.5831	-48.0609		
10	GENE9	100009	731.028	559.3755	54.8658	-397.179	-455.437	-502.652	-49.6559	59.65496		
11	GENE10	100010	-45.0362	18.9389	-38.3608	14.38369	15.2486	-11.1804	16.28611	29.72013		
12	GENE11	100011	9.834633	-23.2836	21.36983	-12.8893	-14.4712	-0.90914	18.5813	1.767441		
13	GENE12	100012	-6.23839	1.852066	-38.8098	17.21245	15.65226	10.75634	7.784335	-8.20923		
14	GENE13	100013	-76.144	-13.8113	-69.4507	32.95067	7.989374	77.77862	16.74748	23.93997		
15	GENE14	100014	-9.927	-10.8887	18.40069	-6.39521	33.53673	-24.7388	13.00964	-12.9974		
16	GENE15	100015	-13.4207	-10.9653	17.48287	-14.5717	0.444259	10.71309	-12.1362	22.45372		
17	GENE16	100016	5.390542	6.5492	0.183867	-28.6276	29.21499	7.455371	-14.9219	-5.24451		
18	GENE17	100017	-4.37465	-9.78979	-24.063	2.157462	15.46833	5.195613	7.346479	8.059526		
19	GENE18	100018	4.719704	-26.8786	-46.2658	22.75123	5.88362	16.66018	22.21394	0.91574		
20	GENE19	100019	221.0974	886.1662	510.7216	272.3527	-661.247	-778.351	-225.942	-224.799		
21	GENE20	100020	-20.7535	-12.1355	-12.8156	8.862412	1.872274	18.56255	2.523591	13.88369		
22	GENE21	100021	18.6053	-132.26	7.50856	29.29971	26.14037	26.2445	13.1474	11.31456		
23	GENE22	100022	12.0019	8.481101	7.235629	-7.32278	9.258583	-6.47511	-7.18451	-15.9948		
24	GENE23	100023	-8.29982	-0.29207	-2.60111	-1.5994	-3.00659	4.697772	5.364697	5.736515		

Figure 1: Highlighting and invoking SAM



Welcome to SAM Version 1.20

Significance Analysis of Microarrays

(C) Trustees of Leland Stanford Junior University

All Rights Reserved

Academic version

Choose Response Type

- Quantitative Response
- Two class, unpaired data
- Censored Survival data
- Multiclass Response
- One class Response
- Paired data

Data in Log Scale? ○ Logged (base 2) ○ Unlogged

Web Link Option ○ Clone ID ● Name ○ Accession No. ○ UniGene Cluster ID

Number of Permutations

- 100
- 200
- 300

Additional Sheets

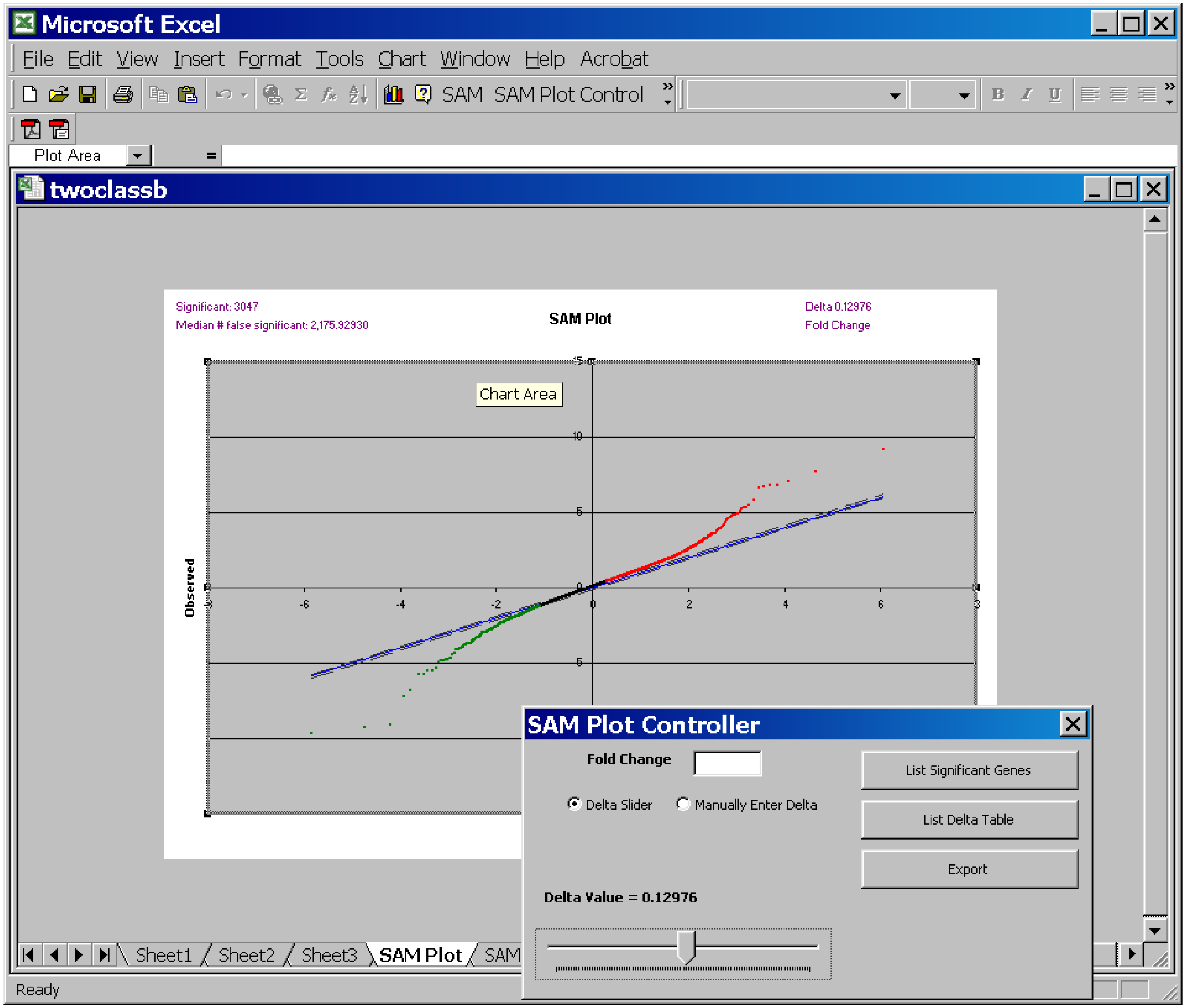
- Sheet2
- Sheet3

Imputation Engine ● K-Nearest Neighbors Imputer ○ Row Average Imputer

Number of Neighbors 10

Random Number Seed 1234567

Figure 2: The SAM Dialog Box



Significant: 3047

Median # false significant: 2,175.92930

Delta 0.12976

Fold Change

Fold Change □

● Delta Slider ○ Manually Enter Delta

Delta Value = 0.12976

A dialog form shown in figure 2 now pops up. You have to select the type of response variable, and if desired, change any of values of the default parameters. For twoclass and paired data, one has to specify if the data is in the logged (base 2) scale or not. Click the `OK` button to do the analysis.

If you had any missing data in your spreadsheet, a new worksheet named **SAM Imputed dataset** containing the imputed dataset is added to the workbook. This data can be used in subsequent analyses to save time. If there is no missing data, this worksheet is not added.

The software adds three more worksheets to the workbook. There is one which is hidden called **SAM Plot data** and should be left alone. The sheet named **SAM Plot** contains the plot that the user can interact with. The sheet named **SAM Output** is used for writing any output.

Initially a slider pops up along with the plot shown in figure 3 that allows one to change the Δ parameter and examine the effect on the false positive rate. It you want a more stringent criterion, try setting a non-zero *fold change* parameter (see section 14 for details). Positive significant genes are labelled in red on the SAM plot, negative significant genes are green. When you have settled on a value for Δ, click on the **List Significant Genes** button, for a list of significant genes. The **List Delta Table** button lists the number of significant genes and the false positive rate for a number of values of Δ. Please note that all output tables are sent to the worksheet named **SAM Output** erasing whatever was previously present in the worksheet.

While the slider is present, all interaction with the workbook is only possible via the slider. It can be killed anytime and recreated by clicking on the `SAM Plot Control` button.

12.1 Using data in Multiple Sheets

The maximum number of columns one can have in an Excel worksheet is 256 columns (`A` through `IV`). If you have more than 256 samples, you can arrange the data in multiple sheets before invoking SAM.

For example, consider the situation where you have 5000 genes and 300 samples. Per the data format required by SAM, this means that the data set would contain $300 + 2 = 302$ columns and 5001 rows. The extra two columns contain the gene name and identifier and the top row contains the response labels.

One possibility is to put the first 256 columns in one sheet and the remaining 46 in another sheet. Or a 100, 100, 102 split over three, not necessarily contiguous worksheets, is also possible—it is your call. Then, highlight the regions in each sheet as usual by clicking on the top-left corner of the rectangle and shift-clicking on the right-bottom corner. Then switch back to the sheet that contains the gene names and ids.

SAM must be invoked from the sheet that contains the gene names and ids. Failure to do so will result in all kinds of hell breaking loose.

The SAM dialog will offer you the option of choosing the additional sheets. Control-click on the sheets that contain the additional data. Proceed as usual after this point.

If any of the n input sheets contains missing data, please note that SAM will add n sheets named **SAM Imputed Dataset**, **SAM Imputed Dataset 1**$\ldots$

12.2 Format of the Significant gene list

For reference, SAM numbers the original genes, in their original order, as 1,2,3, etc. In the output, this is the **Row number**. The output for list of Significant genes has the following format:

Row Number The row in the selected data rectangle.

Gene Name The gene name specified in the first column selected data rectangle. This is for the user's reference.

Gene Id The gene id specified in the second column selected data rectangle. This is for the user's reference, but is also linked to the SOURCE web-site for gene information.

SAM score(d) The T-statistic value.

Numerator The numerator of the T-statistic.

Denominator($s + s_0$) The denominator of the T-statistic.

q-value This is the lowest False Discovery Rate at which the gene is called significant. It is like the familiar "p-value", adapted to the analysis of a large number of genes. The q-value measures how significant the gene is: as $d_i > 0$ increases, the corresponding q-value decreases.

The numerator, denominator and q-value are further explained in the technical section below. The list is divided into positive and negative genes, having positive or negative score d_i. Positive score means positive correlation with the response variable: e.g. for group response 1,2, positive score means expression is higher for group 2 than group 1. For a survival time response, positive score means people with higher expression have longer survival times. The statements are all reversed for negative scores.

13 Interpretation of SAM output

The three panels of figure 4 shows the SAM plots for three different datasets. There are 1000 genes in each of the datasets, and 8 samples, 4 each in control and treatment conditions. We carried out SAM analysis using the unpaired (2 class) option. The corresponding false positive tables are shown in table 4.

In dataset (A) there a number of genes above the band in the upper right and below the band in the bottom left. Looking at table 4, we chose $\Delta = .5$. producing about 65 significant genes and about 5.9 false positives on the average. The choice of Δ is up to the user, depending how many false positives he/she is comfortable with. Note the SAM plots can be asymmetric, in that sometimes there will be significant genes in the top right, but not bottom left, or vice-versa.

In dataset (B) there may be no significant genes. With $\Delta = .5$ (shown in the plot), there are 2 called genes but about 1.3 false positive genes on average.

In dataset (C), there are many significant genes. If $\Delta = 0.3$, then nearly 800 genes are called significant and there are only about 23 false positives on the average. This data was generated as

$$x_{ij} = z_{ij} + \mu_{ij} \tag{13.1}$$

for gene $i = 1, 2, \ldots 1000$, sample $j = 1, 2, \ldots 8$. The first four samples are from group 1, the second four from group 2, Here $z_{ij} \sim N(0, 1)$ (standard normal), $\mu_{ij} = 0$ for $j \leq 4$, $\mu_{ij} = \theta_i \sim N(0, 4)$ for $j > 4$. Hence all genes have a true change θ_i in expression from group 2 vs group 1, although it may be small. In the interpretation of the SAM results, one should also look at the score d_i, which is the standardized change in expression. A value of $d_i = 0.5$ (say) may be called statistically significant in example (C), but is it biologically significant? That is up the scientist. Another way to address this issue: set a non-zero `fold change` for calling genes. With a moderate fold change (say 2), far fewer genes will be called in this example.

14 Technical details of the SAM procedure

The data is x_{ij}, $i = 1, 2, \ldots p$ genes, $j = 1, 2, \ldots n$ samples, and response data y_j, $j = 1, 2, \ldots n$ (y_j may be a vector).

Here is the generic SAM procedure:

1. Compute a statistic

$$d_i = \frac{r_i}{s_i + s_0};\ i = 1, 2, \ldots p \tag{14.1}$$

   r_i is a score, s_i is a standard deviation, and s_0 is a fudge factor. Details of these quantities are given later in this note.

2. Compute order statistics $d_{(1)} \leq d_{(2)} \cdots \leq d_{(p)}$

3. Take B sets of permutations of the response values y_j. For each permutation b compute statistics d_i^{*b} and corresponding order statistics $d_{(1)}^{*b} \leq d_{(2)}^{*b} \cdots \leq d_{(p)}^{*b}$.

4. From the set of B permutations, estimate the expected order statistics by $\bar{d}_{(i)} = (1/B)\sum_b d_{(i)}^{*b}$ for $i = 1, 2, \ldots p$.

5. Plot the $d_{(i)}$ values versus the $\bar{d}_{(i)}$.

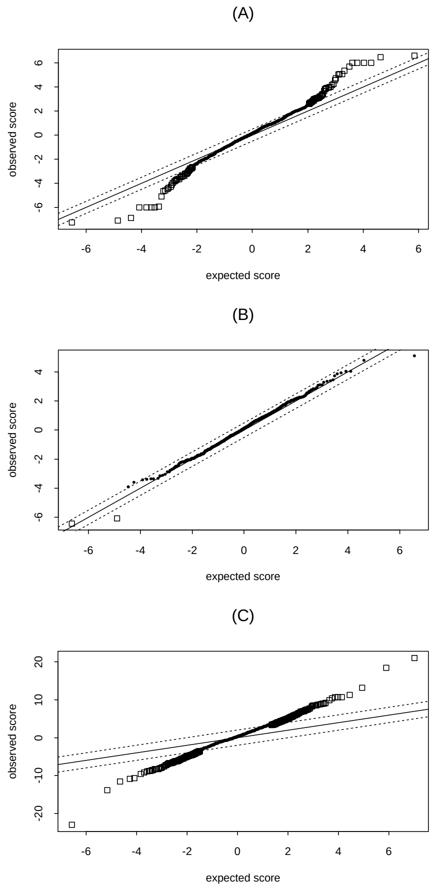


Figure 4: SAM results for 3 different datasets

(A)			
Δ	#false pos	# called	FDR
0.3	11.7	100	0.117
0.4	9.3	76	0.122
0.5	5.9	65	0.091
0.6	4.4	39	0.113
0.7	3.5	33	0.106
0.8	2.1	29	0.072
0.9	1.6	17	0.094
1.0	1.3	16	0.081

(B)			
Δ	#false pos	# called	FDR
0.3	4.8	2	2.40
0.4	1.8	2	0.90
0.5	1.3	2	0.65
0.6	0.6	2	0.30
0.7	0.3	2	0.15
0.8	0.2	0	Inf
0.9	0.2	0	Inf
1.0	0.2	0	Inf

(C)			
Δ	#false pos	# called	FDR
0.3	23.4	894	0.026
0.4	10.6	840	0.013
0.5	5.0	818	0.006
0.6	3.1	780	0.004
0.7	1.9	741	0.003
0.8	1.6	708	0.002
0.9	1.4	674	0.002
1.0	0.9	636	0.001

Table 4: SAM false positive results for 3 scenarios

6. For a fixed threshold Δ, starting at the origin, and moving up to the right find the first $i = i_1$ such that $d_{(i)} - \bar{d}_{(i)} > \Delta$. All genes past i_1 are called "significant positive". Similarly, start at origin, move down to the left and find the first $i = i_2$ such that $\bar{d}_{(i)} - d_{(i)} > \Delta$. All genes past i_2 are called "significant negative". For each Δ define the upper cut-point $\mathrm{cut}_{up}(\Delta)$ as the smallest d_i among the significant positive genes, and similarly define the lower cut-point $\mathrm{cut}_{low}(\Delta)$.

7. For a grid of Δ values, compute the total number of significant genes (from the previous step), and the median number of falsely called genes, by computing the median number of values among each of the B sets of $d^{*b}_{(i)}$, $i = 1, 2, \ldots p$, that fall above $\mathrm{cut}_{up}(\Delta)$ or below $\mathrm{cut}_{low}(\Delta)$. Similarly for the 90th percentile of falsely called genes.

8. Estimate π_0, the proportion of true null (unaffected) genes in the data set, as follows:

   (a) Compute $q25, q75 = 25\%$ and 75% points of the permuted d values (if $p = \#$ genes, $B = \#$ permutations, there are pB such d values).

   (b) Compute $\hat{\pi}_0 = \#\{d_i \in (q25, q75)\}/(.5p)$ (the d_i are the values for the original dataset: there are p such values.)

   (c) Let $\hat{\pi}_0 = \min(\hat{\pi}_0, 1)$ (i.e., truncate at 1).

9. The median and 90th percentile of the number of falsely called genes from step 6, are multiplied by $\hat{\pi}_0$,

10. User then picks a Δ and the significant genes are listed.

11. The False Discovery Rate (FDR) is computed as [median (or 90th percentile) of the number of falsely called genes] divided by [the number of genes called significant].

12. **Fold change**. Suppose $\bar{x}_{i1}$ and $\bar{x}_{i2}$ are the average expression levels of a gene i under each of two conditions. These averages refer to raw (unlogged) data. Then if a nonzero fold change t is also specified, then a positive gene must also satisfy $|\bar{x}_{i2}/\bar{x}_{i1}| \geq t$ in order to be called significant and a negative gene must also satisfy $|\bar{x}_{i1}/\bar{x}_{i2}| \leq 1/t$ to be called significant. When a fold change is specified, genes with either $\bar{x}_{i1} \leq 0$ or $\bar{x}_{i2} \leq 0$ (or both) are automatically left off the significant gene list, as their fold change cannot be unambiguously determined. When such fold changes are reported in output, they are indicated by `NA`.

14.1 Computation of s_0

1. Let s^{α} be the α percentile of the s_i values. Let $d_i^{\alpha} = r_i/(s_i + s^{\alpha})$.

2. Compute the 100 quantiles of the s_i values, denoted by $q_1 < q_2 \ldots < q_{100}$.

3. For $\alpha \in (0, .05, .10 \ldots 1.0)$

   (a) Compute $v_j = \mathrm{mad}(d_i^\alpha | s_i \in [q_j, q_{j+1}))$, $j = 1, 2, \ldots n$, where mad is the median absolute deviation from the median, divided by .64

   (b) Compute $\mathrm{cv}(\alpha)$= coefficient of variation of the v_j values

4. Choose $\hat{\alpha} = \mathrm{argmin}[\mathrm{cv}(\alpha)]$. Finally compute $\hat{s}_0 = s^{\hat{\alpha}}$. s_0 is henceforth fixed at the value $\hat{s}_0$.

14.2 Details of r_i and s_i for different response types.

Quantitative response r_i is the linear regression coefficient

$$r_i = \frac{\sum_j y_j(x_{ij} - \bar{x}_i)}{\sum_j (x_{ij} - \bar{x}_i)^2} \tag{14.2}$$

where $\bar{x}_i = \sum_j x_{ij}/n$ and s_i is the standard error of r_i:

$$s_i = \frac{\hat{\sigma}_i}{[\sum_j (x_{ij} - \bar{x}_i)^2]^{1/2}}, \tag{14.3}$$

and $\hat{\sigma}_i$ is the square root of residual error:

$$\begin{aligned} \hat{\sigma}_i &= \left[\frac{\sum_j (y_j - \hat{y}_{ij})^2}{n-2}\right]^{1/2} \\ \hat{y}_{ij} &= \hat{\beta}_{i0} + r_i x_{ij} \\ \hat{\beta}_{i0} &= \bar{y}_j - r_i \bar{x}_i \end{aligned} \tag{14.4}$$

Two class, unpaired data $y_j = 1$ or 2. Let $C_k = \{j : y_j = k\}$ for $k = 1, 2$. Let $n_k = \#$ of observations in C_k. Let $\bar{x}_{i1} = \sum_{j \in C_1} x_{ij}/n_1, \bar{x}_{i2} = \sum_{j \in C_2} x_{ij}/n_2$.

$$\begin{aligned} r_i &= \bar{x}_{i2} - \bar{x}_{i1} \\ s_i &= [(1/n_1 + 1/n_2)\{\sum_{j \in C_1} (x_{ij} - \bar{x}_{i1})^2 + \sum_{j \in C_2} (x_{ij} - \bar{x}_{i2})^2\}/(n_1 + n_2 - 2)]^{1/2} \end{aligned}$$

Censored survival data $y_j = (t_j, \Delta_j)$. t_j is time, $\Delta_j = 1$ if observation is a death, 0 if censored. Let D be the indices of the K unique death times $z_1, z_2, \ldots z_K$. Let $R_1, R_2, \ldots R_K$ be the indices of the observations at risk at these unique death times, that is $R_k = \{i : t_i \geq z_k\}$). Let $m_k = \#$ in R_k. Let d_k be the number of deaths at time z_k and $x_{ik}^* = \sum_{t_j = z_k} x_{ij}$ and $\bar{x}_{ik} = \sum_{j \in R_k} x_{ij}/m_k$.

$$r_i = \sum_{k=1}^{K} [x_{ik}^* - d_k \bar{x}_{ik}]$$

$$s_i \;=\; [\sum_{k=1}^{K}(d_k/m_k)\sum_{j\in R_k}(x_{ij}-\bar{x}_{ik})^2]^{1/2} \tag{14.5}$$

Multiclass response $y_j \in \{1,2,\ldots K\}$. Let C_k= indices of observations in class k, $n_k = \#$ in C_k, $\bar{x}_{ik} = \sum_{j\in C_k} x_{ij}/n_k$, $\bar{x}_i = \sum_j x_{ij}/n$.

$$r_i \;=\; [\{\sum n_k/\prod n_k\}\sum_{k=1}^{K} n_k(\bar{x}_{ik}-\bar{x}_i)^2]^{1/2} \tag{14.6}$$

$$s_i \;=\; [\frac{1}{\sum(n_k-1)}\cdot(\sum\frac{1}{n_k})\sum_{k=1}^{K}\sum_{j\in C_k}(x_{ij}-\bar{x}_{ik})^2]^{1/2} \tag{14.7}$$

$$\tag{14.8}$$

Paired data $y_j \in \{-1,1,-2,2\ldots-K,K\}$. Observation $-k$ is paired with observation k. Let $j(d)$ be index of the observation having $y_j = d$.

$$z_{ik} \;=\; x_{ij(k)} - x_{ij(-k)} \tag{14.9}$$

$$r_i \;=\; \sum_k z_{ik}/K \tag{14.10}$$

$$s_i \;=\; [\sum_k (z_{ik}-r_i)^2/\{K(K-1)\}]^{1/2} \tag{14.11}$$

One class data $y_j = 1\forall j$.

$$r_i \;=\; \bar{x}_i = \sum_j x_{ij}/n$$

$$s_i \;=\; \{\sum_j (x_{ij}-\bar{x}_i)^2/(n(n-1))\}^{1/2} \tag{14.12}$$

14.3 Details of Permutation Schemes

For *unpaired*, *quantitative*, *Multiclass* and *Survival* data we do simple permutations of the n values y_j. For *Paired data*, random exchanges are performed within each $-k,k$ pair. For *One-class* data, the set of the expression values for each experiment are multiplied by $+1$ or -1, with equal probability. If blocks are specified, the permutations are restricted to be within blocks, as described earlier.

15 Frequently Asked Questions

15.1 General Questions

1. How is SAM licensed? Whom should I contact?

   SAM is distributed without cost to Academic Institutions for research purposes. Academic users of SAM should cite the article [4]. They can download the software after registration directly from `http://www-stat.stanford.edu/~tibs/SAM`.

   Commercial users of SAM should contact Kirsten Leute of the Stanford University Office of Licensing (`http://otl.stanford.edu`) via phone at (650) 725-9407 or via email at `kirsten.leute@stanford.edu`. A limited version of SAM is available for download from `http://www-stat.stanford.edu/~tibs/SAM`.

2. Is there a version of SAM that works on Macintosh computers?

   Unfortunately no. Since the Excel version of SAM makes extensive use of Microsoft Component architecture on Windows (COM), it is not easy to port it to Macs.

   One suggestion that has been made is to use a Windows emulator on Macs such as Virtual PC from Connectix Corporation. We have not confirmed that this works although the folks at Connectix say it should do so with the 4.0.2 update of their virtual PC product.

15.2 SAM Registration Questions

1. I registered for SAM and I have still not received an email confirming my registration.

   This is most likely due to your email server being down. Hundreds of requests have been successfully sent out to people. Our registration server tries every hour to remail the pending requests.

   If you do not receive your registration user-id and password within the day, you may always register again and use another email address that works.

15.3 Installation, Uninstallation Questions

1. How do I uninstall SAM?

   To uninstall, one pretty much reverses the steps in the install process. However, please make sure you do it in the following order.

   (a) First you must unlink SAM from the list of **Addins** loaded into Excel. The list of addins is available by choosing the **Addins** item from the **Tools** menu.

   (b) SAM can be uninstalled via the **Control Panel**. Double Click **Add/Remove Programs** and double click on **Significance Analysis of Microarrays**.

2. How do I install a newly released version of SAM? Do I just install it on top of the old version?

   Installing new software on top of old versions is a good way to hose your Windows machine. If you want to preserve the little sanity that Windows has, you must first uninstall the old version and then install the new version.

3. I just downloaded your SAM program from your website and am having difficulty installing it. When I try to run the `setup.exe` it says it says something about not finding a folder!

   This is most likely due to the peculiarities of your computer.

   - First make sure that your computer has sufficient disk space. It's an easy thing to forget, especially with the amount of crud that Internet Explorer keeps piling up in temporary folders.
   - Extract **s**am.zip to a directory on your C or D drive, say **s**am. You'll need an extractor like **WinZip** or equivalent. We'll assume this on the next step below.
   - Double click on **My Computer** followed by `C:` followed by `sam` and finally `Setup.exe`.

   If even this doesn't work, send email to `sam-bug@stat.stanford.edu` with complete details including

   (a) The error message

   (b) The system you are using (Windows 95, Windows 98, Windows ME, Windows NT or Windows 2000)

   (c) Whether you have installed all the prerequisites mentioned in the SAM manual. In particular, please make sure you have installed the Microsoft Java Virtual machine and the Data Access Components specified in the section 5.

   (d) The dataset you used that generated the error.

4. I would like to revert back to the old version of SAM. How should I go about it?

   We strongly recommend against this. We have expended quite a bit of effort to make the new version of SAM bug-free and correct.

   However, if you really need to do so for other reasons, you must uninstall SAM as usual. Then you must locate the file called `SAMProject.dll` and remove it manually. This file tends to be left orphaned by the uninstallation process.

5. Is there an easier way to detect if Microsoft Java virtual machine is installed on my computer?

   We don't know of any easy way except to set up an applet on our web-site and sniff out your JVM when you access it using Internet Explorer. If there is enough demand, we might do so. The surest bet is to download the latest Microsoft JVM as indicated in section 5.

6. When I install SAM, I get an error that a library was not registered. However, at the end, the program says that the installation was successful. Does this mean that SAM is installed correctly?

   No! Anytime an error occurs, it means that SAM is not installed properly. The problem must be fixed before you can rely on SAM working for you. This often happens when the prerequisites are not met. It also happens if your system doesn't have Microsoft Data Access Components. installed. See section 5.

7. I am using office 97. Where can I download the Service packs for it?

   The last time we looked, it was at the following URL: `http://office.microsoft.com/downloads/9798/sr2off97detail.aspx`.

   If you don't find it there, search for the words **office97 service release** at the web-site `http://office.microsoft.com`. Beware these things keep being moved around!

15.4 SAM Usage Questions

1. SAM generates an error when I run it on my dataset. What should I do?

   Most often, errors are due to improper data formats.

   - Please make sure that your data is formatted exactly as described in section 10. Particular attention needs to be paid to the format of the response in the first row as described in section 10.1.
   - Please make sure that the response type you chose in the SAM dialog box shown in figure 2 matches the format of your response.

     In our testing, about 95% of the problems have been due to the wrong response format.
   - Please make sure that you have chosen your data area appropriately as discussed in 12. It is easy to highlight the wrong area or accidentally highlight some blank cells.
   - Is there a gene with only one or zero *non-missing* value? If so, the imputation will fail.

   Sometimes SAM will run out of memory, especially if the dataset is large. The memory demands during the imputation phase coupled with other demands during the SAM phase can cause SAM to bomb. In such cases, typically, the imputation goes through. One can save the workbook, exit Excel and then rerun SAM on the imputed data.

2. Why does the random number seed stay the same? Can you not generate a new seed automatically?

   The random number seed allows one to reproduce an analysis. By default, it is set to 1234567. However, if one uses the default seed for every analysis, then the *same sequence of permutations* are generated. This is not always desirable. It would appear that generating a seed randomly using the clock or some such mechanism without bothering the user for input might be better. Not necessarily. If reproducibility is important, then asking the user to set the seed is preferable so that any analysis can be rerun to confirm results. We have come down on the side of reproducibility. The user always has a choice of requesting a randomly generated seed based on the clock by clicking on the **Generate Random Seed** button. Please also note that the random number generator seed used in any analysis is always listed in the output to ensure reproducibility of results.

3. How large a dataset can SAM handle?

   There is really no hard limit *per se* in SAM. Excel itself has some limit on the number of rows and columns it can handle. There are additional overheads involved in marshalling the data between Excel and the core of SAM. Therefore, the practical limit is lower. In general, the more memory you have, the larger problems you can handle.

4. I set the value of fold change to some value and now I want to analyze my data without fold change. I seem to be unable to do so.

   To analyze your data without using fold change, completely erase the value for the fold change and leave it empty. You can now hit **Enter** or move your delta slider to recompute the results.

5. Why does SAM take so long to show results when I change the value of fold change?

   Whenever a new value is entered for fold change, SAM has to recompute the q-value bounds for each gene. This is computationally intensive.

6. Why doesn't Excel allow me to enter a response label like **-1.4block1**? It seems to think it is a formula!

   Use quotes around the response label to work around this problem. SAM strips off quotes at the ends of the label.

7. When I enter a different number for the fold change, it seems to have no effect on the number of significant genes!

   This usually happens if one indicates the data is logged when they are actually not logged. Make sure that you specify the correct scale for the data.

8. This document does not answer my questions. Where should I look?

   As we get asked new questions, we update this list of frequently asked questions with answers. Please visit the url `http://www-stat.stanford.edu/˜tibs/SAM` where you may find further information.

9. I get an error that says that

   ```
   cannot create Active X component
   ```

   This is usually due to the prerequisites not being met. Try downloading the Microsoft Java VM as indicated in section 5. We have seen this problem with Office XP, especially.

10. Every time I run SAM I get the message

    ```
    Run-time error ’429’: Active X component can’t create object.
    ```

    How do I fix this problem?

    This error message can occur if the Microsoft Data Access Components are not available on your system. Follow the suggestions in section 5.

11. Where is the SAM manual?

    It should be located in `C:\Program Files\SAMVB\doc` in the default installation. If you used a different directory, then it should be in the analogous place.

    In the worst case, search for the file **sam.pdf**.

12. Where are the examples?

    They should be located in the `C:\Program Files\SAMVB\Examples` in the default installation. If you used a different directory, then it should be in the analogous place.

    In the worst case, search for the file **twoclass.xls**.

13. What does the gene hyperlink lookup do? Does it mean that my identified genes are snooped by Stanford?

    The web lookup facility is provided merely a convenience. One doesn’t have to use it. Just don’t click on it! Please remember that all websites have logs and surely your query gets recorded somewhere. But as to what happens to it, we cannot answer as we have really no affiliation with that site.

    So the bottom line is that if you are really concerned, you should just refrain from using that feature.

14. Where can I go for help if I just cannot get SAM to work?

    We are very interested in making SAM work for all users. However, before reporting problems or bugs, we'd really like you to make sure that the problem is really with SAM. The following checklist should help.

    - Please make sure you have installed all the prerequisites. See section 5.
    - If the problem is with SAM usage, please make sure that you have formatted your data exactly as mentioned in the SAM manual.
    - If you are having problem on a particular type of data, please make sure that you have formatted the response labels appropriately and have chosen the correct applicable data type.

    If you still cannot get SAM to work, send email to `sam-bug@stat.stanford.edu` with complete details including

    (a) The error message

    (b) The system you are using (Windows 95, Windows 98, Windows ME, Windows NT or Windows 2000)

    (c) Whether you have installed all the prerequisites mentioned in the SAM manual. In particular, please make sure you have installed the Microsoft Java Virtual machine and the Data Access Components specified in the section 5.

    (d) The dataset you used that generated the error.

References

[1] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *J. Amer. Stat. Assoc.*, to appear.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data mining, Inference and Predict ion*. Springer Verlag, New York, 2001.

[3] J. D. Storey. A direct approach to false discovery rates. Submitted. Available at `http://www-stat.stanford.edu/~jstorey`.

[4] V. Tusher, R. Tibshirani, and C Chu. Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences*, 2001. First published April 17, 2001, 10.1073/pnas.091062498.