t-SNE Analysis Tool – User Guide

For Visualization and Analysis of gene expression data
This takes advantage of the t-SNE algorithm developed by L.J.P. van der Maaten and G.E. Hinton.
Full details and code are available from http://homepage.tudelft.nl/19j49/t-SNE.html
L. van der Maaten and G. Hinton, Journal of Machine Learning Research 9, 2579 (2008).

Overview:

The starting input is externally prepared data in .csv files. Expected input is a table with each row corresponding to a gene/probe and each column representing an individual experiment/condition. The first row of the table is expected to contain the names of the conditions. The first column the IDs of the genes/probes.
The package is divided into four GUIs (graphical user interfaces) to guide a user through the process.

Annotation Processing – for processing annotation files. This takes the table that describes a specific microarray and processes it into a form that can be associated with the results of the t-SNE during the analysis stage.

Repository Stage - to create a repository file

The repository imports all the data from the .csv file and converts it into a format that Matlab can use. It also performs some preliminary preparation of the data if needed. This includes an optional log2 transformation of all the expression values and editing of the gene names.

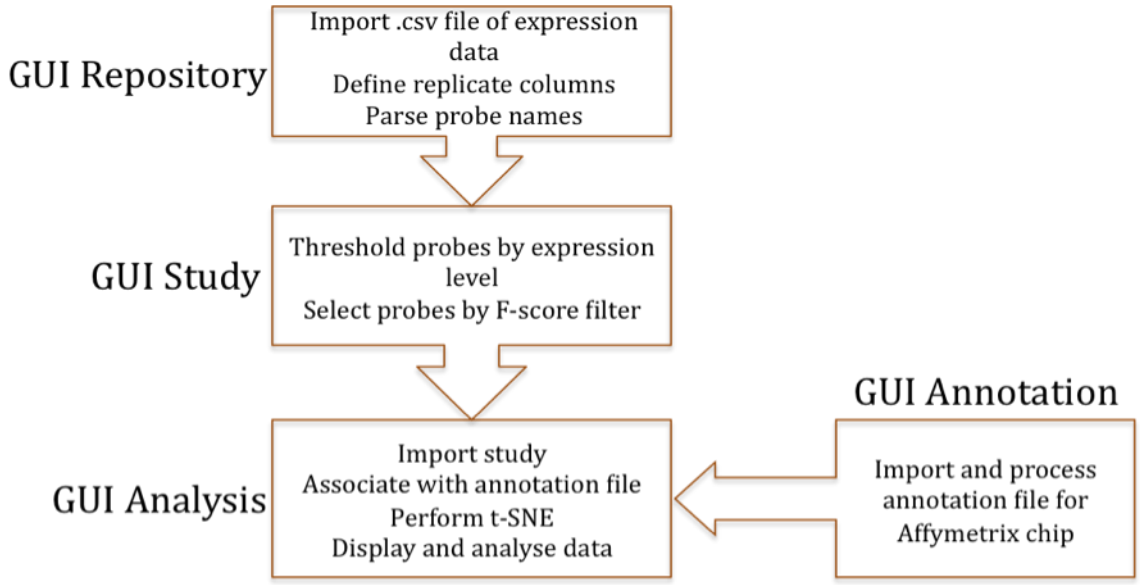
Study Stage - to filter genes from the repository and create a study

The repository file is passed through user controlled filters to generate a study file that contains a subset of probes/genes from the original repository. (These filters are optional and it is possible to convert an entire repository file into a study file). The two filters are:

- Threshold filter. Removes genes that are expressed below a user-defined threshold across all conditions.
- F-score filter. F-Scores are calculated for the remaining genes and a user-defined filter is applied to remove all the genes with F-scores below the defined limit.

Analysis Stage – for interaction and analysis of a study file

This stage imports a study file and performs a t-SNE analysis of the data. The result is processed, linked with the specified annotation files and the result displayed to allow analysis with various tools.



System Requirements

- A valid install of MATLAB – all the scripts run in MATLAB.
    - 2GB of Ram, any operating system.
    - Recommended- Intel 2.0ghz Dual core or higher, 4gb of RAM or higher
    - Best Result – Intel 2.0ghz Dual core or higher, 10gb of RAM or higher, with Linux 64bit as the Operating System.
- These scripts can be quite resource intensive, due to the magnitude of microarray data most datasets will have at least 30,000 probes worth of data to begin with. The scripts should run on any system that runs MATLAB however the program may seem to freeze from time to time, in the majority of places progress information is provided if it is a tool that takes a longer time to process. Errors can be checked in the command window, if these are encountered shut down MATLAB and re-run it.
- The t-SNE algorithm uses more RAM the more genes you put into it. The t-SNE implementation provided with the Analysis GUI is the MATLAB version. This version is written for a 32-bit system, consequently a maximum of ~4Gb of RAM can be used. The maximum limit is therefore ~8000-10,000 probes (this number of probes may take ≥1h to process). For study files containing larger numbers of genes after filtering, a 64-bit Linux version of the algorithm is available from the authors website: http://homepage.tudelft.nl/19j49/t-SNE.html. (For example, ~15,000 probes requires ~7gb of RAM (9GB with MATLAB running, and operating system overheads). The t-SNE algorithm program will throw errors into the MATLAB console if it tries to use more RAM than is available or addressable.
- The MATLAB scripts and t-SNE algorithm all run single threaded.

Setup Instructions

- The folder containing the scripts should be copied to your Matlab folder, which is located in the documents folder for the logged in folder. (MAC OSX - Documents/MATLAB/ , Windows – My Documents/MATLAB/). This should be done once.
- The current folder should be changed to the t-SNE Analysis Tool folder,
- To run the scripts, in the command window at the bottom type the following commands, each step is better documented later in this guide.
    - gui_repository – executes the GUI (Graphical User Interface) repository.
    - gui_study – GUI for the study step
    - gui_analysis – the most recent version of the analyse part of the program
    - gui_annotation – GUI for transforming AffyMetrix tables into a MATLAB format
    - gui_supplementary – dedicated GUI for creating supplementary annotation files.
    - clc – clears the command window
    - clear – deletes all the variables in the 'workspace' (it is a good idea to call this command if you wish to free up RAM, however using it whilst using the analysis GUI will break it, and require to reload the GUI.)

Folder Structure

\utils – the code is all stored in here
\data – data folder
\data\raw data – this is where the processed (.csv microarray) files should be kept
\data\studies – this is where the studies are stored, each repository has its own folder (named after the name given to that repository in the repository GUI) with the all studies for that repository located inside it.
\data\repositories – home of the compiled repositories (*.mat)
\data\libraries – this is where the compiled primary (*.mat) annotation files and compiled supplementary annotation files (*.mat) belong.

Processing and Analysing Data

Pre Matlab

Data should be extracted and summarized from the microarrays using your method of choice (e.g. MAS 5.0, RMA etc). A single .csv should then be generated in which the first column contains the names of the genes/probes and the first row contains the names conditions (or column headers). Eg:

AffyID	ACCN	EntrezGene	Symbol	Description	S9_R1	S9_R2	S9_R3	S10_R1
1053_at	M87338	5982	RFC2	replication factor C (activator 1) 2, 40kDa	7.855188281	7.814563717	7.8073647	7.82892935
121_at	X69699	7849	PAX8	paired box 8	8.043493279	7.943990194	8.046733514	8.2315752
1255_g_at	L36861	2978	GUCA1A	guanylate cyclase activator 1A (retina)	5.219751346	5.68049948	5.278763727	5.0402462
1438_at	X75208	2049	EPHB3	EPH receptor B3	7.243616994	7.118100893	7.073754005	7.4533859
1861_at	U66879	572	BAD	BCL2-antagonist of cell death	5.021842678	5.192448039	5.029581654	5.2741690
200002_at	NM_007209	11224	RPL35	ribosomal protein L35	11.17740846	11.26370309	11.23411272	11.314795

Repository stage – command: gui_repository

The .csv file containing the data is selected and uploaded. The GUI guides the user through the following steps:

Column Selection

From the uploaded data the user chooses which columns to ignore and which to 'comment' ('comment' columns are saved into a supplementary annotation file and used when looking up data in the analysis stage). The user then specifies the number of columns to group together for each condition. For example it is typical to have 3 columns for each condition (triplicates). Each column is assigned to a group using the pull down menu beside each column.

Log Transformation

The input file can be log base 2 transformed. Data that is not already log transformed (e.g. MAS 5.0) should be log transformed, whereas e.g. RMA is already log transformed.

Naming

The repository name is specified. This will be the default folder name for all studies generated from this repository. Studies generated from the repository will be placed in ' \data\studies\<Name Specified>\ ' by default.

Editing Gene IDs

Some Affymetrix microarray tables contain a prefix to the AffyMetrix ID, this needs to be removed so that it matches the exact text from the primary annotation file (see below) in order that it can be cross referenced. The GUI displays the first entry in the table to help you determine the prefix, but it is recommended to have the .csv file open in excel on the side, as this allows you to easily spot if a prefix exists. For example, 'Affymetrix:CompositeSequence:HG-U133B:AFFX-BioB-3_at' could be displayed as the first entry 'Affymetrix:CompositeSequence:HG-U133B:' is the prefix and 'AFFX-BioB-3_at' is the AffyMetrix ID for the probe. To remove the prefix edit the displayed text so that just the AffyMetrix ID is left in the text box then click the button to remove the prefix. Leave blank if no prefix.

Notes

You can save notes for your repository here, they are viewable in both the study stage and the analyse stage.

Study stage – command: gui_study

This stage filters the data uploaded into the repository to select the subset of genes on which to perform a t-SNE mapping. This allows the user to select those probes that are expressed above a defined threshold and vary significantly between conditions.

The GUI will initially request a repository file to be loaded.

Threshold Filtering
The Threshold filter provides a histogram of the distribution of the gene expression values (log2 transformed) in the loaded repository. The highest value of each group/condition for each gene is the value used to set the threshold cut off. This filter removes genes only where all values are below the user-defined threshold in all conditions.

F-Score Filtering
This filtering step calculates and displays a histogram of the F-Scores for each gene that passes the threshold filter. The filter selects those genes with F-scores above the specified value.

Notes
Notes may be added to the study file. There is no need to write down the thresholds as these are saved in the study files and are viewable in the comments pane on the Analysis part of the program.

Analysis stage – command: gui_analysis

This GUI allows calculation and plotting of t-SNE maps and analysis using various tools.

This GUI is resizable. The right side of the GUI is designed to contain one of 4 panels. The panel displayed is specified by the "Right pane" pull down menu in the bottom left of GUI. The "New Figures" pull down menu, allows you to specify whether the next plot specified by the user will be plotted inside the GUI or exported to a Fig file. Export will generate separate figures that you can then save as jpegs as well as in zoom in on. 'Keep inside' will use the embedded figure. (To produce a copy of an uncoloured t-SNE, select 'export', make sure the right pane is set to 'Colour t-SNE', and click the 'Remove Colours / Redraw' button.)

t-SNE Plot
To perform a t-SNE plot. First upload the Study file. A primary annotation file (see below) that matches the data used to generate the expression profiles should then be uploaded. The primary annotation is used to associate each data point with its annotation and to display this information in the table when you select points. You can also search through the annotation via the table, to find specific genes and see where they are plotted as well as viewing their expression and z-scores.

The data is then ready for t-SNE analysis. The t-SNE plot is the centrepiece of the Analysis Tool, it uses the t-SNE algorithm to produce a 2-dimensional projection of the original high dimensional data. The data passed to the t-SNE algorithm can be treated in one of several ways:
Expression means – unmodified values contained in the Study file are passed to t-SNE
Z scores – data converted to standardized (z) scores prior to t-SNE
Zero-row means – the expression levels of each gene zero-meaned before t-SNE
Fold-induction – the fold induction from the lowest value of each gene calculated prior to t-SNE
Percent Induction – the percent induction for each gene calculated so that for each gene the lowest value is set at 0 and the highest value set at 1.
Choose the desired data treatment to input these into the t-SNE algorithm by selecting the appropriate "Plot t-SNE" button, which then initiates the algorithm. Progress can be followed in the main MATLAB window, all the t-SNE plots run for 1000 iterations. The user can specify perplexity, however a value of 30 is the default. (Perplexity is an effective measure of the number of nearby data points that each gene uses to assign its position.)

Once a t-SNE map has been produced it will be displayed in the embedded figure. The output data of the t-SNE algorithm can be exported using "Export t-SNE data". Previously exported data can be imported with "Import t-SNE data".

Analysis and visulization tools

To investigate and visualize groups of genes in the t-SNE plot switch the right panel to "Table of selected genes". Use the select tool (to the right above the embedded figure) to select groups of points from the t-SNE plot. Information regarding the selected points are then placed in the table to the right. Selecting genes from the table will highlight them on the t-SNE plot. The mean expression levels for selected genes in each condition are also displayed in the table. The z-scores/fold-induction/percent-induction/zero-mean/expression levels of genes selected in the table can be displayed in an external figure using the appropriate buttons. The export button on the table saves the contents of the table with all the annotation data as a separate .csv file. The clear button empties the table, and removes highlighted points from the plot.

The entire annotation data for the experiment can be searched for genes/probes using the search box below the table. The remove button removes selected genes from the list. The import list button is used to import a .csv list (e.g. a list of probes in a cluster generated by a separate clustering algorithm) and then search for each of the contents of the list in the t-SNE map.

Neighbour Plots

The neighbour plots link data points in the t-SNE mapping based on their closeness in high dimensional space. These plots can be used as a guide through the t-SNE mapping of the data and to give an indication of the validity of the mapping. The Euclidean distance between every pair of points in high dimensional space is calculated; then for each point these are ranked from closest to furthest point. This gives a rank ordering of the neighbours of each data point. The neighbour plot tool allows the selection of a range (x - y) of neighbours, and each data point (gene) is then connected in the t-SNE plot. Thus the xth to yth nearest neighbour of each gene from the underlying high-dimensional space is linked. For example, entering the values 1 - 5 connects each point (gene) with its first, second, third, fourth and fifth nearest neighbours, resulting in five lines originating from each point. The lines are coloured according to the distance between the connected points in high dimensional space. Red indicates short distances, while blue indicates long distances. To visualize directionality, connectors are wide at their origin and narrow to their neighbour.

The data points in the neighbour plot are selectable and selected points can be imported into the main GUI via the IMPORT button.

Mosaic Plots

If you are unfamiliar with mosaic plots a good explanation is available here: http://www.childrensmercy.org/stats/definitions/mosaic.htm

The Mosaic plot tool can be used to identify groups of genes with similar behaviour in the study. The user specifies threshold z-scores for up to 4 conditions and the mosaic plot divides the genes in the study into groups of genes with the specified characteristics (setting the upper and lower value to the same, eliminates the middle group). The size of each rectangle reflects the number of genes it comprises. By following the labels it is possible to determine which block corresponds to which behaviour. A rectangle(s) of interest can be selected using the selection to lasso the genes within the rectangle. These gene IDs can then be imported into the main Analysis window using the import mosaic data button. These genes are imported into the table and can then be selected and displayed on the t-SNE. Colours are used for the final cut to minimize usage of labels, a legend is displayed stating what the colours mean.

Supplementary Annotations

Supplementary annotations are tables (.csv format or .mat if you used the repository stage or the supplementary annotation GUI to create them) that contain either a list of genes or extra information regarding specific genes/probes, first column in the lists/table should always be the AffyMetrix ID. Multiple supplementary annotation files can be loaded at once, the title of the analysis window will change depending on how many are loaded. All the data from the supplementary annotation files is displayed in the table, in the case of lists, a 1 means it is contained in the list and a blank entry means it is not. You can use the colour tools to select various columns to colour the points on the t-SNE according to these annotations.

Colour Tools

These tools colour the t-SNE plot. Options are expression level/z-scores across 3 conditions as well as data from the annotation files.

Bull's-eye tool
The bull's-eye tool identifies the high dimensional near neighbours of a user defined data point. Select the tool, then select a data point in the embedded t-SNE plot, the tool highlights the nearest neighbours of this data point (from high dimensional space). This selected point is coloured red, it will then calculate the Euclidean distance, from this point to every other point and take the nearest 9, the nearest 4 are coloured green, and the next nearest 5 are coloured blue.

Notes + saving / loading rendered t-SNE plot data
This panel displays notes saved from previous steps, as well as the thresholds values used for the filters in the study step. The analysis comments are saved on exporting a t-SNE map, this saves the data from the t-SNE plot. To load a presaved t-SNE export the primary annotation must be loaded first.

Generation of (Primary) Annotation Files – command: gui_annotation

Primary annotation files, are matlab readable files that contain the 'lookup' table with information on every gene in the dataset. For each dataset there should be one primary annotation file, the scripts use it as a definition file for that dataset, repositories and studies are subsets of the probes in the primary annotation. For Affymetrix chips, the annotation file, from Affymetrix should be converted to.csv format. The first column should be the AffyMetrix ID. Any row that starts with '#' will be ignored. The first row without a '#' character is assumed to be the header row. For other data the first column should consist of the primary key that links each annotation of the gene in the experiment. Primary annotation files only need be to processed and saved once, after that you can use the saved file repeaedly. It will take a few minutes to load and parse the .csv file due to its size.

Known Issues

- select Tool – this tool is not stable, you should not move into other figures whilst selecting points as this can cause issues, as the code will try and use that figure instead. An error catcher attempts to rest the tool if problems are detected. If problems persist its best to restart MATLAB.

FootNote

Created for the National Institute for Medical Research,
Programmed by James N Smith and Chris Watkins, Royal Holloway College, London University.
Developed with James Briscoe, and Natascha Bushati
t-SNE Algorithm developed by: L.J.P van der Maaten and G.E. Hinton
more info available at http://homepage.tudelft.nl/19j49/t-SNE.html
selectdata.m - John d'Errico - http://www.mathworks.com/matlabcentral/fileexchange/13857

Last update – 12 August 2011