MATLAB & Microarray Data Analysis

Yidong Chen
Computational Biology and Bioinformatics
cheny8@uthscsa.edu

WE MAKE LIVES BETTER
UT HEALTH SCIENCE CENTER
SAN ANTONIO

MATLAB in 1 minutes

- A computer language for numerical calculation *faster* than C/C++ and other languages
- **A VERY fancy calculator**
- A simple and user friendly programming language
- A tool for data analysis and statistics
- **A tool to produce plots**

WE MAKE LIVES BETTER
UT HEALTH SCIENCE CENTER
SAN ANTONIO

Statistics Toolbox

Apply statistical algorithms and probability models

Key Features

- Regression techniques, including linear, generalized linear, nonlinear, robust, regularized, ANOVA, and mixed-effects models
- Univariate and multivariate probability distributions
- Random and quasi-random number generators and Markov chain samplers
- Hypothesis tests for distributions, dispersion, and location, and design of experiments (DOE) techniques
- Supervised/Unsupervised machine learning algorithms, including support vector machines (SVMs), k-means and hierarchical clustering, Gaussian mixtures, and hidden Markov models

WE MAKE LIVES BETTER
UT HEALTH SCIENCE CENTER
SAN ANTONIO

Bioinformatics Toolbox

Read, analyze, and visualize genomic, proteomic, and high throughput profiling data

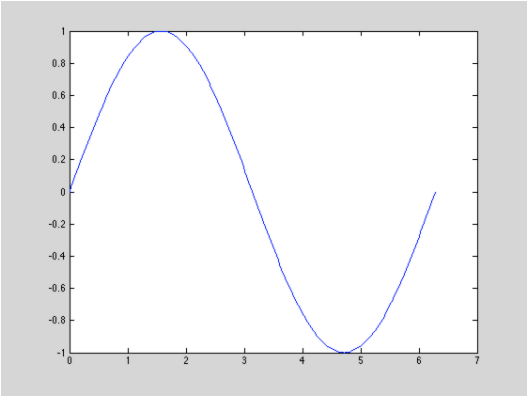
Key Features

- **Next Generation Sequencing** analysis and browser
- Sequence alignment analysis and visualization
- **Microarray,** including reading, filtering, normalizing, and visualization
- **Mass spectrometry**, preprocessing, classification, and marker identification
- **Phylogenetic tree analysis**
- Graph theory functions, including interaction maps, hierarchy plots, and pathways
- Data import from genomic, proteomic, and gene expression files, including SAM, FASTA, CEL, and CDF, and from databases such as NCBI and GenBank

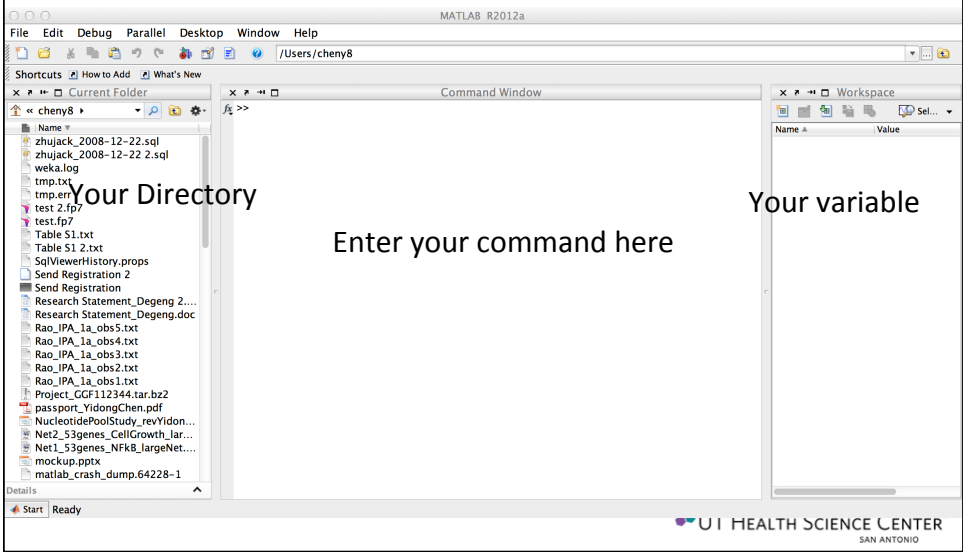
WE MAKE LIVES BETTER
UT HEALTH SCIENCE CENTER
SAN ANTONIO

What does MATLAB code look like?

```
t = 0:pi/100:2*pi;
y = sin(t);
plot(t,y)
```



MATLAB Desktop (Mac, R2012a)



Some simple functions

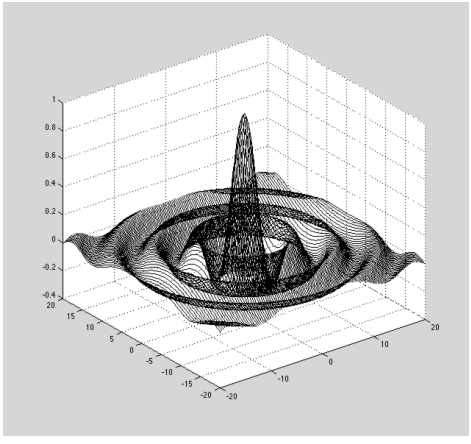
```
>> A = [1 3 2 1; 5 10 11 8; 9 6 7 1; 4 15 14 1]
MATLAB displays the matrix you just entered.
A =     1  3  2  1
        5 10 11 8
        9  6  7  1
        4 15 14 1
>> sum(A)
ans =

   19   34   34   11

>> B = sum(A')
B =

    7   34   23   34
```

```
>> mean( A )
ans =

    4.7500    8.5000    8.5000    2.7500

>> sigma = std( A )
sigma =

    3.3040    5.1962    5.1962    3.5000
```

WE MAKE LIVES BETTER
UT HEALTH SCIENCE CENTER SAN ANTONIO

Fancy plots

```
x1 = -20:0.3:20;
y1 = -20:0.3:20;
[x, y] = meshgrid(x1,y1);
r = sqrt( x.^2 + y.^2 );
s = sin(r)./r;
s(find(r==0)) = 1;
plot3( x, y, s );
grid on;

Try these:
plot3( x, y, s, 'k' );
mesh( s );
surf( s );
shading interp
```



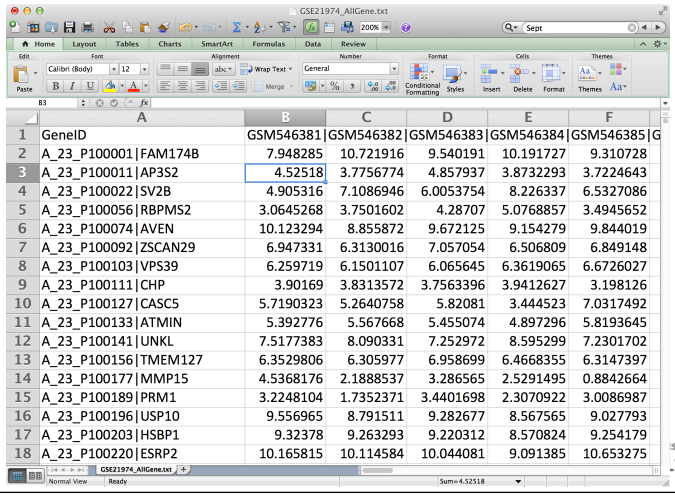
WE MAKE LIVES BETTER
UT HEALTH SCIENCE CENTER SAN ANTONIO

Loading Microarray Data

WE MAKE LIVES BETTER
UT HEALTH SCIENCE CENTER
SAN ANTONIO

Reading generic text, tab-delimited file

- GSE21974_AllGene.txt – Using excel.



	A	B	C	D	E	F
1	GeneID	GSM546381	GSM546382	GSM546383	GSM546384	GSM546385	G
2	A_23_P100001|FAM174B	7.948285	10.721916	9.540191	10.191727	9.310728	
3	A_23_P100011|AP3S2	4.52518	3.7756774	4.857937	3.8732293	3.7224643	
4	A_23_P100022|SV2B	4.905316	7.1086946	6.0053754	8.226337	6.5327086	
5	A_23_P100056|RBPMS2	3.0645268	3.7501602	4.28707	5.0768857	3.4945652	
6	A_23_P100074|AVEN	10.123294	8.855872	9.672125	9.154279	9.844019	
7	A_23_P100092|ZSCAN29	6.947331	6.3130016	7.057054	6.506809	6.849148	
8	A_23_P100103|VPS39	6.259719	6.1501107	6.065645	6.3619065	6.6726027	
9	A_23_P100111|CHP	3.90169	3.8313572	3.7563396	3.9412627	3.198126	
10	A_23_P100127|CASC5	5.7190323	5.2640758	5.82081	3.444523	7.0317492	
11	A_23_P100133|ATMIN	5.392776	5.567668	5.455074	4.897296	5.8193645	
12	A_23_P100141|UNKL	7.5177383	8.090331	7.252972	8.595299	7.2301702	
13	A_23_P100156|TMEM127	6.3529806	6.305977	6.958699	6.4668355	6.3147397	
14	A_23_P100177|MMP15	4.5368176	2.1888537	3.286565	2.5291495	0.8842664	
15	A_23_P100189|PRM1	3.2248104	1.7352371	3.4401698	2.3070922	3.0086987	
16	A_23_P100196|USP10	9.556965	8.791511	9.282677	8.567565	9.027793	
17	A_23_P100203|HSBP1	9.32378	9.263293	9.220312	8.570824	9.254179	
18	A_23_P100220|ESRP2	10.165815	10.114584	10.044081	9.091385	10.653275	

Importdata, load, textscan, fscanf, xlsread

```
importdata( 'GSE21974_AllGene.txt' );
```

Problem: because the first line is text, this command does not work

```
load 'GSE21974_AllGene.txt'
```

Use textscan:

```
fid = fopen('GSE21974_AllGene.txt' , 'r' );
A = textscan( fid, '%s', 58 );
fclose( fid );
```


Read Microarray data

- Affyread: affymetrix CEL file
- Agferead: agilent feature extraction
- Ilmnbsread: illumina beadStudio output file (text file)

We use custom code

- For generic text
  - `ReadTextWithoutHeader()`

  e.g.

```
data = ReadTextWithHeader( 'GSE21974_AllGene.txt', 3 );
```

- For agilent microarray
  - `ReadAgilentResult()`

  e.g.

```
data = ReadAgilentResult(
'GSM546381_US91803681_251485049822_S01_GE1_105_Dec08_1_1.txt' );
```

  Request source code @ CBBI

Examination of Raw Agilent Microarray Data

- Download one of example Agilent gene expression data
  - http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM546381

TYPE	FEPARAMS	DATA *	TYPE	STATS	DATA *	TYPE	FEATURES
text	Protocol_Name	GE1_105_Dec08 (Read Only)	float	gDarkOffsetAverage	6.16137	integer	FeatureNum
text	Protocol_date	11/21/08 12:35	float	gDarkOffsetMedian	6.16137	integer	Row
text	Scan_Date	9/16/09 11:06	float	gDarkOffsetStdDev	1.39298	integer	Col
text	Scan_ScannerName	Agilent Technologies Scanner G2505(	integer	gDarkOffsetNumPts	0	text	accessions
integer	Scan_NumChannels	1	integer	gSaturationValue	779631	text	chr_coord
float	Scan_MicronsPerPixelX	5	float	gAvgSig2BkgeQC	589.932	integer	SubTypeMask
float	Scan_MicronsPerPixelY	5	float	gAvgSig2BkgNegCtrl	1.47761	text	SubTypeName
text	Scan_OriginalGUID	d7cf9523-ff83-43d3-bc16-4c98d5173	float	gRatioSig2BkgeQC_NegCtrl	399.247	integer	Start
text	Grid_Name	014850_D_F_20090416	integer	gNumSatFeat	0	text	Sequence
text	Grid_Date	4/16/09 0:00	float	gLocalBGInlierNetAve	12.7466	integer	ProbeUID
integer	Grid_NumSubGridRows	1	float	gLocalBGInlierAve	18.908	integer	ControlType
integer	Grid_NumSubGridCols	1	float	gLocalBGInlierSDev	1.73806	text	ProbeName
integer	Grid_NumRows	532	integer	gLocalBGInlierNum	44645	text	GeneName
integer	Grid_NumCols	85	float	gGlobalBGInlierAve	18.908	text	SystematicName
float	Grid_RowSpacing	73.3235	float	gGlobalBGInlierSDev	1.73806	text	Description
float	Grid_ColSpacing	63.5	integer	gGlobalBGInlierNum	44645	float	PositionX
float	Grid_OffsetX	0	integer	gNumFeatureNonUnifOL	1	float	PositionY
float	Grid_OffsetY	36.6617	integer	gNumPopnOL	138	float	gSurrogateUsed
float	Grid_NomSpotWidth	65	integer	gNumNonUnifBGOL	0	boolean	gIsFound
float	Grid_NomSpotHeight	65	integer	gNumPopnBGOL	370	float	gProcessedSignal
text	Grid_GenomicBuild	hg18:NCBI36:Mar2006	float	gOffsetUsed	6.16137	float	gProcessedSigError
text	FeatureExtractor_Barcode	251485049822_1_1	float	gGlobalFeatInlierAve	2228.19	integer	gNumPixOLHi
text	FeatureExtractor_Sample		float	gGlobalFeatInlierSDev	11292.8	integer	gNumPixOLLo
text	FeatureExtractor_ScanFileName	D:\Daten_Agilent_Rechner\ScanData	integer	gGlobalFeatInlierNum	44876	integer	gNumPix
text	FeatureExtractor_ArrayName	US91803681_251485049822_S01	float	AnyColorPrcntFeatNonUnifOL	0.00222148	float	gMeanSignal
text	FeatureExtractor_DesignFileName	014850_D_F_20090416.xml	float	AnyColorPrcntBGNonUnifOL	0	float	gMedianSignal
text	FeatureExtractor_PrintingFileName		float	AnyColorPrcntFeatPopnOL	0.306564	float	gPixSDev
text	FeatureExtractor_PatternName	Agilent-014850	float	AnyColorPrcntBGPopnOL	0.821948	float	gPixNormIQR
text	FeatureExtractor_ExtractionTime	9/16/09 12:07	float	TotalPrcntFeatOL	0.00230542	integer	gBGNumPix
text	FeatureExtractor_UserName	sa_agilent	integer	gNumNegBGSubFeat	4210	float	gBGMeanSignal
text	FeatureExtractor_ComputerName	AGILENT02	integer	gNonCtrlNumNegFeatBGSubSig	3344	float	gBGMedianSignal
text	FeatureExtractor_ScanFileGUID	957a1497-1aa5-4f45-ba2e-2ab8537d	float	gSpatialDetrendRMSFit	1.17339	float	gBGPixSDev

FirstLook

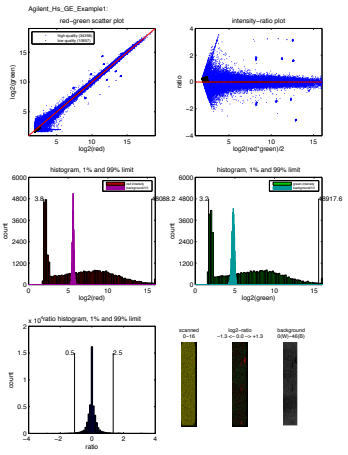
When you use

```
ReadAgilentResult('Agilent_Hs_
GE_Example1.txt'),
```

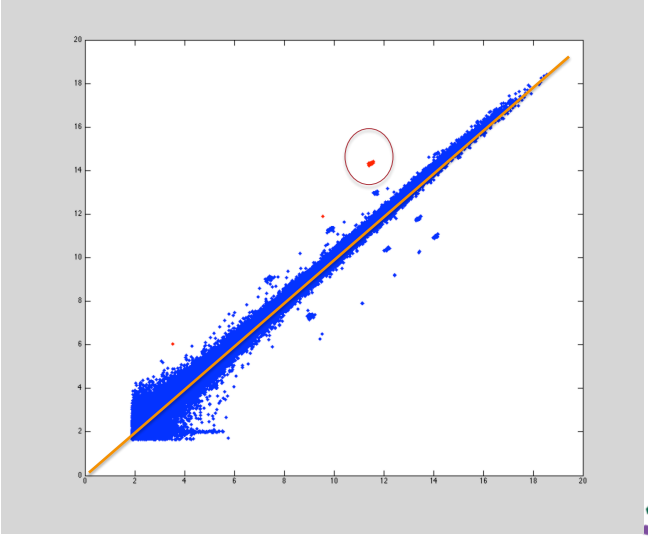
it generate a pdf file,

```
Agilent_Hs_GE_Example1.txt_qFig.p
df
```

It is mainly for quality control purpose.



Control probes, background fluorescent



What are these??

Normalization

- Boxplot, Lowess, quantile

```
>> boxplot( data.value, 'plotstyle', 'compact' );
>> data.normValue = quantilenorm( data.value );
```

(for agilent microarrays, lowess correction is not needed)

```
hold on;
plot( data.value(8698,:), 'r' );
text( 1, data.value(8698,1), data.GeneID(8698), 'interpret',
'none', 'horizontalalignment', 'right');
```


Intensity, median/mean, ratio, etc

- This is single channel data file with data in gProcessedSignal
- For 2 channels, data are in `rProcessedSignal`, and `gProcessedSignal`. With ratio in log10 transformation.
  - http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM785046

PCA, MDS, etc

- Principal component analysis

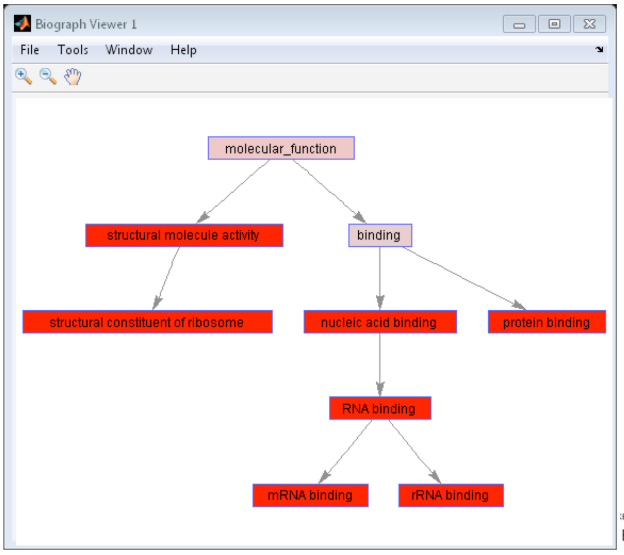
```
[coefs, scores, sigma2] = princomp( data.value' );
```

- Multidimensional scaling

Differential Gene Expression

```
p = mattest( data.value( 1:2:57), data.value(2:2:57) );
adjp = mafdr( p, 'BHFDR', true );

Or

[fdr, q] = mafdr( p );
```

Gene Ontology



Other subjects

- Phylogenetic tree
- Sequence alignment
- Motif search
- Mass Spec data analysis
- And many more…