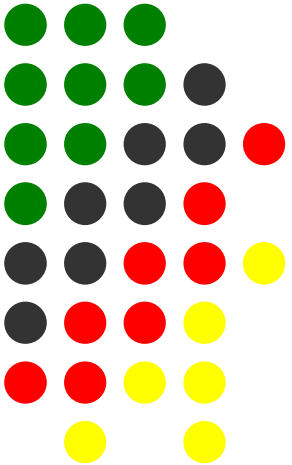


Microarray Data Analysis

Gene Filtering, Missing Values Imputation

國立臺灣大學 資訊所

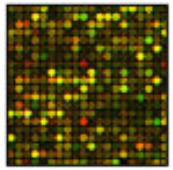
Course: 生物資訊之統計與計算方法

2006/04/11

吳漢銘
hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu/

Institute of Statistical Science, Academia Sinica
中央研究院 統計科學研究所

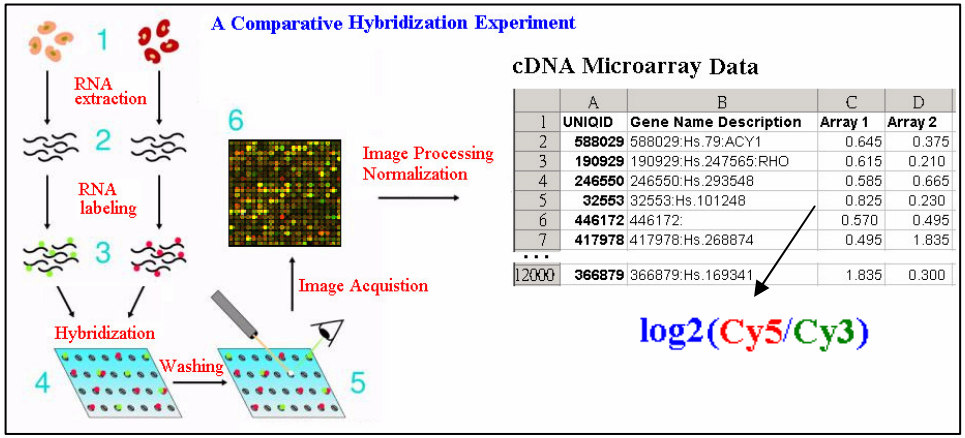
Outlines



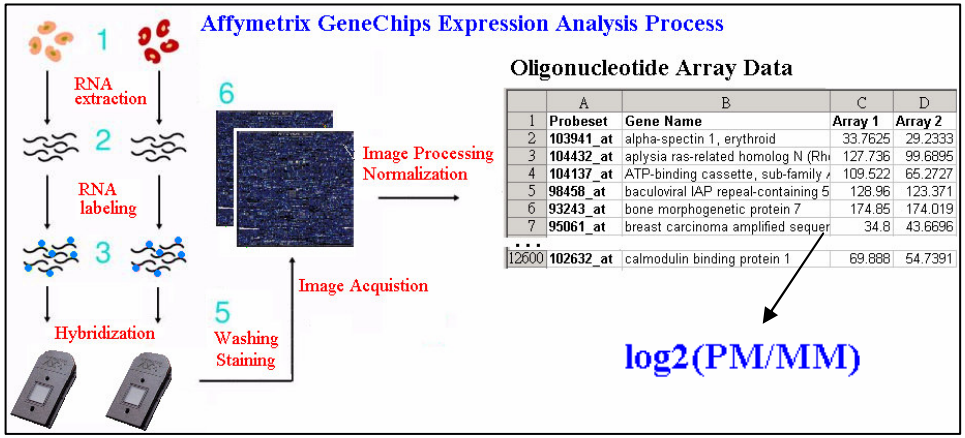
- Quality Control
- Gene Filtering

- MSA5: Detection Calls
- dChip: Filter Genes
- Missing Values Imputation

Gene Expression Microarray Data



	A	B	C	D
1	UNIQID	Gene Name Description	Array 1	Array 2
2	588029	588029:Hs.79:ACY1	0.645	0.375
3	190929	190929:Hs.247565:RHO	0.615	0.210
4	246550	246550:Hs.293548	0.585	0.665
5	32553	32553:Hs.101248	0.825	0.230
6	446172	446172:	0.570	0.495
7	417978	417978:Hs.268874	0.495	1.835
...				
12000	366879	366879:Hs.169341	1.835	0.300



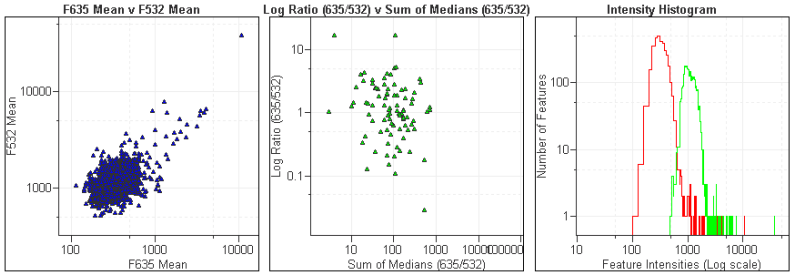
	A	B	C	D
1	Probeset	Gene Name	Array 1	Array 2
2	103941_at	alpha-spectin 1, erythroid	33.7625	29.2333
3	104432_at	aplysia ras-related homolog N (Rh	127.736	99.6895
4	104137_at	ATP-binding cassette, sub-family	109.522	65.2727
5	98458_at	baculoviral IAP repeal-containing 5	128.96	123.371
6	93243_at	bone morphogenetic protein 7	174.85	174.019
7	95061_at	breast carcinoma amplified sequer	34.8	43.6696
...				
12600	102632_at	calmodulin binding protein 1	69.888	54.7391

Software for Quality Control

GenePix Pro

Analysis Reports and Quality Control Measure

- Array Quality Control
- Controls Quality Control
- Feature Quality Control
- Replicates Quality Control
- Signal-To-Noise



(untitled)

Scanned by:	GenePix 4000B [81990]	Analyzed by:	GenePix Pro 6.0.0.54
Scanned on:	2003/9/29 下午 09:51:37	GPS file:	
Image wavelengths:	635, 532	GAL file:	0701.gal
PMT:	850, 750 V	Temperature:	27.6 V
Laser Power:	3.2, 3.9 V	Laser On-time:	661, 453
Scan Power:	100, 100 %	Barcode:	none
Normalization Factors:	1, 1	Normalization Method:	None
Wavelength Image Files:	7-01.tif		
Comment:	none		

Vital Statistics

	635	532	Threshold	Results 635	Results 532
Median signal-to-background	0.9	0.9	> 10	Fail	Fail
Mean of median background	283.21	1065.585	< 500	Pass	Fail
Median signal-to-noise	0	0	> 10	Fail	Fail
Median % > B+1SD	8	8	> 90	Fail	Fail
Feature variation	0.915	0.622	< 0.5	Fail	Fail
Background variation	1.181	0.855	< 0.5	Fail	Fail
Features with saturated pixels	0 %	0.12 %	< 0.1 %	Pass	Fail
Not Found features	5855/9248 (63.3%).		< 7 %	Fail	
Bad features	0/9248 (0%).		< 7 %	Pass	

Bioconductor

arrayMagic:
http://www.dkfz-heidelberg.de/abt0840/home/buness/WWW/Software/

arrayQuality
http://www.ugrad.stat.ubc.ca/R/library/arrayQuality/doc/guide.html

GeneFilter
http://www.bioconductor.org/packages/bioc/1.8/html/genefilter.html

Gene Filtering

Some common defects

1. Feature is smeared into a neighboring feature;
2. Feature is very close to background;
3. Feature has a hair or a scratch through it;
4. Feature is in pieces;
5. Feature is saturated;
6. Feature pixels have highly non-uniform intensities;
7. Feature has a highly non-uniform background.

Note

- Each of these defects is evaluated for each feature individually.
- All filtering conditions should be applied to microarrays that have already been normalized.

Condition 1:

Flags >= 0

Condition 2:

SNR 635 > 3 AND SNR 532 > 3
Sum of Medians (635/532) > 200
% > B635+2SD > 55 AND % > B532+2SD > 55

Conditions 3 and 4:

Circularity > 80

Condition 5:

F635 % Sat. < 2 AND F532 % Sat. < 2

Conditions 6 and 7:

B635 CV < 25 AND B532 CV < 25
Rgn R2 > 0.6

See:

GPR - GenePix Results format (*.gpr)

MSA5: Detection Calls

- Answers: "Is the transcript of a particular gene Present or Absent?"
- Absent means that the expression level is below the threshold of detection. That is, the expression level is not provably different from zero.
- Advantage: easy to filter and easy to interpret: we may only want to look at genes whose transcripts are detectable in a particular experiment.

Method

There are four steps to the method:

1. Remove saturated probe pairs and ignore probe pairs wherein PM ~ MM + *tau*
2. Calculate the discrimination scores. (This tells us how different the PM and MM cells are.)
3. Use Wilcoxon's rank test to calculate a significance or *p*-value. (This tells us how confident we can be about a certain result.)
4. Compare the *p*-value with our preset significance levels to make the call.

Saturation

If a mismatch cell is saturated $MM \geq 46000$, the corresponding probe pair is not used in further computations. We also discard pairs where PM and MM are within *tau* of each other.

Discrimination Score

$$R_i = \frac{PM_i - MM_i}{PM_i + MM_i}$$

Computing *p*-values:

The one-sided Wilcoxon's Signed Rank Test

H_0: median$(R_i - \tau) = 0$

H_1: median$(R_i - \tau) > 0$

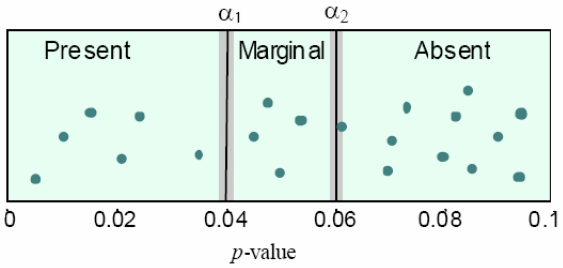
The null hypothesis is that the target is absent (zero effect on the probes).

Making the call

We set two significance levels α_1 and α_2 such that $0 < \alpha_1 < \alpha_2 < 0.5$

default $\alpha_1 = 0.04$ (16-20 probe pairs)

default $\alpha_2 = 0.06$ (16-20 probe pairs)



Significance levels α_1 and α_2 define cut-offs of *p*-values for making calls.

dChip: Filter Genes

1. A < SD/mean < B

A < SD (for logged data) < B

A gene is variable enough compared to its mean expression level to contain interesting information (> A), but not so variable that nothing can be learned (< B).

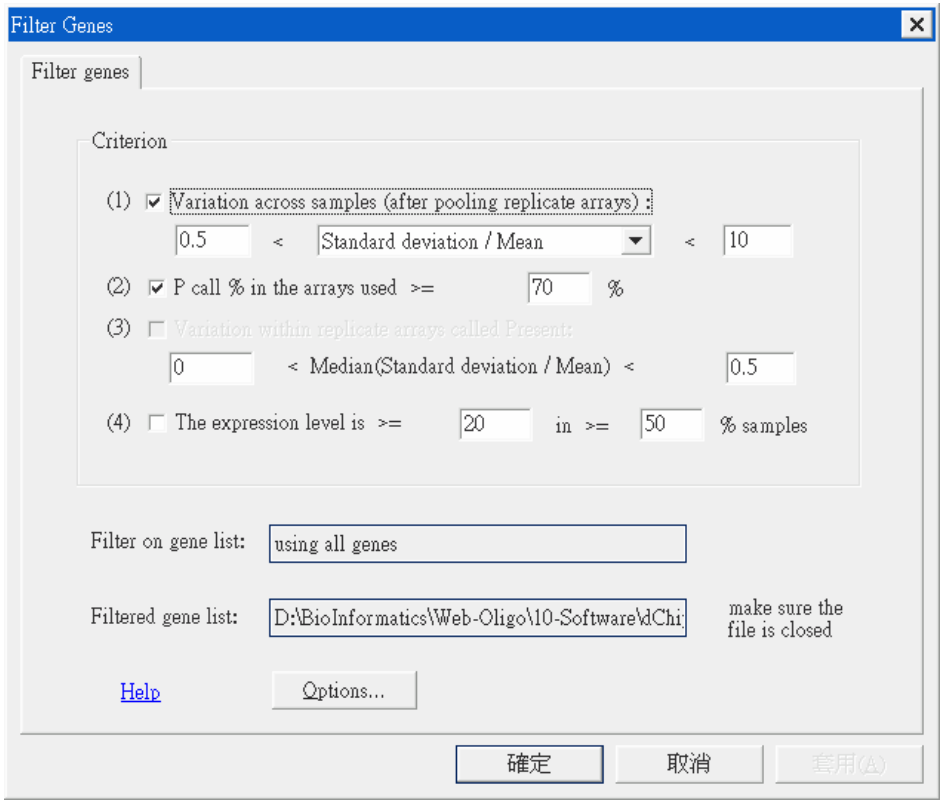
2. Presence call > X%

Narrows genes with a positive presence call in a certain percentage (> X%) of the samples.

3. A < Median(SD/Mean) < B

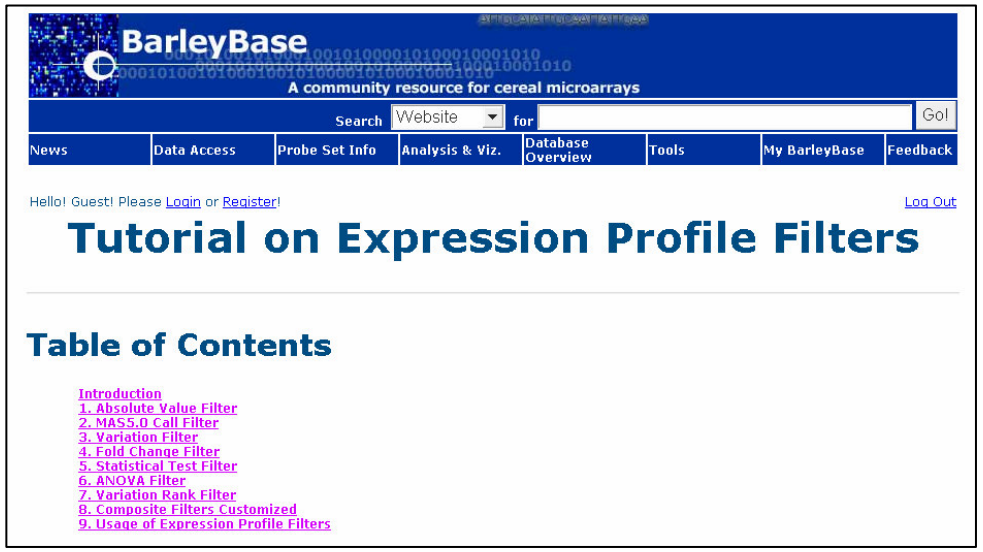
4. Expression level > Y in X%

Since low expression estimates are sometimes unreliable, we may want to limit our analysis to genes that are expressed above some threshold (>Y) in a certain percentage (X%) of the samples.

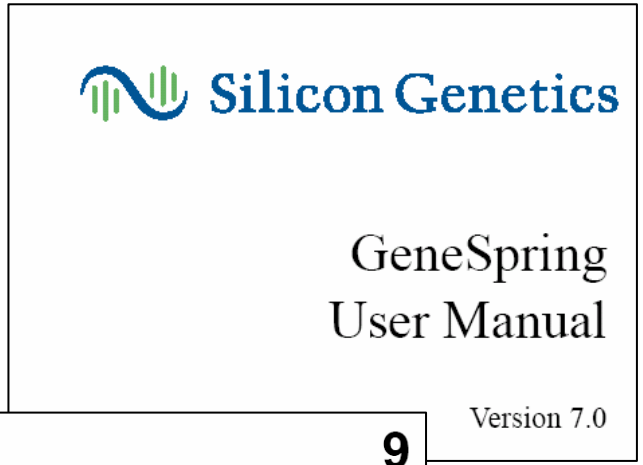


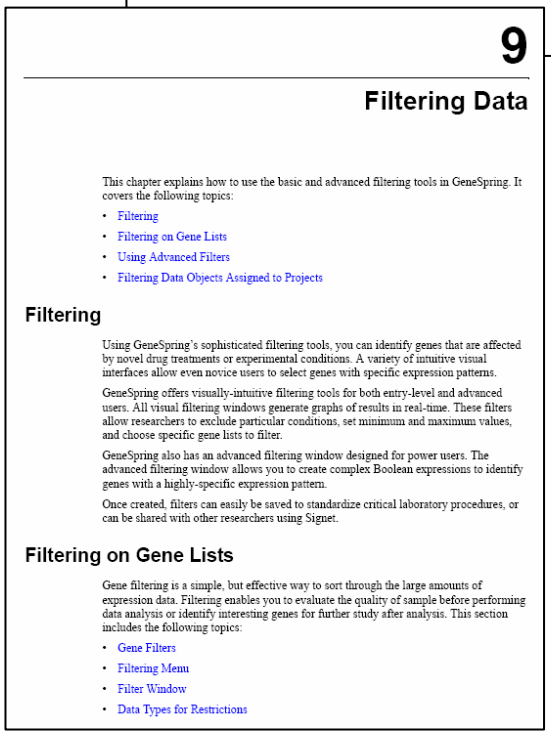
http://www.biostat.harvard.edu/complab/dchip/

Useful Reference



http://www.barleybase.org/filtertut.php





GeneSpring Tutorials
http://www.chem.agilent.com/Scripts/Generic.ASP?lPage=34743&indcol=Y&prodcol=Y

GeneSpring User Manual
http://www.chem.agilent.com/cfusion/faq/faq2.cfm?subsection=78&section=20&faq=1118&lang=en

Missing Values Estimation for Microarray Data

Missing values imply a loss of information

- Many analysis techniques that require complete data matrices: such as hierarchical clustering, k-means clustering, and self-organizing maps.
- May benefit from using more accurately estimated missing values.

Possible Solution

1. Exclude missing values from subsequent analysis.

May be of scientific interest !

2. Repeat the experiment **Expensive.**
3. Missing values in replicated design.

4. Adjust dissimilarity measures. (e.g., pairwise deletion.)
5. Modify clustering methods that can deal with missing values.

6. Imputation of missing values.

Sources of Missing Values

- **Various Reasons**
  - a feature of the robotic apparatus may fail,
  - a scanner may have insufficient resolution,
  - simply dust or scratches on the slide (image corruption),
  - spots with dust particles, irregularities, …

- **Mathematical transformation**
  - undefined mathematical transformed:
    e.g., corrected intensities values that are negative or zero, a subsequent log-transformation will yield missing values.

- **Flag**

spots may be flagged as *absent* or *feature not found* when nothing is printed in the location of a spot.

- the imaging software cannot detect any fluorescence at the spot,
- expression readings that are barely above the background correction,
- the expression intensity ratio is undefined: */0, 0/*.

GenePix

Good=100. Bad=-100. Not Found=-50. Absent=-75. unflagged=0.

Statistical Classification of Missing Data

It helpful to classify missing values on the basis of the **stochastic mechanism** that produces them.

Missing Completely At Random (MCAR)

- Missingness is **independent** of their own unobserved values and the observed data.
- Arising from chance events that are **unrelated to the nature** of the investigation.
- **e.g.,** A spot that is obscured accidentally by a dust particle.

Missing At Random (MAR)

- Missingness does not depend on their on unobserved value but does dependent on the observed data.

Missing Not At Random (MNAR)

- Missingness depend on their own unobserved values.
- missingness depdents on the fact that their raw intensity values are zero or small.
- **e.g.,** Spots that show no fluorescence or that have undefined log-intensities because their background-corrected intensities are negative.

- The missing values may give clues to systematic aspects of the problem.
- If missing values do occur by chance among a set of replicates, the observed members of the set can stand in for the missing, albeit with some loss of statistical precision.

Imputation: methods rely on the missingness being of the **MCAR** type.

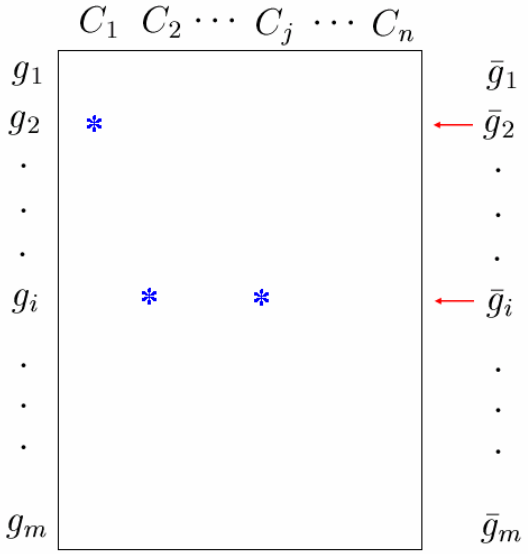
Imputation of Missing Values

- Missing log2 transformed data are replaced by *zeros* or by an *average* expression over the row ("row average“).

- Row average assumes that the expression of a gene in one of the experiments is similar to its expression in a different experiment, which is often not true in microarray experiments.

  - **Main weakness:**
    - it makes no serious attempt to model the connection of the missing values to the observed data.
    - since these methods do not take into consideration the correlation structure of the data.
    - not very effective (Troyanskaya et al, 2001)

  - **Useful:** where an initial imputation is required an iterative imputation method.



$$\widehat{C_j(g_i)} = \bar{g}_i$$

$i = 1, 2, \cdots, m.$

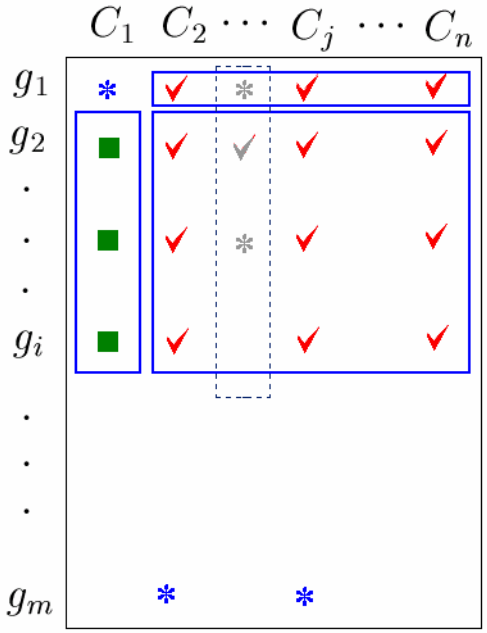
$j = 1, 2, \cdots, n.$

zero's
row average
row median

K-Nearest Neighbors Imputation

KNNImpute: a missing value estimation method to minimize data modeling assumptions and take advantage of the correlation structure of the gene expression data.

- Results are adequate and relatively insensitive to values of k between 10 and 20. (Troyanskaya et al, 2001)
- Euclidean distance appeared to be a sufficiently accurate norm.



KNNimpute

Model:

$$\{g_{(k)}, k = 1, 2, \cdots, K\} = \underset{k}{\text{args}} \max_{i \in C} \text{Corr}(g_1, g_i)$$

$$\{g_{(k)}, k = 1, 2, \cdots, K\} = \underset{k}{\text{args}} \min_{i \in C} \text{Dist}(g_1, g_i)$$

C: Observed C_i's without missing values

Imputation:

Average $\quad \widehat{C_1(g_1)} = \frac{1}{K} \sum_{k=1}^{K} C_1(g_k)$

Weighted Average $\quad \widehat{C_1(g_1)} = \frac{\sum_{k=1}^{K} w_k C_1(g_k)}{\sum_{k=1}^{K} w_k}$

$$w_k = \frac{1}{\sum_{i \in C} [C_j(g_k) - C_1(g_1)]^2}$$

- Euclidean distance measure is often sensitive to outliers, which could be present in microarray data.
- Log-transformed data seems to sufficiently reduce the effect of outliers on genes similarity determination.

Evaluation of Imputation Methods

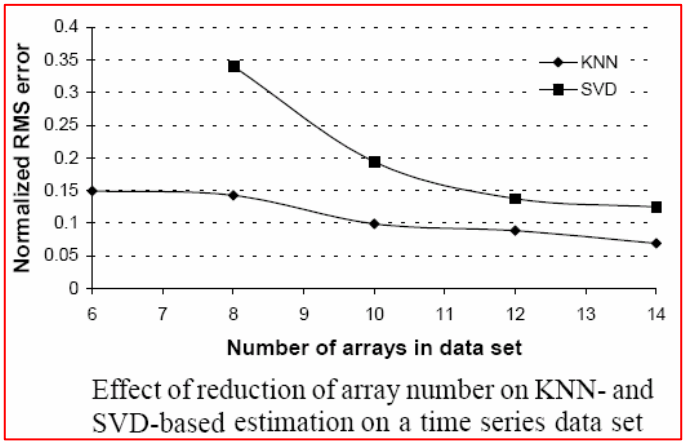
Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. (**2001**), Missing value estimation methods for DNA microarrays. Bioinformatics 17(6), 520-525.

Data sets:

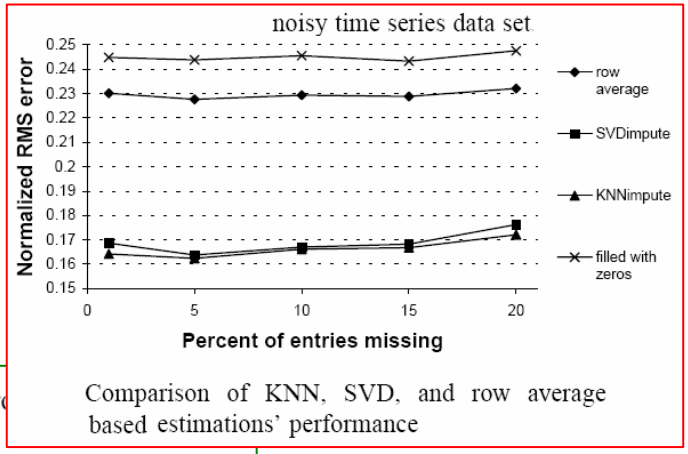
- Non-Time Series: Gasch et al., (2000): 755 genes, 173 arrays.
- Time Series: DeRisi et al., (1997): 6135 genes, 7 arrays.
- Noisy Times Series: Spellman et al., (1998): 509 genes, 77 arrays.

Criterions: normalized root mean squared error (NRMSE)

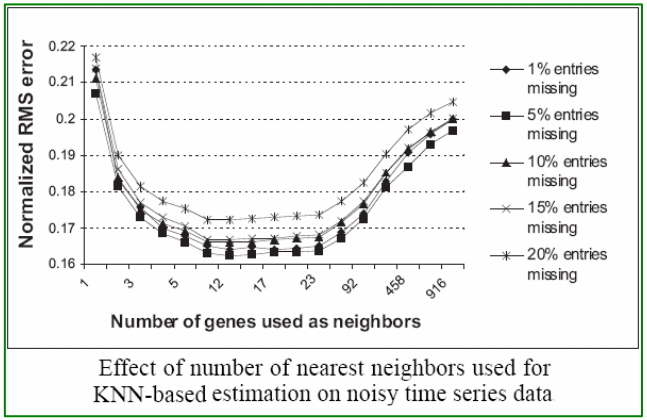
$$\mathrm{NRMSE} = \frac{\sqrt{\mathrm{mean}[(\mathrm{y}_{guess} - \mathrm{y}_{ans})^2]}}{\mathrm{std}[\mathrm{y}_{ans}]}$$



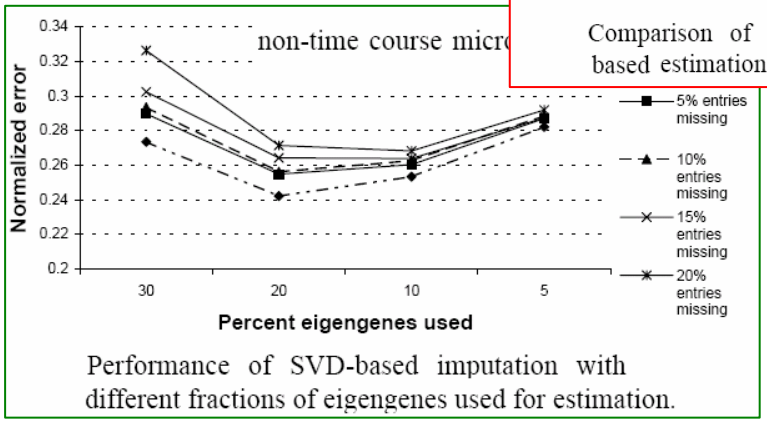
Effect of reduction of array number on KNN- and SVD-based estimation on a time series data set



Comparison of KNN, SVD, and row average based estimations' performance



Effect of number of nearest neighbors used for KNN-based estimation on noisy time series data



Performance of SVD-based imputation with different fractions of eigengenes used for estimation.

Reference

Singular Value Decomposition Imputation

- Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. (2001), Missing value estimation methods for DNA microarrays. Bioinformatics 17(6), 520-525.
- Trevor Hastie , Robert Tibshirani, Gavin Sherlock , Michael Eisen , Patrick Brown , David Botstein. (1999). Imputing Missing Data for Gene Expression Arrays, Technical Report.

Local Least Square Imputation

- Bo TH, Dysvik B, Jonassen I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Res. 2004 Feb 20;32(3):e34.
- Hyunsoo Kimy, Gene H. Golubz, and Haesun Parky. (2004). Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation, Bioinformatics Advance Access published August 27, 2004.

Bayesian

- Oba S, Sato M-A, Takemasa I, Monden M, Matsubara K-I, Ishii S: A Bayesian missing value estimation method for gene expression profile data,. Bioinformatics 2003, 19:2088-2096.
- Zhou X, Wang X, Dougherty ER: Missing-value estimation using linear and non-linear regression with Bayesian gene selection. Bioinformatics 2003, 19:2302-2307.

GMCimpute

- Ouyang M, Welsh WJ, Georgopoulos P. Gaussian mixture clustering and imputation of microarray data. Bioinformatics. 2004 Apr 12;20(6):917-23. Epub 2004 Jan 29.

Others

- Kim KY, Kim BJ, Yi GS. Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics. 2004 Oct 26;5(1):160.
- Shmuel Friedland, Amir Niknejad, and Laura Chiharaz. (2004). A Simultaneous Reconstruction of Missing Data in DNA Microarrays, Institute for Mathematics and its Applications,.
- Alexandre G de Brevern, Serge Hazout and Alain Malpertuy. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering, BMC Bioinformatics Volume 5.