

Tutorial: IMM 101

basic introduction on how to use IMM
this tutorial assumes you have a windows machine

Ka Yee Yeung

July 28, 2003

What is IMM?

- Infinite Mixture Model-based clustering algorithm
- The algorithm is developed by Mario Medvedovic, who wrote the IMM code in C++

Why do I want to use IMM?

- IMM produces better clusters than other heuristic clustering algorithms, especially when there are repeated measurements.
- See our paper in Genome Biology 2003 4(5): R34 at <http://genomebiology.com/2003/4/5/R34>
- Also see Medvedovic *et al.* in Bioinformatics 2002, 18: 1194-1206

7/29/2003

3

Quick intro to IMM

- Model-based approach
 - Assume data points from each cluster are generated from a multivariate Gaussian distribution.
- Built-in error model
 - Assume the repeated measurements are generated from another multivariate Gaussian distribution.
- Infinite mixture model:
 - Do not assume the number of clusters

7/29/2003

4

Output of IMM

- Output: pairwise probabilities (P_{ij}) for two genes (i,j) to belong to the same cluster
- Advantages of IMM:
 - These pairwise probabilities are generated using the entire dataset, using repeated measurements as well.
 - Cluster structure is clearer using these pairwise probabilities than simple Euclidean distance
- To form clusters: cluster P_{ij} with a heuristic clustering algorithm (eg. complete-link)

7/29/2003

5

Limitations of IMM

- Current status: Command-line program without a graphical interface.
- Multi-step process.
- Assume input data has no missing values.
- Require repeated measurements of data, *not* just estimated error of measurements.

7/29/2003

6

What should I do if my data has missing values?

- There are 2 options:
 - If a significant fraction of the data does **not** contain any missing values, you might filter out the genes (if you are clustering the genes) with missing values.
 - If most data points contain missing values, you can **impute** missing values.

7/29/2003

7

How to impute missing values?

- A literature of its own.
- In our GB paper, we adopted the *k-nearest neighbor approach* [Troyanskaya et al. 2001]
 - Manuscript: <http://bioinformatics.oupjournals.org/cgi/reprint/17/6/520.pdf>
 - Software "KNNimpute" is available from <http://smi-web.stanford.edu/projects/helix/pubs/impute/>
 - KNNimpute: a command-line program written in C

7/29/2003

8

Software requirements

- To run the IMM executable written by Mario Medvedovic, you need a windows machine.
- To run the java bytecode files written by Ka Yee Yeung, you need **Java SDK 1.4 or above** on any your favorite OS:
 - Windows/Linux: <http://java.sun.com/j2se/1.4.2/download.html>
 - Mac OS10: should come with 10.2
- To run the **Perl** scripts written by Ka Yee:
 - Windows: <http://www.activeperl.com/>
 - Linux: <http://www.perl.com/CPAN/>
 - Mac OS10: should come with 10.2

7/29/2003

9

To-do list before you start using IMM

- Install Java 1.4 or above
- Install Perl
- Install WinZip (optional)
- Download IMM from <http://homepages.uc.edu/~medvedm/BioinformaticsSupplement.htm>
- Download Perl scripts + Java bytecode files "pre_post_process.tar.gz" from <http://expression.microslu.washington.edu/expression/kayee/yeunggb2003.html>

7/29/2003

10

Steps of using IMM

1. Message input data → ".inp" format
 - Perl script written by Ka Yee Yeung
2. Run IMM
 - C++ code written by Mario Medvedovic
3. pairwise probabilities → similarity matrix
 - Perl script written by Ka Yee Yeung
4. Cluster the similarity matrix
 - Java code written by Ka Yee Yeung
 - Alternatively, you can write your own code or use your favorite clustering program

7/29/2003

11

Ka Yee's favorite input file format

Example: test_10gene_5expt_2rep.txt

*tab-delimited text file, can be viewed in Excel

		Header row									
YORF	NAME	expt1	expt1	expt2	expt2	expt3	expt3	expt4	expt4	expt5	expt5
G1	G1	0.98	0.99	0.78	0.78	0.54	0.55	-0.12	-0.13	-1.2	-1.25
G2	G2	0.97	0.98	0.78	0.8	0.55	0.55	-0.12	-0.13	-1.2	-1.25
G3	G3	0.98	0.98	0.77	0.78	0.55	0.56	-0.12	-0.11	-1.2	-1.3
G4	G4	0.99	0.98	0.79	0.8	0.56	0.55	-0.11	-0.13	-1.2	-1.21
G5	G5	0.97	0.97	0.78	0.77	0.54	0.54	-0.12	-0.12	-1.1	-1.2
G6	G6	0.16	0.16	-0.11	-0.15	-0.78	-0.76	1.3	1.4	0.54	0.53
G7	G7	0.16	0.16	-0.16	-0.16	-0.76	-0.76	1.35	1.4	0.53	0.53
G8	G8	0.16	0.17	-0.15	-0.14	-0.78	-0.77	1.36	1.3	0.55	0.54
G9	G9	0.15	0.16	-0.16	-0.16	-0.78	-0.76	1.35	1.3	0.53	0.54
G10	G10	0	-0.03	2.8	2.43	-0.8	-0.84	0.02	0.03	1.67	1.68

2 documentation columns

2 repeated measurements for gene G10 under expt1

7/29/2003

12

1. Convert to “.inp” format

- This step can be skipped if your data is already in the “.inp” format.
- If your input data is in KaYee’s favorite input format, you can use her Perl script “txt2inp_win.pl”:

```
perl txt2inp_win.pl <input filename> <# experiments> <# replicates>
```

 - **Example:**

```
perl txt2inp_win.pl test_10gene_5expt_2rep.txt t 5 2
```
- Output: “test_10gene_5expt_2rep.inp”

7/29/2003

13

Steps of using IMM

1. Massage input data → “.inp” format
 - Perl script written by Ka Yee Yeung
2. Run IMM
 - C++ code written by Mario Medvedovic
3. pairwise probabilities → similarity matrix
 - Perl script written by Ka Yee Yeung
4. Cluster the similarity matrix
 - Java code written by Ka Yee Yeung
 - Alternatively, you can write your own code or use your favorite clustering program

7/29/2003

14

2. Run IMM

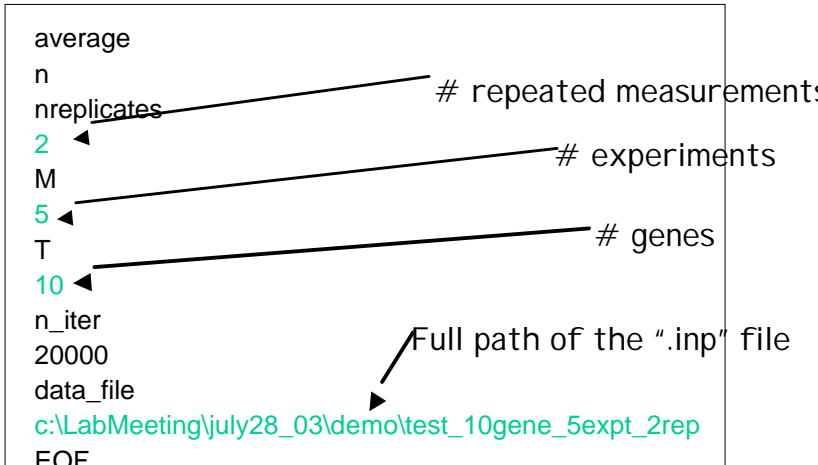
- GI MM.exe:
 - parameters.prm
- post_hoc.exe:
 - posthoc_parameters.prm
- What are the ".prm" files?
 - They are the parameter files that specify
 - The size of the data: # genes, # experiments, # repeated measurements etc.
 - Convergence parameters (Unless you know what you are doing, please leave them alone.)

7/29/2003

15

Example: "parameters.prm"

```
average
n
nreplicates # repeated measurements
2
M # experiments
5
T # genes
10
n_iter
20000
data_file
c:\LabMeeting\july28_03\demo\test_10gene_5expt_2rep
EOF
```

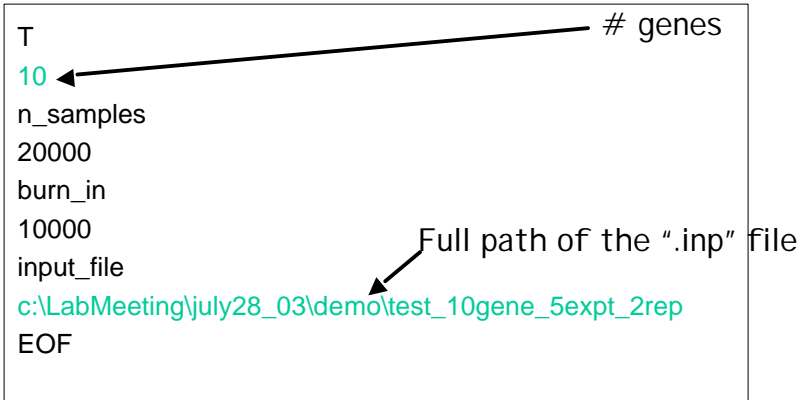


7/29/2003

16

Example: "posthoc_parameters.prm"

```
T # genes
10
n_samples
20000
burn_in
10000
input_file
c:\LabMeeting\july28_03\demo\test_10gene_5expt_2rep
EOF
```



7/29/2003

17

Running IMM

1. Edit the ".prm" files with your favorite text editor (emacs, notepad, word etc.)
 - # genes
 - # experiments
 - # repeated measurements
 - Full path of the files without the suffix
2. Double-click on `GIMM.exe`
 - This process can take a good while... I t's now time for a coffee.
3. Double-click on `post_hoc.exe`
4. Output: "test_10gene_5expt_2rep.zm"

7/29/2003

18

Steps of using IMM

1. Massage input data --> ".inp" format
 - Perl script written by Ka Yee Yeung
2. Run IMM
 - C++ code written by Mario Medvedovic
3. pairwise probabilities --> similarity matrix
 - Perl script written by Ka Yee Yeung
4. Cluster the similarity matrix
 - Java code written by Ka Yee Yeung
 - Alternatively, you can write your own code or use your favorite clustering program

7/29/2003

19

3. IMM output → similarity matrix

- Perl script "zm2simMatrix_win.pl" converts the output zm file from "post_hoc" to a similarity matrix for clustering algorithms.

```
perl zm2simMatrix_win.pl <zm filename> <# genes> <# samples> <# burnin>
```

- <# samples> is equivalent to "n_samples", and <# burnin> is equivalent to "burn_in" in the parameter file "posthoc_parameters.prm"

- **Example:**

```
perl zm2simMatrix_win.pl test_10gene_5expt_2rep.zm 10 20000 10000
```

Output: sim_matrix.txt

7/29/2003

20

Steps of using IMM

1. Massage input data → “.inp” format
 - Perl script written by Ka Yee Yeung
2. Run IMM
 - C++ code written by Mario Medvedovic
3. pairwise probabilities → similarity matrix
 - Perl script written by Ka Yee Yeung
4. Cluster the similarity matrix
 - Java code written by Ka Yee Yeung
 - Alternatively, you can write your own code or use your favorite clustering program

7/29/2003

21

4. Cluster the similarity matrix

- Apply hierarchical average-link or complete-link to cluster the pairwise similarities from IMM.
- To run the algorithm:

```
java hieclustSim -r <# genes> -NoC_range <range of # clusters> -step <step size> -alg <algorithm> -doc <doc file> [sim_matrix.txt]
```
- Example: to produce 2 to 3 clusters from our similarity matrix using average-link

```
java hieclustSim -r 10 -NoC_range 2 3 -step 1 -alg avg -doc test_10gene_5expt_2rep.txt sim_matrix.txt
```
- Output: OutSim_AvgLink_2.txt & OutSim_AvgLink_3.txt

7/29/2003

22

5. Interpretation of results

- Output from autoHieClust:
 - A tab-delimited text file, which can be viewed in Excel
 - Genes in the same cluster are assigned the same cluster number
 - Can concatenate with the raw data file
 - Column A: order in which genes appear in the raw data input file
 - Sort on Column A
 - Copy & Paste expression values

7/29/2003

23

Visualization of clusters 1 (if you have Matlab)

- Plots the average expression profile of each cluster
- Compute the average expression value over all repeated measurements
 - Perl script
avgRepData_win.pl <filename> <#expts> <# rep>
 - Example:
perl avgRepData test_10gene_5expt_2rep.txt 5 2
 - Output: avg_test_10gene_5expt_2rep.txt

7/29/2003

24

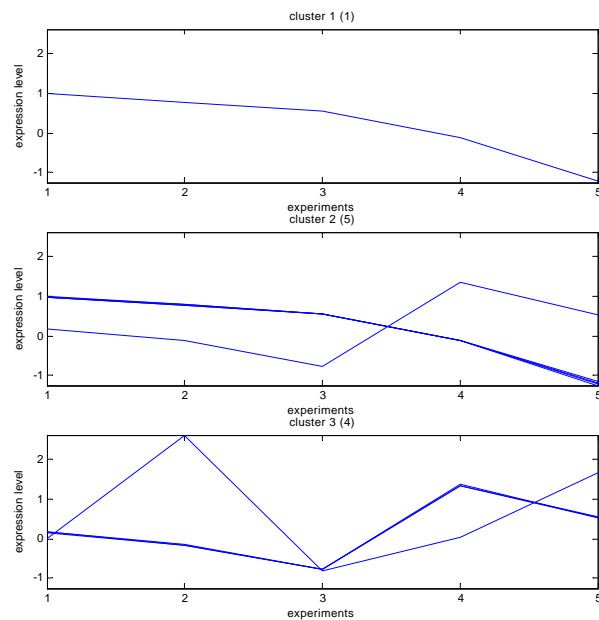
Visualization of clusters 2 (if you have Matlab)

- Matlab script "viewClusterResult.m"
- viewClusterResult (avg_data_filename, #genes, #expt, #Cluster, cluster_file, NoC_lo, NoC_hi)
- Example:
 1. Start Matlab
 2. Change to the directory containing the cluster output file and "viewClusterResult.m"
 3. Start the script:

```
>> viewClusterResult  
('avg_test_10gene_5expt_2rep.txt', 10, 5, 3,  
'OutSim_AvgLink_3.txt', 1, 3);
```

7/29/2003

25



7/29/2003

26