CRC tutorial

Steve Qin

10/13/06

This document provided a guided tour through the main functions of CRC.

1. Introduction

Chinese restaurant cluster (CRC) is a model-based clustering tool for analyzing microarray gene expression data. CRC is designed specifically for clustering genes according to their expression profiles across multiple experiments.

2. Input data

CRC assumes that the input data has already been properly preprocessed. For example, data have been log-transformed and if multiple probe-sets represent one gene, the expression level of that gene has been summarized. But Noted that no cross-chip normalization step is necessary and sporadic missing data is allowed.

CRC requires a simple input data format which is a tab delimited plain text file. No header line is allowed, and each row represents one gene. The first column contains all the gene names, and each subsequent column contains gene expression level of that gene under a specific experimental condition. Missing data is denoated as -999. The beginning part of a sample input file can be found in Box 1.

Box 1. Sample CRC input file.

```
CDC19  -0.118 -0.729 -0.136 -0.202 -0.036 0.248 0.395 -0.034 -0.127 0.024 -0.088 -0.31 -0.066 -0.292 -0.049 -0.143 0.446 -0.36 -0.255 -0.795
HAP3  -9999 0.157 -0.365 0.152 -0.479 -0.054 0.035 0.223 0.27 -0.353 0.052 -0.116 -0.264 0.14 -0.134 0.0 0.225 0.248 0.199 -0.146
RRN10  0.175 0.133 -0.341 0.064 -0.314 0.055 0.123 0.189 0.358 -0.216 0.127 -0.165 -0.083 0.114 -0.13 -0.0040 0.117 0.168 0.271 0.354
RPS8A  -0.018 -0.461 -0.212 -0.458 -0.392 0.0080 -0.164 -0.422 -0.593 -0.162 -0.527 -0.249 -0.457 -0.427 -0.523 -0.239 -0.407 -0.389 -0.445 -0.645
RPL23A  -0.244 -0.488 -0.243 -0.435 -0.358 -0.085 -0.145 -0.44 -0.417 0.048 -0.631 -0.297 -0.631 -0.479 -0.655 -0.357 -0.494 -0.379 -0.458 -0.796
…
```

A useful pre-processing step is to filter out genes that are of little interest for clustering. For example, some genes show little variation across all experimental conditions. They may either not expressing or expressing at a constant rate under all the experimental conditions. Therefore including those genes in the clustering analysis will not produce desirable results. A commonly used filtering strategy is described in Tamayo et al. 1999. which requires that the absolute differences (|maximum expression level-minimum expression levell) is higher than a threshold and the relative differences (|maximum expression level/minimum expression levell) is higher than another threshold. To perform this filtering, the users may use the perl script "filter.pl" in the package provided or at the CRC website. The syntax for this command is:

Perl filter.pl [input file name] [output filename]

3. Run CRC

After input file has been created. We run CRC to generate clustering result. The syntax for this command is the follows:

./crc [input file name] [output file name] num_chains num_cycles invertion_flag max_shift prob_cutoff

num_chains: integers, number of parallel chains to run. Higher number indicates that the MCMC algorithm is more likely to find a solution that has higher likelihood. This argument may be tuned to achieve the best performance while maintaining acceptable computation cost.

num_cycles: integers, number of cycles to run in each Markov chain. Higher number indicates that the chain will run longer which give it better chance to navigate to the solution with higher likelihood. This argument may be tuned to achieve the best performance while maintaining acceptable computation cost.

Invertion_flag:integer: 0 or 1. 0 means use no invert relations, and ignore nonsynexpression relationships such as inverted or time-shifted. 1 means full model, complex relationships are considered. When the experiments are independent, no inversion flag is recommnded; while the experiments are conducted over time, then it ispossible that genes display complex correlation pattern, therefore the inversion flag may be turned on to include complex correlation structure into consideration.

max_shift: integer: 0,1,2,… 0 means no shift. When the experiments are independent, no shift is recommnded; while the experiments are conducted over time, then it ispossible that genes display time shift expression pattern, therefore non-zero shifts are recommended to include complex correlation structure into consideration.

prob_cutoff: floating point 0-1, indicates the posterior probability threshold for a gene to be included in a cluster in the final result. 0.9 is a stringent threshold, and 0.5 is a liberal one. For an exploration study, a lower threashold is recommneded to consider more genes as likely grouped together. For a study that need accurate functional annotation, a higher threshold is recommended, such that loose genes will be filtered out and the cluster will look more compact.

4. Analyzing CRC clustering results

After CRC clustering is completed. One can used various output files to generated reports and figures. The main output file is shown in Box 2. At the beginning of the file, the basic information is summarized: number of genes, number of experiments, proportion of missing expression levels and number of clusters. alpha is the tuning parameter in the Dirichlet process prior distribution.

In the section summarizing all clusters, log Bayes ratio indicates the tightness of the clusters. Higher value indicates stronger similarity in the expression profiles of genes in this cluster. Average co-occurrences index indicates the stability of cluster members. Higher value indicates stronger correlation among genes in this cluster.

Each subsequent line provides information of a gene in this cluster. The first item is the order of this gene, followed by the gene name, the integer inside () indicates whether the correlation relationship between this gene and its cluster is synexpresion. “+” indicates positive relation, “-“ indicates other wise. “0” indicates simultaneous correlation, others indicate amount of shift. The value inside [] is the posterior probability that this gene belong to this cluster. If missing data is contained in this gene, “*” will be placed by this probability to warn the user that this probability need to be intepret with caution since missing data will affect the calculation of this value.

Box 2. A sample CRC output file.

```
*****************************
*                           *
*      CRC 1.0 clustering Result      *
*                           *
*****************************

Total number of genes: 205
Total number of experiments: 20
Rate of missing data: 0%

There are total of 4 clusters
apha = 1.  posterior probability threshold = 0.5
log likelihood = 2883.88

cluster 1 size = 93
log Bayes ratio = 26.5906
average co-occurrences =  1

1 HAP3 (+0) [0.999999]
2 RRN10 (+0) [0.990918]
5 PRP6 (+0) [0.999989]
6 TFC1 (+0) [0.999945]
8 NTC20 (+0) [0.999958]
…
```

A files named “member.txt” will be generated automatically. This file can be used to generate a trace plot as shown in Figure 1. As shown in Box 3, "member.txt" has the following format: The first column is the cluster ID (e.g., order), the second column is the geneID, the third column is gene name, and the last column indicated the correlation relationship: 1 indicates synexpressionl; 2 indicates time-shifted; 3 indicates inverted; 4 indicates time-shifted and inverted. This file will be used by the provided R program plot.r to generate trace plots.

Box 3. Sample output file "member.txt".

1	974	BA4479	1
1	975	BA4480	1
1	1143	BA5159	1
1	1165	BA5219	1
1	1166	BA5220	1
2	1242	BA5569	1
2	1246	BA5591	2
2	1302	pXO1-16	2

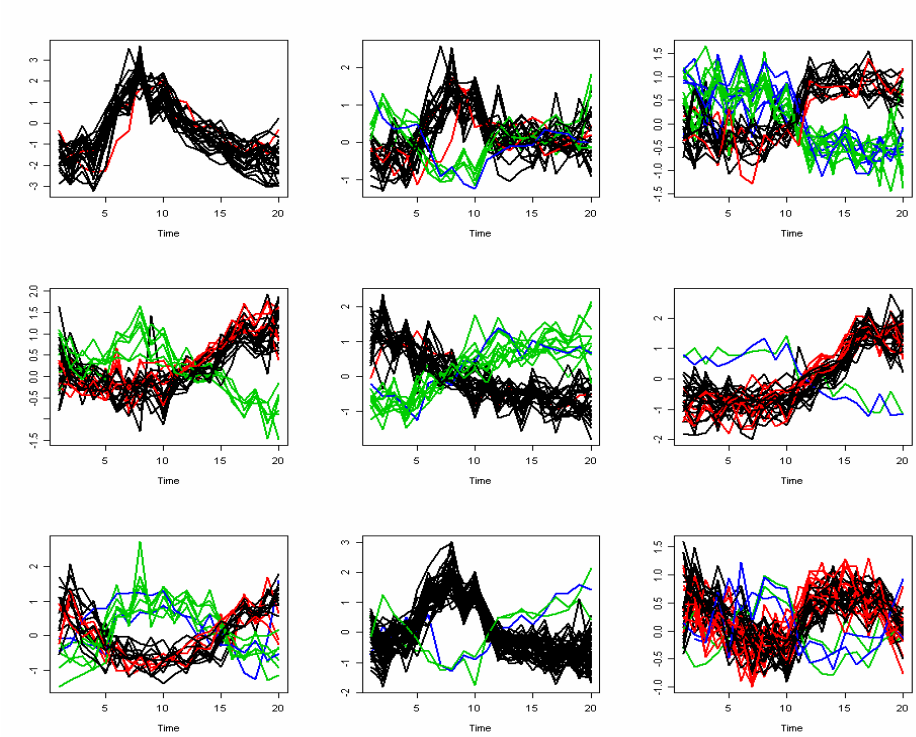

Figure 1. Trace plots of 9 clusters. Black line represent genes display no shift and positive correlation with the cluster pattern; green line represent genes display no shift but negative correlation with the cluster pattern; red line represent genes display shift and positive correlation with the cluster pattern; blue line represent genes display shift and negative correlation with the cluster pattern.

Another file named “like.txt” will be generated automatically contains all the likelihood value at the end of each iteration. This will be used by the provided R program diag.r to draw trace plots (Figure 2) and autocorrelation plots to performe diagnostic check on convergence.

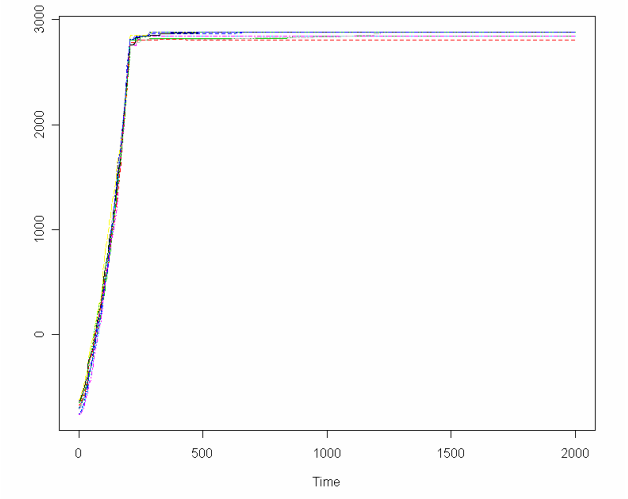

Figure 2. Marginal likelihood trace plots of 10 independent Markov chains.

5. Further questions.

We are committed to make CRC a powerful and easy-to-use clustering tool. More functions will be added. If you have any question whiling using this tool, please contact Steve Qin at qin@umich.edu.

6. Reference

Qin ZS (2006) Clustering microarray gene expression data using weighted Chinese restaurant process. Bioinformatics. **22**. 1988-97.