CRC Users' Guide
Steve Qin
Last update: Oct 13, 2006

Introudction

CRC (Chinese restaurant cluster) implements a model-based Bayesian clustering algorithm where assignment procedure can be regarded as following a iterative weighted Chinese restaurant process. This program is designed to cluster microarray gene expression data collected from multiple experiments. Cluster assignments and total cluster number are inferred simultaneously. An advantage of CRC is that, for data collected over a time course, CRC is able to cluser genes displaying synexpression (positive and simultaneous) and non-synexpreesion (inverted and/or time-delayed) (Qian et al. 2001. *J Mol Biol* 2001; **314**: 1053-1066) relationships together in the same cluster. Other notable fatures including automatic handling of missing data, and various measures to assess the confidence of cluster assignments and cluster quality. The program is written in C++, and it is available for Linux, Unix, Windows, MAC OSX operating systems as a command line exeutable.

Command

The full command is
./crc input_filename output_filename num_chains num_cycles invertion_flag max_shift prob_cutoff.

input_filename:	string of characters, input file name.
output_filename:	string of characters, output file name.
num_chains:	integers, number of parallel chains to run. Recommanded value is 10.
num_cycles:	integers, number of cycles to run in each Markov chain. Recommanded value is 20.
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.
max_shift:	integer: 0,1,2,… 0 means no shift.
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 experiments conducted over time, such as cell cycle experiments, where complex correlation relationships other than synexpression, e.g., inverted and/or time-delayed, may be expected, it is recommended that the pattern selction option of the program to be turned on. That is, set invertion_flag and max_shift to 1. Otherwise, fix invertion_flag and max_shift to 0.

After CRC runs, one can run the two R program to obtain cluster time series plots and convergence diagnostic trace plot and autocorrelation plot.
R CMD BATCH plot.r plot.out.txt

R CMD BATCH diag.r diag.out.txt

To do this, R software is required. R is an open source statistics package, and can be downloaded for free at http://www.r-project.org/.

Input file format

Input file contains gene expression profiles. Each line represents a gene. The first column is gene name or gene ID. Each subsequent column represents an experiment, stores gene expression level of that gene observed from that experiment. Entries are separated by one or more white space. Missing data is represented by -9999. a sample input file is shown below:

```
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
…
```

Output file format

Output file has the following format:

```
******************************
*                            *
*      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]
…
```

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. Posterior probability threshold is the cutoff value such that only genes having posterior probability greater than this value will be included in the results. This is to filter out loose genes and to make the cluster more compact and specific. The marginal log likelihood value of the clustering output is also displayed.

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.

Other additional output files

A files named "member.txt" will be generated automatically, it has the followun format:

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

The first column is the cluster ID (e.g., order), the second column is the geneID (e.g., order), 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.

A 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 and autocorrelation plots to performe diagnostic check on convergence.

Reference

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

Contact

Comments, suggestions, questions are welcomed, and should be directed to Steve Qin. Email: qin@umich.edu. Phone: 734-763-5965.