Spectrum

A software for inferring population structure and recombination events

Kyung-Ah Sohn and Eric P. Xing
School of Computer Science
Carnegie Mellon University
{ksohn,epxing}@cs.cmu.edu

1 Overview

Spectrum is a software for joint inference of population structure and recombination events from multi-locus SNP haplotypes. Under non-parametric Bayesian framework using Hidden Markov Dirichlet process, the genetic inheritance process under mutation and recombination event is inferred. Assuming a number of founder haplotypes, it recovers the association of each individual haplotype with founders across all the loci so that the individual frequency vector with respect to recovered founders can be estimated as well as the recombination rate. Moreover, users need not specify the number of founders in advance since the number itself is inferred as the result of inference under Dirichlet process prior. Details of the algorithm can be found in [1].

The program was written in C++ and the executables (for Linux and Windows platforms) are available at `http://www.cs.cmu.edu/~ksohn/Spectrum/`. After extracting the downloadable zip file, you can run the program by typing `./Spectrum InputFile` (in Linux), or `Spectrum.exe Inputfile` (in Windows). The following sections describe detailed input/output file format, and the program parameters that users can change.

2 Input file

An input file for *Spectrum* is a simple text file in which the first line contains two integers of the number of individuals I and the length of each sequence T. The rest part consists of a binary matrix for SNP sequences where rows correspond to individuals and columns correspond to SNPs. Data for each individual is represented as two consecutive rows, so the dimension of the data matrix is $2I$ by T. The following is a sample input file of data from 3 individuals with 5 SNPs.

```
3 5
0 0 0 0 0
1 1 0 0 0
0 1 0 1 0
0 0 0 0 0
0 1 0 1 0
1 1 0 0 1
```

3 Output file

Spectrum generates four kinds of information: recovered founder haplotypes, individual frequencies with respect to founders, founder frequencies along chromosome positions, and the estimated recombination rates.

- **Founders:**
  The first part shows the recovered founder haplotypes with their population frequencies and the founder-specific mutation rates. The frequency of each founder is computed as the total number of loci associated with that founder across all the individuals divided by the total number of loci across all the individuals (i.e. the sequence length $\times$ the number of individual $\times$ 2). The mutation rate means the ratio of loci in the associated individual haplotypes whose allele is different from that of the founder. Each founder is represented in each row, where each column indicates the label, frequency, mutation rate, and alleles of the founder, respectively.

[Founders]

ID	Frequency	Theta	Haplotype
0	0.241857	0.002704	1001010000
1	0.157530	0.011613	0110101100
2	0.121860	0.003080	0100010110
3	0.134923	0.005321	1111111111
4	0.343829	0.010483	0000000000

...

- **Individual Spectrum:**
  Individual spectrum is defined as the fractions of individual haplotype associated with recovered founders. In output file, each individual frequency vector is shown in a row with respect to each founders so that the element (i, k) corresponds to the frequency of founder k in the haplotype of individual i.

[Individual Spectrum]

0.0000	0.0161	0.0000	0.0000	0.0075	0.5313	0.4424	0.0027	0.0000
0.0000	0.0000	0.0000	0.0291	0.0000	0.0010	0.3362	0.0005	0.6332
0.0000	0.0473	0.0000	0.2808	0.0003	0.0000	0.6481	0.0235	0.0000
0.0000	0.0000	0.0000	0.3965	0.0000	0.0000	0.6035	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.5577	0.4397	0.0000	0.0026
0.0000	0.0000	0.0000	0.0000	0.0000	0.7162	0.2800	0.0000	0.0038
0.1400	0.0000	0.0000	0.6400	0.0000	0.0400	0.1800	0.0000	0.0000
0.0001	0.0000	0.0000	0.2799	0.0000	0.0000	0.7100	0.0100	0.0000
0.0000	0.5817	0.0000	0.0000	0.4183	0.0000	0.0000	0.0000	0.0000
0.2613	0.0637	0.3737	0.0000	0.1188	0.0000	0.0000	0.1824	0.0000
0.1147	0.0000	0.1053	0.0000	0.7800	0.0000	0.0000	0.0000	0.0000
0.4657	0.3104	0.2202	0.0000	0.0037	0.0000	0.0000	0.0000	0.0000

...

- **Founder Frequency:**
  Next, *Spectrum* shows population frequencies of each founder along chromosome positions. This can be computed for each locus as the ratio of individual haplotypes associated with each founder at that locus. Therefore, the result is written as K by T matrix where K is the number of recovered founders and T is the number of SNPs.

[Ancestor Frequency]

0.3458	0.3458	0.3000	0.3000	0.3000	0.3000	0.3000	0.3000	0.3000	0.3125
0.0792	0.0792	0.0792	0.0792	0.0792	0.0792	0.0792	0.0792	0.0792	0.0792
0.0000	0.0000	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.0208	0.1042
0.2875	0.2875	0.2708	0.2708	0.2708	0.2708	0.2708	0.2708	0.2708	0.1625
0.0542	0.0542	0.0875	0.0875	0.0875	0.0875	0.0875	0.0875	0.0875	0.0875
0.0917	0.0917	0.0875	0.0875	0.0875	0.0875	0.0875	0.0875	0.0875	0.1042
0.1417	0.1417	0.1458	0.1458	0.1458	0.1458	0.1458	0.1458	0.1458	0.1500
0.0000	0.0000	0.0083	0.0083	0.0083	0.0083	0.0083	0.0083	0.0083	0.0000

...

- **Recombination Rate:** The last part consists of the estimated recombination rate (per bp) between every pair of adjacent SNPs.

[rec]

1.4921e-005	1.1088e-005	1.5674e-005	3.3333e-004	1.5281e-005	2.1552e-005	1.3156e-005
2.3156e-005	5.9201e-006					

4 Program parameters

Some parameters of *Spectrum* can be changed by command-line arguments.

- -loci (LOCFILE): specify filename for physical location of each SNP (in bp). Default positions: every 1Kbp.
- -dscale (SCALEFACTOR): loci positions in (LOCFILE) with "-loci" option are expected to be in base pairs, but if not, users can specify the scale factor of the positions. For example, if positions in (POSFILE) are in range of (0,1) and the length of the entire sequence was $20K$, then use "-dscale 20000". Default: 1
- -r (RECRATE): mean recombination rate per bp used in Beta prior. It is recommended to try different magnitudes of prior recombination rate (e.g., 1.e-3, 1.e-4, ... 1.e-7) if the result from the default setting shows relatively large mutation rates (e.g. over 0.1) for some founders. Default: 1.e-5.
- -m (MUTRATE): mean mutation rate per bp used in Beta prior mutation rate. Default: 0.0001
- -nbr (NBURNIN) : the number of burnin iterations. Default: 4000.
- -nc (NCUM) : the number of cumulative iterations. Default: 2000.
- -dir (OUTDIR): directory name where output will be written. Default: "./output/InputFileBaseName/".

If input sequences are too long (e.g. over a few hundreds of SNPs), then convergence of *Spectrum* can be very slow. In this case, it is recommended to partition the sequences into blocks of moderate size and then merge the local result into one final estimation. This partitioning and merging can be done internally within *Spectrum* if user explicitly specifies the length of maximum block length (BLOCKLENGTH). Each intermediate result using (BLOCKLENGTH) loci will be written as well as the final estimation.

- -bl (BLOCKLENGTH): maximum block length to process internally as unit inference. Default: length of each sequence.

5 Plotting tool

Matlab files for plotting *Spectrum* output are also provided. Function PlotResult.m will produce three figures using individual frequencies, founder frequencies, and recombination rates, respectively and save as pdf files. Figure 1 shows the generated example figures using *Spectrum* output on HapMap four population dataset.

Reference

[1] Kyung-Ah Sohn and Eric P. Xing, Spectrum:Joint Bayesian Inference of Population Structure and Recombination Events. The Fifteenth International Conference on Intelligence Systems for Molecular Biology (ISMB 2007)

[2] Eric P. Xing and Kyung-Ah Sohn, Hidden Markov Dirichlet Process: Modeling Genetic Recombination in Open Ancestral Space, Bayesian Analysis, Volunn 2, Number 2, 2007

[3] Kyung-Ah Sohn and Eric P. Xing, Hidden Markov Dirichlet Process: Modeling Genetic Recombination in Open Ancestral Space, Advances in Neural Information Processing Systems 19 (NIPS 2006)

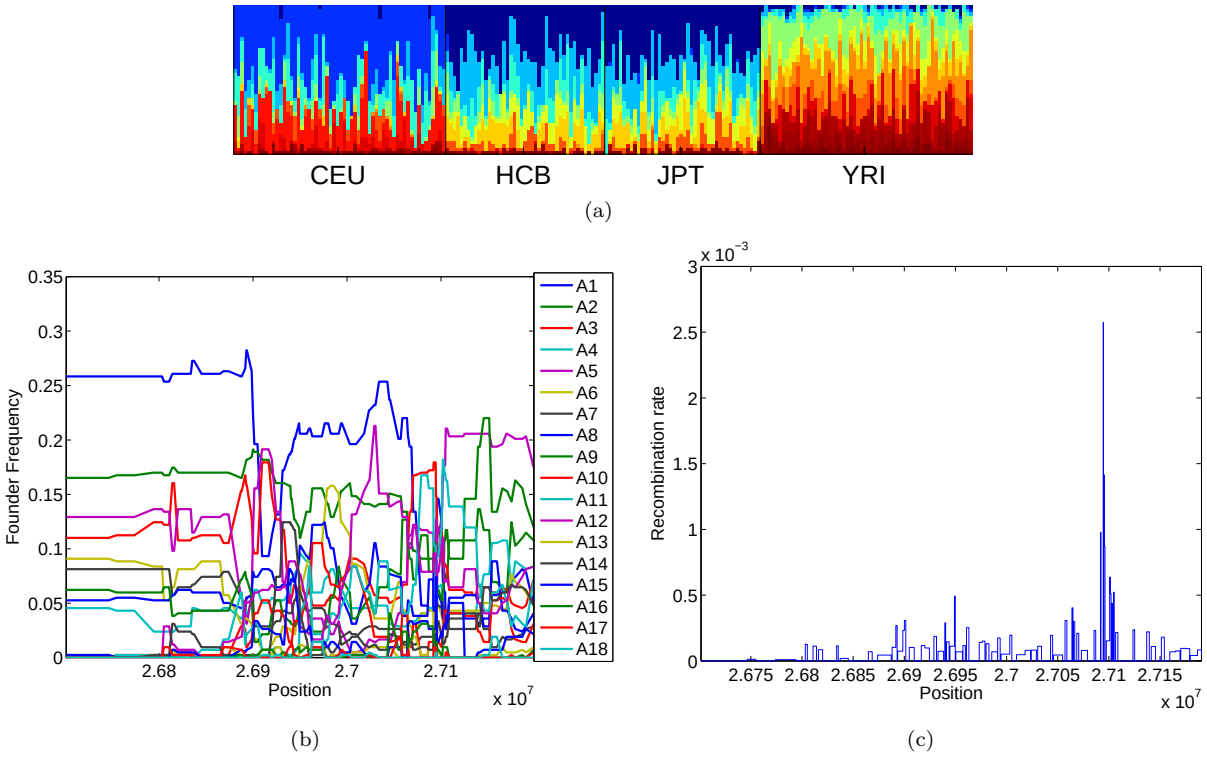

Figure 1: Example plots from Spectrum output. (a) Population structure in terms of individual frequencies. (b) Founder frequencies. (c) Recombination rates.