Last modified: 08 April 2020

Manual of parallelized BLSOM for multithreaded environment.

Takashi Abe Graduate School of Science and Technology, Niigata University
E-mail : takaabe@ie.niigata-u.ac.jp

An unsupervised neural network algorithm, Kohonen's Self-organizing Map (SOM), is a powerful tool for clustering and visualizing high-dimensional complex data on a two-dimensional map (Kohonen, 1982 and 1990; Kohonen et al., 1996). On the basis of Batch-Learning SOM (BLSOM), we have developed a modification of the conventional SOM for genome sequence analyses, which makes the learning process and resulting map independent of the order of data input (Kanaya et al, 2001; Abe et al, 2003). We used the BLSOM for phylogenetic classification of metagenomic sequences obtained from mixed genomes of environmental microorganisms by analyzing tetranucleotide frequencies (Abe et al, 2005 and 2006) and protein function prediction of metagenomic sequences by analyzing oligopeptide frequencies (Abe et al, 2009).

Reference

1. Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, Nishi T, Mori H, Ikemura T (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map: characterization of horizontally transferred genes with emphasis on *E. coli* O157 genome. ***Gene*** 276:89-99.
2. Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. ***Genome Research*** 13:693-702.
3. Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T (2005) Novel Phylogenetic Studies of Genomic Sequence Fragments Derived from Uncultured Microbe Mixtures in Environmental and Clinical Samples. ***DNA research*** 12:281-290.
4. Abe T, Sugawara H, Kanaya S, Ikemura T (2006) Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale Self-Organizing Map constructed with the Earth Simulator. ***Journal of the earth simulator*** 6:17-23.
5. Abe T, Kanaya S, Uehara H, Ikemura T (2009) A novel bioinformatics strategy for function prediction of poorly-characterized protein genes obtained from metagenome analyses. ***DNA Research*** 16:287-298.
6. (Review) Iwasaki Y, Abe K, Wada K, Wada Y, and Ikemura T. A Novel Bioinformatics Strategy to Analyze Microbial Big Sequence Data for Efficient Knowledge Discovery: Batch-Learning Self-Organizing Map (BLSOM). ***Microorganisms*** 2013; 1:137-157.

System Requirements

1. Hardware:

a) 64-bit x86-64 CPUs with SSE instructions.

2. OS and Software:

a) 64-bit Linux system (kernel >=2.6).

b) Intel C/C++ compiler (>= 12.0) or GNU C/C++ (gcc) complier (>=4.4.7) (for using OpenMP library).

How to usage

1. Install BLSOM on Linux

1-1. Download the BLSOM.tar.gz

1-2. Unzip file. For example, on the Linux console, type:

```
cd <TheDirectoryYouPutTheTarball>
tar zxvf BLSOM.tar.gz
cd BLSOM
```

1-3. In "BLSOM" directory, there are 5 executable files, **run.csh, PCA, makeweight, SOM and PLOT**. The **param.dat** is parameter file for performing BLSOM.

The sample data are as followings

*input.dat : input data of degenerated tetranucleotide frequencies (merged Bacteria.frq, Virus.frq and Eukaryote.frq).

*Bacteria.frq : frequencies obtained from Bacteria genomes.

*Virus.frq : frequencies obtained from Virus genomes.

*Eukaryote.frq : frequencies obtained from Eukaryote genomes.

*weight.final : output file of weight vector data.

*CLSOM.txt : output sample of classification result using "input.dat".

CLSOM-1.txt : output sample of classification result using ".frq".

2. Quick Start

The commands required for simple execution are as follows.

```
./run.csh    <Your input file>    <Number of threads>
```

Sample command:

```
cd BLSOM
./run.csh    input.dat    8
```

All executable files and "param.dat" put under same directory.

For visualization of output results, please check "3.3. **Visualization of BLSOM result"**.

3. Introduction to detail execution

3.1 Preparation of oligonucleotide frequency for each fragment sequence

We prepare the program for calcultating oligonucleotide frequency. The program is "**freq_analysis_ver202004.rb**" under "**frq_program**" directory (Software requirement: ruby 1.8.7 or more).

Usage: ruby ./freq_analysis_ver202004.rb <Genome seq. FASTA file> <Pattern_file>

Sample command: ruby ./freq_analysis_ver202004.rb Ecoli_K12_MG1655.fna ./ptn/ Dinucleotide.ptn

After exacution, "Ecoli_K12_MG1655.fna.frq" was created as an output file.

If you want to change the calculation conditions, such as window size, normalization, etc, you can change the **parameters** in this program.

The format of "**Pattern file**" is as followings.

```
--
2    #Oligonucleotide number. Ex: Dinucleotide: 2, Trinucleotide: 3, etc.
16   #Pattern number. Ex: Dinucleotide: 16, Trinucleotide: 64, etc.
TT   #Oligonucleotide Pattern per line.
TC
TA
TG
CT
CC
CA
CG
AT
AC
AA
AG
GT
GC
GA
--
```

In the case of digenerated dinucleotide, we can use "+" to write "TC+GA". Please check "**./ptn/ DegeTetranucleotide.ptn**" for details.

We have prepared the following oligonucleotide patterns.

・Dinucleotide.ptn : Dinucleotide
・Tetranucleotide.ptn : Tetranucleotide
・DegeTetranucleotide.ptn : Degerated tetranucleotide

3.1.1 Format of input data

The data set must be constructed as a text file.
The format of Input data is as followings.

First line is written in sample name, for example, gene and protein name, sequence region etc. Second line is written in comment. Third line is written in vector values.
An example of input data set is shown in the file “input.dat”.

Format of input data.

1st line: sample name
2nd line: Comment
3rd line: Vector values (Each column in this line is separated by a space.)

***sample data of input data**

```
[1-5000]
1 [1-5000]
23 55 31 15 21
[5001-10000]
2 [5001-10000]
10 42 23 11 23
 ・ ・ ・ ・ ・ ・ ・
 ・ ・ ・ ・ ・ ・ ・
```

3.2. **Execution of BLSOM**

3.2.1 Parameter setting

You can set the prameter file (“**param.dat**”) according to input data.
DIMENSION, MaxNumberOfData, and MaxMapSize, can be changed depending on the memory size of your server.

The format of “**param.dat**” is as follows.

```
//DIMENSION (Integer. Input to dimension size of your vector data. Maximum number: 1300)
        136
//NumberOfDataInLine (Integer. Input to dimension size.)
        136
//NumberOfLines (Integer)
        1
//MaxNumberOfData (Integer. It changes according to volume of data.)
        4000000
//NumberOfIteration (Integer. We used “NeighborParam+20”)
        30
//MaxMapSize (Integer. Size of horizontal axis. It changes according to volume of data.)
        40
//NeighborParam (Integer. Neighborhood function. We used “MaxMapSize/4”)
        10
```

3.2.2 Calculation of BLSOM

In the linux console, type:

```
cd BLSOM
./run.csh <Your input file> <Number of threads>
```

*All executable file and “param.dat” put under same directory.

Sample command: ./run.csh input.dat 8

3.2.3 Output files

Six output files “PCAre”, “weight.start”, “weight.even”, “weight.odd”, “weight.final”, and “CLSOM.txt” are constructed by the present software, which has been designed to analyze data based on BLSOM.

Intoroduction to output files

PCAre : PCA results for creating initial weight vector file
weight.start : Initial weight vector file
weight.odd : Intermediate file during BLSOM execution
(odd number of itaration times)
weight.even : Intermediate file during BLSOM execution
(even number of itaration times)

3.2.3.1. Weight vector files

Four output files “weight.start”, “weight.even”, “weight.odd”, and “weight.final” are constructed as a weight vector files. “weight.start” contains initial vectors generated by PCA. “weight.final” contains the final weight vector generated by the learning process of BLSOM. “weight.odd” and “weight.even” contains generated by the middle learning process of BLSOM.

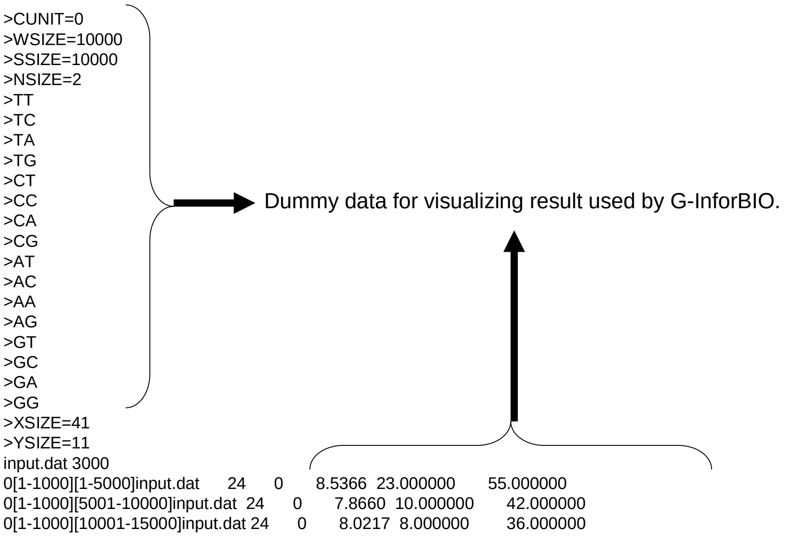
The format of these files is as followings. First line represents the number of lattice points in the horizontal (=40) and vertical (=11) axes. The second line represents the coordinates of the lattice points. Third line represents the mulitidimensional vector values.

*Format of weight vector

```
40 11
0 0
245.463042 88.117128 128.828534 108.208598
0 1
242.031187 81.040610 151.853100 107.007287
 • • • • • • • • • • • • • • • • •
 • • • • • • • • • • • • • • • • •
```

3.2.3.2 CLSOM file.

CLSOM file accumulates information on classification of objects in BLSOM (Fig.1). The “>XSIZE=40” and second line “YSIZE=11” represent the number of lattice points in the first and second axes. In the following lines, the first column corresponds to object name, the second and third columns correspond to the coordinate of the lattice point to which the object is classified. The dummy data for visualizing map used by G-InforBIO is shown in from forth to the last columns. The first letter in the object name is data type ID.



```
>CUNIT=0
>WSIZE=10000
>SSIZE=10000
>NSIZE=2
>TT
>TC
>TA
>TG
>CT
>CC
>CA
>CG
>AT
>AC
>AA
>AG
>GT
>GC
>GA
>GG
>XSIZE=41
>YSIZE=11
input.dat 3000
0[1-1000][1-5000]input.dat      24      0       8.5366  23.000000       55.000000
0[1-1000][5001-10000]input.dat  24      0       7.8660  10.000000       42.000000
0[1-1000][10001-15000]input.dat 24      0       8.0217  8.000000        36.000000
```

Dummy data for visualizing result used by G-InforBIO.

Fig.1 Sample data of “CLSOM.txt”

Attention: In the sample program, three data types (bacteria, virus, and eukaryote) are used. If you want to check according to three data types, you make the CLSOM file by making the frequency file according to the data type. The making CLSOM file is as followings.

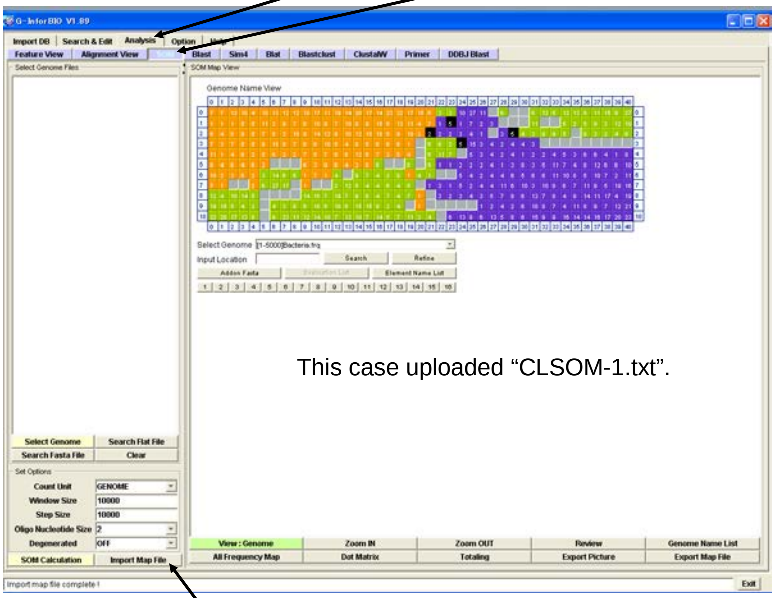
```
cat weight.final | ./PLOT Bacteria.frq Virus.frq Eukaryote.frq > CLSOM-1.txt
          or
cat weight.final | ./PLOT *.frq > CLSOM-1.txt
```

3.3. Visualization of BLSOM result

BLSOM result (CLSOM.txt or CLSOM-1.txt) was visualized by G-InforBIO.
G-InforBIO is freely available at
http://bioinfo.ie.niigata-u.ac.jp/GInforBIO/GInforBIO.zip

The process for uploading "CLSOM.txt" file is shown in Fig. 2.
Attention: In this case, you can use a part of visualization system in G-InforBIO. Please check the G-InforBIO's manual for the detail usage.

1. Select "Analysis" and "SOM"



Select Genome [1-5000]Bacteria.frq

Count Unit: GENOME
Window Size: 10000
Step Size: 10000
Oligo Nucleotide Size: 2
Degenerated: OFF

Import map file complete !

This case uploaded "CLSOM-1.txt".

2. Select "Import Map File" and upload "CLSOM.txt".

Fig. 2 The process for uploading CLSOM file using G-InforBIO.