FileFormat.doc

This file explains the file format for the input file for GenoMap.

An extended GTF file can be prepared by the following two steps.

First, a GTF file describing basic genome information (location of CDS and RNA genes) is produced from a GenBank or EMBL genome file by a tool such as the SISEQ (Sato 2000). The SISEQ command for this is:

‘genlist <infile> <outfile> t’.

The option ‘t’ directs production of a GTF table.

The quantitative data section is normally prepared by a software such as EXCEL and finally saved as a tab-delimited text file. The output of EXCEL contains DOS type end of line codes, which can be corrected by either apropriate tools of UNIX, a SISEQ command ‘txtr’, or the ‘txtr’ tool of GenoMap.

Then the two files are joined by the UNIX ‘cat’ command.

The GC content or GC skew data are produced by the SISEQ command:

‘tofast <infile> <outfile> b <window> <slide>’,

where option ‘b’ directs output of a table, <window> is the window size for the calculation of base composition, and <slide> is the size of sliding window, most commonly identical to the window size.

A short example is shown in Fig. 1. All items within the table should be delimited by a tab, not a space or something else. Within the description, multiple words should be delimited by a space.

New feature:

In a new version, color can be specified in each TAG entries. This is useful when the data with non-significant difference (judged by t-test or other tests) are displayed without highlight. The color is optional. If the color is not specified, the color settings of the GenoMap is used.

References

Sato, N. (2000) SISEQ: Manipulation of multiple sequence and large database files for common platforms, *Bioinformatics,* **16**, 180-181.

Sicheritz-Ponten, T. and Andersson, S. G. (1997) GRS: a graphic tool for genome retrieval and segment analysis, *Microb. Comp. Genomics*, **2**, 123-139.

Web site

http://www.molbiol.saitama-u.ac.jp/~naoki/GenoMap/

```
#GTF
Organism: Anabaena sp. BA000019
Type: type_unknown
Size: 6413771
Contigs: 0

definition of format:

name    type  orient  start   stop    length  description   color
all0002 CDS   R       1718    981     737     some gene 1
asl0003 CDS   R       2805    2617    188     some gene ABC
atpC    CDS   R       4365    3418    947     ATP synthase subunit
chlP    CDS   F       131490  132710  1220    geranylgeranyl hydrogenase
rpaA    CDS   R       133777  133034  743     response regulator

trnL    RNA   R       179948  179865  83      tRNA-Leu
trnL    RNA   F       185099  185180  81      tRNA-Leu
rrn5Sa  RNA   F       2382093 2382211 118     5S ribosomal RNA

1       TAG   1       1       2565    2565    1.288
2       TAG   1       1613    5067    3455    0.723
3       TAG   1       4475    7618    3144    0.9825

230     TAG   2       603266  606407  3142    1.013         gray
231     TAG   2       606213  609389  3177    0.923         gray
232     TAG   2       607164  610508  3345    1.199
```

Figure 1. A simplified example of an extended GTF file.