Computer Programs for the Detection of DNA Repeats Using Moving Window Spectral Analysis

Liping Du[1], Hongxia Zhou[1] and Hong Yan[1,2]
[1]Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong
[2]School of Electrical & Information Engineering, University of Sydney, NSW2006, Australia

Contract Email: h.yan@cityu.edu.hk, Website: http://www.hy8.com/~tec/papers/omwsa01.zip

Introduction

We have developed two computer programs for the visualization and detection of DNA repeats suing moving window spectral analysis:

`repeat_ft01:` the fast Fourier transform (FFT) is used for spectral analysis
`repeat_ar01:` the autoregressive (AR) model is used for spectral analysis

The programs are written in C and have been compiled for:

Red Hat Linux, version 4.2.13, Intel processor, using the GCC compiler
Solaris, version 5.7, Sun processor, using the GCC compiler
Windows XP, Intel processor, using the MS Visual C/C++ compiler

We believe that the program should also work on other versions of Linux, Solaris (SunOS) and Windows systems. They can be recompiled for other operating systems as well.

Installation

Unzip the file "omwsa01.zip". Use "linux.tar" under Linux, "solaris.tar" under Solaris, and "win32.zip" under Windows.

To extract the files:
On Linux, type "`tar xvf linux.tar`".
On Solaris, type "`tar xvf solaris.tar`".
On Windows, unzip "`win32.zip`".

If you use FTP to move files between a Windows system and a Unix system, make sure to use ASCII mode to transfer text files and binary mode for other files.

The Parameter File

The text file "`parameters1.txt`" contains three parameters. This file is used to reduce the number of input parameters on the command line for the programs. At the same time, it provides the flexibility for the user to change the parameters.

The first parameter is the maximum sequence length. This is used for allocating an array storing the sequence data. The current value, 10000, is enough to deal with two examples in

"ah005824_6800_7756.txt" and "x64775.txt". If you need to analyze a longer sequence, just increase the value of this parameter.

The second and third parameters in the file are the height and width of the output spectrogram image. If you set them smaller than 256, the programs will change them to 256. For the width, if you set it to 0 or a negative number, the programs will change it to the length of the sequence.

How to Use the Programs

You need to use the programs on a command-line window. Open a "Terminal" window on Linux or Solaris, and a "Command Prompt" window (MSDOS) on a Windows system.

Type "**repeat_ft01**", you will see

```
repeat_ft01  input_file  window_length  [seq_offset]  [entry]
```

Type "**repeat_ar01**", you will see

```
repeat_ar01  input_file  window_length  prediction_order  [seq_offset]  [entry]
```

These lines show you how to use the programs.

"input_file" is the name of the input sequence file in ASCII text format. They can be "ah005824_6800_7756.txt", "x64775.txt", "x_combined1.txt", or a file created by the user.

"window_length" is the moving window length.

"prediction_order" is the prediction order used in the AR model.

"seq_offset" is an optional parameter. It's the index of the first base pair in the input sequence. For example, it's 6800 for the sequence in "ah005824_6800_7756.txt", which contains a segment of the AH005824 sequence between 6800 and 7756 bp.

"entry" is an optional parameter. Note that both [seq_offset] and [entry] can be omitted in the command line. You can also omit [entry] alone. However, you cannot omit [seq_offset] alone. That is, if you specify [entry], you must specify [seq_offset] as well. The input file can contain several sequences, and each of them should have a one-line header starting with the sign ">". The parameter [entry] means the number of header lines before the sequence you're interested in.

The output is a BMP image, which has the file name consisting of the input file name, the window length, the prediction order and the entry number.

Examples of command lines for using the programs are shown below:

```
repeat_ft01 ah005824_6800_7756.txt 100 0 1
repeat_ar01 ah005824_6800_7756.txt 100 18 0 1

repeat_ft01 x64775.txt 100 0 1
```

```
repeat_ar01 x64775.txt 100 18 0 1

repeat_ft01 x_combined1.txt 100 0 1
repeat_ar01 x_combined1.txt 100 18 0 1

repeat_ft01 x_combined1.txt 100 0 2
repeat_ar01 x_combined1.txt 100 18 0 2
```

The file "`xxt01.bat`" ("`xxt01`" on Unix) is a batch file containing all above command lines. You can simply type "`xxt01`" to run the programs according to various options above.

If you want to create your own batch file, for example, "`xxt02`", similar to "`xxt01`", make sure it's executable under Unix. You can do so by typing "`chmod u+x xxt02`". On Windows, you can just name it "`xxt02.bat`" to make it executable.