User's Guide

MACML: Model Averaging Clustering by Maximum Likelihood

Version 1.1 (June 2009)

Content

MACML: Model Averaging Clustering by Maximum Likelihood 1
Content I
1 Introduction 1
2 Installation 1
2.1 Compiled Executables 1
2.2 Linux/Unix/Mac/Windows 1
3 Format of Input File 1
4 Parameters Setting 2
5 Output Format 2
6 Acknowledgements 3
7 Contact 3

1 Introduction

MACML is a program that clusters sequences into heterogeneous regions with specific site types, without requiring any prior knowledge, such as cluster count or length, etc. It features maximum likelihood estimation, model selection and model averaging and adopts a divide-and-conquer approach to hierarchically detect clusters within sequences.

2 Installation

For high efficiency and compatibility with more platforms, the kernel codes of MACML are written in standard C++. You can download from the MACML webpage at http://www.yale.edu/townsend/software.html the newest package, normally named MACMLXXX.tar.gz (XXX stands for the version).

2.1 Compiled Executables

There have compiled executables on Linux/Unix/Mac/Windows. Please unpack the package of MACMLXXX.tar.gz (see below) and then you will find compiled executables in the folder of “MACMLXXX/bin/”.

2.2 Linux/Unix/Mac/Windows

For compilation on your specific platform, please follow the steps below.

- Unpack the package of MACMLXXX.tar.gz by the following commands.

```
tar –zxf MACMLXXX.tar.gz
```

- If you use other Linux/Unix OS, you have to compile the program in the source codes folder with the help of g++/gcc compiler.

```
cd MACMLXXX/src
make
```

3 Format of Input File

MACML accepts the fasta file and two kinds of sequences.

- Aligned sequences, with equal length. For example, four aligned sequences are shown as follows.

```
>dmel-FBpp0100045
MSLTNKNVIFVAGLGGIGLDTSKELLKRDLKNLVILDRIENPAAI
```

```
>dsec-FBpp0197133
MALTNKNVIFVAGLGGIGLDTSKELLKRDLKNLVILDRIENPAAI
>dsim-FBpp0222370
MALTNKNVIFVAGLGGIGLDTSKELLKRDLKNLVILDRIENPAAI
>dyak-FBpp0264047
MALTNKNVVFVAGLGGIGLDTSKELVKRDLKNLVILDRIENPAAI
```

- Vectors of 0's and 1's. Each vector is independent from each other, without the requirement of equal length of all sequences, so that MACML can work as a batch. For example,

```
>gene1
0000000000000000000011111111111110000000000000000000000
01000000000000000000010000000
>gene2
0000111111110001
>gene3
1111111111111000000010000100000111111111001
```

You can also see two example files, named "sequences.fasta" and "vectors.fasta" in the folder of "MACMLXXX/example/".

4 Parameters Setting

MACML allows user to choose different criteria for significance test. The following is the parameters' setting.

- -i Input sequence file name [string, required]
- -c Criterion for clustering [integer, optional], {0:BIC || 1:AIC || 2:AICc}, default = 0
- -t Type of input sequence [integer, optional], {0: aligned sequences || 1: a vector of 0's and 1's], default = 0
- -m Model selection and model averaging [integer, optional], {0: use both model selection and model averaging || 1: use only model selection], default = 0

For more details, please see the file, named "commands.txt" in the folder of "MACMLXXX/example/".

5 Output Format

MACML provides comprehensive information estimated from input sequences. The following is the detailed output information.

- cs: Start position of cluster

- ce: End position of cluster
- p0: Variant rates outside of cluster
- pc: Variant rates within cluster
- InL0: Log-likelihood value under null model
- InL: Maximum log-likelihood value under clustering model
- AIC0: AIC under null model
- AIC: AIC under clustering model
- AICc0: AICc under null model
- AICc: AICc under clustering model
- BIC0: BIC under null model
- BIC: BIC under clustering model

MACML profiles sequence heterogeneity by model averaging, so that it also provides variant rate for each site. More details can be found at the folder of ““MACMLXXX/example/”.

6 Acknowledgements

We thank three reviewers for their constructive comments on this work. We also thank Zheng Wang, Francesc López, Aleksandra Adomas, Gina Wilpiszeski and Andrea Hodgins-Davis for valuable discussions.

7 Contact

Please send bugs or advice to Zhang Zhang at zhang.zhang@yale.edu.