Documentation for **Hetero2**

A program to simulate the evolution of DNA under

the mixture models of heterogeneity across lineage

and heterogeneity across sites

May, 2017

Disclaimer

CSIRO Open Source Software License Agreement (variation of the BSD / MIT License)

All rights reserved. CSIRO is willing to grant you a license to this Hetero Version 2 on the following terms, except where otherwise indicated for third party material.

Redistribution and use of this software in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

- Neither the name of CSIRO nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission of CSIRO.

EXCEPT AS EXPRESSLY STATED IN THIS AGREEMENT AND TO THE FULL EXTENT PERMITTED BY APPLICABLE LAW, THE SOFTWARE IS PROVIDED "AS-IS". CSIRO MAKES NO REPRESENTATIONS, WARRANTIES OR CONDITIONS OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY REPRESENTATIONS, WARRANTIES OR CONDITIONS REGARDING THE CONTENTS OR ACCURACY OF THE SOFTWARE, OR OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT, THE ABSENCE OF LATENT OR OTHER DEFECTS, OR THE PRESENCE OR ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE.

TO THE FULL EXTENT PERMITTED BY APPLICABLE LAW, IN NO EVENT SHALL CSIRO BE LIABLE ON ANY LEGAL THEORY (INCLUDING, WITHOUT LIMITATION, IN AN ACTION FOR BREACH OF CONTRACT, NEGLIGENCE OR OTHERWISE) FOR ANY CLAIM, LOSS, DAMAGES OR OTHER LIABILITY HOWSOEVER INCURRED. WITHOUT LIMITING THE SCOPE OF THE PREVIOUS SENTENCE THE EXCLUSION OF LIABILITY SHALL INCLUDE: LOSS OF PRODUCTION OR OPERATION TIME, LOSS, DAMAGE OR CORRUPTION OF DATA OR RECORDS; OR LOSS OF ANTICIPATED SAVINGS, OPPORTUNITY, REVENUE, PROFIT OR GOODWILL, OR OTHER ECONOMIC LOSS; OR ANY SPECIAL, INCIDENTAL, INDIRECT, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES, ARISING OUT OF OR IN CONNECTION WITH THIS AGREEMENT, ACCESS OF THE SOFTWARE OR ANY OTHER DEALINGS WITH THE SOFTWARE, EVEN IF CSIRO HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH CLAIM, LOSS, DAMAGES OR OTHER LIABILITY.

APPLICABLE LEGISLATION SUCH AS THE AUSTRALIAN CONSUMER LAW MAY APPLY REPRESENTATIONS, WARRANTIES, OR CONDITIONS, OR IMPOSES OBLIGATIONS OR LIABILITY ON CSIRO THAT CANNOT BE EXCLUDED, RESTRICTED OR MODIFIED TO THE FULL EXTENT SET OUT IN THE EXPRESS TERMS OF THIS CLAUSE ABOVE "CONSUMER GUARANTEES". TO THE EXTENT THAT SUCH CONSUMER GUARANTEES CONTINUE TO APPLY, THEN TO THE FULL EXTENT PERMITTED BY THE APPLICABLE LEGISLATION, THE LIABILITY OF CSIRO UNDER THE RELEVANT CONSUMER GUARANTEE IS LIMITED (WHERE PERMITTED AT CSIRO'S OPTION) TO ONE OF FOLLOWING REMEDIES OR SUBSTANTIALLY EQUIVALENT REMEDIES:

(a) THE REPLACEMENT OF THE SOFTWARE, THE SUPPLY OF EQUIVALENT SOFTWARE, OR SUPPLYING RELEVANT SERVICES AGAIN;

Credits

This software was developed by the bioinformatics and phylogenomics team in Ecosystem Sciences, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, Australia.

Introduction

We present a software to simulate the evolution of nucleotide sequences under mixture models of heterogeneity across lineages and heterogeneity across sites. Hetero2 allows users to assign lineage-specific (and also site-specific) differences in the rate matrices used to describe the evolutionary process.

Installation of the software

The software was written in C++, and it has been tested under linux and MacOS platform. You need to have C++ compiler installed in the machine in order to compile the source codes. The compilation steps are shown as follows:

```
$ tar -zxvf Hetero-2.2.tar.gz

$ cd Hetero-2.2

$ make
```

Then the executable file named *Hetero2* will appear.

Usage of Hetero2

```
Syntax:
  ./Hetero2 <tree file> <site info file> <param file list> <other options>
  ./Hetero2 -h

  <tree file>           : "Tree file" lists the tree of each site category, the
                          edge lengths and the labels of terminal/internal nodes

  <site info file>      : "Site info file" lists the detailed information of
                          each site category, including the site proportion and
                          the nucleotide distribution at the root

  <param file list>     : "Param file list" shows the name of the parameter file
                          of each variant site category
```

```
Other options:
```

```
  -l <sequence length>  : The length of sequences to be simulated
                          (default: 50,000)

  -f <output format>    : The format of simulated multiple sequence alignment
                          1 - FASTA format
                          2 - Sequential PHYLIP format (default)

  -o <output prefix>    : Prefix for output files
                          (default: <tree file> w/o .ext)

  -h                    : The help page
```

The output file will be:

<output prefix>.out, which stores the simulated multiple sequence alignment file. By default, it is in sequential PHYLIP format, and the user can select FASTA as the output format.

Example files

The following example files are available for reference:

trees.txt	An example of "tree file" which lists the tree of each site category, the edge lengths and the labels of terminal/internal nodes
site_info_file.txt	An example of "Site info file" lists the detailed information of each site category, including the site proportion and the nucleotide distribution at the root
param_file_list.txt	An example of "parameter list file" showing the name of the parameter file of each variant site category
parameter_1.txt parameter_2.txt	An example of parameter file showing the detailed parameters of the rate matrix of each edge leading to the corresponding node

1. Example tree file – tree.txt

The "Tree file" should list the tree of each site category, the edge lengths and the labels of the terminal and the internal nodes. Also, the edge lengths represent the average number of substitutions per site.

Format:

[name of variant site category] [newick tree format with internal node labels]

```
Category_1     (A:0.30941,(B:0.33809,(C:0.29115,(D:0.04607,(E:0.11096,(F:0.06955,(G:0.03861,
H:0.04429)1:0.01701)2:0.02719)3:0.04063)4:0.27458)5:0.02226)6:0.20349);

Category_2     (A:0.47928,(B:0.06142,(C:0.06739,(D:0.00915,(E:0.00227,(F:0.00033,(G:0.00114,
H:0.00059)1:0.00076)2:0.00018)3:0.00069)4:0.05522)5:0.06595)6:0.09457);
```

Note:

1. "A,B,C,D,E,F,G,H" are terminal nodes and "1,2,3,4,5,6" are internal nodes.
2. All the trees have to be in the same topology, and with the same set of terminal/internal nodes.

2. Example site info file – site_info_file.txt

The "site info file" should show the detailed information of each site category, including the site proportion and the nucleotide distribution at the root.

Format:

[name of site category] [variant/invariant] [proportion] [freq(A)] [freq(C)] [freq(G)] [freq(T)]

```
Constant_site invariant    0.48100 0.30606    0.16447    0.14259    0.38688
Category_1    variant      0.13214 0.51059    0.18069    0.12901    0.17971
Category_2    variant      0.38686 0.32819    0.26728    0.10713    0.29740
```

Note:

1. The names of the variant categories are same as the names of variant site categories in the tree file.
2. There is at most one invariant category.

3. Example parameter list file – param_file_list.txt

The "parameter list file" shows the name of the corresponding parameter file for each variant site category.

Format:

[name of variant site category] [parameter file name]

```
Category_1    parameter_1.txt
Category_2    parameter_2.txt
```

Note: The names of the variant categories are same as the names of variant site categories in the tree file.

4. Example parameter file – parameter_1.txt

The "parameter file" shows the detailed parameters of the rate matrix of each edge leading to the corresponding node.

Format:

Node S1 S2 S3 S4 S5 S6 Pi_1 Pi_2 Pi_3 Pi_4

```
A       1.91928  2.13915  1.59805  2.58998  1.50701  1.00000  0.75553  0.12277  0.04531  0.07639
B       1.54547  3.27634  1.64800  4.70045  3.30819  1.00000  0.57448  0.15236  0.16912  0.10404
C       2.67365  4.14835  1.73040  4.36875  2.94216  1.00000  0.63957  0.16518  0.10018  0.09507
D       1.54547  3.27634  1.64800  4.70045  3.30819  1.00000  0.57448  0.15236  0.16912  0.10404
E       2.95455  13.8448  2.56968  7.04199  16.4082  1.00000  0.49983  0.14356  0.17577  0.18084
F       2.95455  13.8448  2.56968  7.04199  16.4082  1.00000  0.49983  0.14356  0.17577  0.18084
G       2.95455  13.8448  2.56968  7.04199  16.4082  1.00000  0.49983  0.14356  0.17577  0.18084
H       2.95455  13.8448  2.56968  7.04199  16.4082  1.00000  0.49983  0.14356  0.17577  0.18084
1       2.95455  13.8448  2.56968  7.04199  16.4082  1.00000  0.49983  0.14356  0.17577  0.18084
2       2.95455  13.8448  2.56968  7.04199  16.4082  1.00000  0.49983  0.14356  0.17577  0.18084
3       2.95455  13.8448  2.56968  7.04199  16.4082  1.00000  0.49983  0.14356  0.17577  0.18084
4       2.67365  4.14835  1.73040  4.36875  2.94216  1.00000  0.63957  0.16518  0.10018  0.09507
5       2.67365  4.14835  1.73040  4.36875  2.94216  1.00000  0.63957  0.16518  0.10018  0.09507
6       1.54547  3.27634  1.64800  4.70045  3.30819  1.00000  0.57448  0.15236  0.16912  0.10404
```

Note:

1. Pi_1, Pi_2, Pi_3, Pi_4 are the equilibrium distribution of A, C, G, T respectively. Their sum should equal to 1.
2. The labels of the terminal and the internal nodes have to be same as those in the tree file.
3. The value of S6 does not need to be 1.0, although it is in this example.

To run *Hetero2* for these example files

```
$./Hetero2 trees.txt site_info_file.txt param_file_list.txt
```

The simulated multiple sequence alignment would be in the file: "trees.out".

Contact person

Dr Lars Jermiin

Email: lars.jermiin@anu.edu.au

Dr Thomas Wong

Email: thomas.wong@anu.edu.au

Research School of Biology, Australian National University, Canberra, ACT 0200, Australia