

# **USER GUIDE – BH+I PROGRAM**

**Version 1.0**

**Created By Vivek Jayaswal**

## Table of Contents

<b>Java Programs</b>	
Installing Java	3
BH+I Programs	3
• Compiling Programs	4
• Input	4
• Output	5
Q-matrices Initialization File	6
<b>Sample Files</b>	7
<b>References</b>	8

# 1. JAVA PROGRAMS

## 1.1 Installing Java

Java™ can be downloaded from the website <http://www.java.sun.com>. The programs were developed using JDK 1.4 and can be compiled using either JDK1.4 or JDK1.5. There are several IDEs (Integrated Development Environment) available for developing/ modifying Java programs e.g. JCreator, NetBeans and Eclipse. An advanced user may use any of these IDE's for viewing and/ or modifying the existing BH programs.

## 1.2 BH+I Programs

This sub-section provides some information about the various methods implemented in Java and can be skipped by a user not interested in the actual implementation of the program.

Table 1 provides a brief description of the programs required for computing Maximum-Likelihood using BH+ I model (Barry and Hartigan's model with invariant sites).

**TABLE 1 List of programs for computing maximum-likelihood**

<b>Program Name</b>	<b>Brief Description</b>
BHI.java	This is the main class which calls all the remaining classes for computation of maximum-likelihood
BranchDetails.java	This class contains methods for storing and retrieving a 4*4 matrix of joint probability distribution values along each edge of the phylogenetic tree
InternalNodeRiXi.java	This class contains methods for storing and retrieving RiXi values for each internal node for all 4 nucleotide types
InternalNodeSiXi.java	This class contains methods for storing and retrieving SiXi values for each internal node for all 4 nucleotide types
MixtureLikelihood.java	This class contains functions for reading a set of DNA sequences, initializing the joint probability distribution values along each branch, computing log likelihood based on Q-values and updating Q-values
NodeParameters.java	This class contains methods for storing and retrieving the invariant sites parameters
NodeStructure.java	This class contains methods for storing and retrieving parent node value, child node value and node type (Leaf/ Internal)
NewickTreeTraversal.java	This class contains the methods for constructing a phylogenetic tree from a Newick tree representation
QmatrixGenerator.java	This class is used to generate random Q-matrices. The matrix generated could be either symmetric or non-symmetric

### 1.2.1 Compiling Programs

1. Download and unzip the file BHI.zip
2. Go to the command prompt. Windows users can click on Start -> Run and then type cmd to go to the command prompt
3. Type **set classpath=%classpath%;C:/BHI/Java** at the command prompt. For Unix users the command would be **export CLASSPATH=\$CLASSPATH:/BHI/Java**
4. Change the present working directory to BHI/Java e.g. Windows users can type **cd C:/BHI/Java**
5. Type **javac BHI.java** to compile the java programs
6. Type **java BHI** to execute the program and obtain **log likelihood** value

### 1.2.2 Input

Table 2 describes the input parameters required by the BHI program.

**TABLE 2** Input parameters for the program BHI

Parameter Name	Brief Description	Sample Value
1. No of data sets	Number of data sets in the PHYLIP file. This should be an integer value between 1 and N, where N denotes the total number parametric bootstrap data sets	1
2. No of species	Number of leaf nodes in the phylogenetic tree	5
3. No of sites	Total number of sites	1238
4. Data sets file	Complete path of the PHYLIP format file	C:/javapgms/source/BacData.txt
5. No of trees	Number of Newick format trees	15
6. Tree Topologies File	Complete path of the tree topologies. This file should contain atleast as many as trees as those specified above	C:/javapgms/source/BacTreeTop.txt
7. Options for Q-matrices: <b>1</b> -> Default, <b>2</b> -> User File, <b>3</b> -> Random	<p><b>1:</b> Uses a Q-matrix with the diagonal elements being 1/8 and off-diagonal elements being 1/24 for each edge</p> <p><b>2:</b> Q-values initialization file: Name and complete path of the user-defined Q-matrices file. A blank line should separate the Q-matrices corresponding to different edges. Each line of the Q-matrix should have 4 values separated by “/” character (generated by pressing the “TAB” key). The Q-matrices are read from the source file in a pre-determined order which can be identified by running the program NewickTreeTraversal (refer section 1.2.4)</p> <p><b>3:</b> Random: Randomly generate Q-matrices for each edge. The Q-matrices could be either symmetric or non-symmetric. The Q-values generated are stored in a file “<b>InitMatrices.txt</b>” in the same folder as the destination file</p>	<b>1</b>

8. Output file (name and path)	Name and path of the output file that stores log likelihood value at convergence and Q-matrices along each edge	C:/javapgms/Output/out.txt
9. Specify Iteration cut-off value/ Number of iterations (1/2)	<p>1: Specify a cut-off value (e.g. <b>0.000000001</b>) for convergence such that <math>\sum_{i=1}^4 \sum_{j=1}^4 [Q^{\text{new}}(i,j) - Q^{\text{old}}(i,j)]^2 \leq \text{cut-off value}</math></p> <p>where,  <math>Q^{\text{new}}</math> = new joint probability distribution matrix  <math>Q^{\text{old}}</math> = old joint probability distribution matrix  <math>Q(i,j)</math> = (i,j)<sup>th</sup> element of the Q-matrix</p>	
10. Beta cut-off value	<p>2: Number of iteration: Specify the number of iterations e.g. <b>10</b></p> <p>The minimum difference between the new and old values of beta (the proportion of invariant sites) that is required for convergence</p>	0.00001

### 1.2.3 Output

An output file, stats.txt, is generated in the destination folder (folder specified in parameter 8 of Table 2). This file contains the log likelihood under the BH+I model and the invariant sites parameters for each tree specified in the Tree Topologies file.

If the source file contains more than one data set, then the output is generated for all the tree topologies per data set in a sequential order i.e. all tree topologies are evaluated under the first data sets, followed by an evaluation for the second data set and so forth.

Figure 1 shows the output for a single data set along with an explanation of the values in each column

Log Likelihood (BH+I)	Beta	P(A inv)	P(C inv)	P(G inv)	P(T inv)
-4193.71	0.55443	0.285948	0.222462	0.340415	0.151175
-4202.87	0.553158	0.289716	0.215763	0.338465	0.156056
-4206.92	0.561917	0.280953	0.228224	0.345142	0.145681
-4207.51	0.563212	0.282113	0.228273	0.343536	0.146078
-4223.91	0.566681	0.280442	0.221942	0.339067	0.158549
-4224.55	0.566711	0.280458	0.22199	0.338196	0.159356
-4217.42	0.570128	0.283088	0.220239	0.34084	0.155833
-4230.79	0.576663	0.279285	0.226404	0.343073	0.151238
-4234.73	0.578266	0.275055	0.222452	0.339938	0.162555
-4232.12	0.577766	0.279354	0.228726	0.343104	0.148816
-4236.86	0.579429	0.278223	0.22761	0.343527	0.15064
-4239.81	0.580109	0.279744	0.228495	0.340662	0.151099
-4239.39	0.581992	0.278114	0.22833	0.342283	0.151273
-4240.61	0.585026	0.275947	0.229148	0.340093	0.154811
-4240.66	0.58527	0.276036	0.229045	0.340077	0.154842

**Figure 1.** Log likelihood values for 15 un-rooted tree topologies of a 5-leaf node data set

The output file “out.txt” specified in parameter 8 of Table 2 is useful when a single data set and a single tree topology is considered. It contains the log likelihood values and the Q-matrices per iteration. The Q-matrices obtained at convergence can be analyzed for the marginal probabilities at various nodes (internal nodes and leaf nodes) as well as for identifying substitution biases.

This file is re-written each time a tree topology is evaluated. Therefore, if more than one data set or tree topology is specified, the values in this file correspond to the last data set and the last tree topology.

### 1.2.4 Creating a Q-values Initialization File

1. Download and unzip the file DisplayTree.zip
2. Go to the command prompt. Windows users can click on Start -> Run and then type cmd to go to the command prompt
3. Type **set classpath=C:/DisplayTree** at the command prompt. For Unix users the command would be **export CLASSPATH=/DisplayTree**
4. Change the present working directory to DisplayTree e.g. Windows users can type **cd C:/DisplayTree**
5. Type **javac NewickTreeTraversal.java** to compile the java programs
6. Type **java NewickTreeTraversal** to execute the program

### Input

The input parameters required by the NewickTreeTraversal program are specified in table 3.

**TABLE 3** Input parameters for the program NewickTreeTraversal

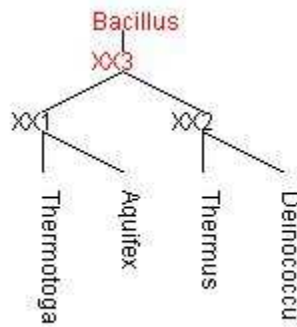
Parameter Name	Brief Description	Sample Value
1. Newick Tree Representation	Newick format tree	(Bacillus((Thermotoga,Aquifex)(Thermus,Deinococcu)))
2. JPG file path	Name and path of JPG file	C:/showtree/Output/tree.jpg

### Output

The console output shows the name and order in which the Q-matrices will be read. The Q-matrices corresponding to the internal edges are read first followed by the leaf node edges. Figure 2 shows a sample output and Figure 3 the corresponding tree structure.

```
C:\Documents and Settings\vivek\My Documents\ShowTree>java NewickTreeTraversal
----- INPUT PARAMETERS -----
Newick Tree Representation: <Bacillus<<Thermotoga,Aquifex><Thermus,Deinococcu>>>
JPG file path: C:\Documents and Settings\vivek\My Documents\ShowTree\Trial.jpg
Leaf Nodes: [Aquifex-Node, Deinococcu-Node, Thermotoga-Node, Thermus-Node]
Internal Nodes: [1-Node, 2-Node, 3-Node]
```

**Figure 2.** Console Output for the program NewickTreeTraversal



**Figure 3.** Phylogenetic tree corresponding to the Newick format tree  
(Bacillus((Thermotoga,Aquifex)(Thermus,Deinococcus)))

The leaf node indicated in Figure 3 in red (Bacillus) is the start node for maximum-likelihood computation. The rows and columns of a Q-matrix are defined based on the distance of the nodes (corresponding to an edge) from the start node. For example, edge 1-Aquifex has node-1 closer to the start node, so the Q-matrix will have rows corresponding to the marginal probabilities at node-1 and the columns corresponding to the marginal probabilities at the node Aquifex.

The Q-matrices for Newick tree specified in Figure 2 and 3 would be read from the user-specified file in the following order:

- a. Internal nodes
  - a. 1: This corresponds to the edge 1-3 since 3 is closer to Bacillus (start node)
  - b. 2: This corresponds to the edge 2-3.
  - c. 3: This corresponds to the edge Bacillus-3
- b. Leaf nodes
  - d. Aquifex: This corresponds to the edge 1-Aquifex
  - e. Deinococcus: This corresponds to the edge 2-Deinococcus
  - f. Thermotoga: This corresponds to the edge 1-Thermotoga
  - g. Thermus: This corresponds to the edge 2-Thermus

The Q-matrices selected should maintain the consistency at internal nodes. For example, the Q-matrices along the edges Bacillus-3, 3-1 and 3-2 should be such that the marginal probability for node-3 is the same in each case.

## 2. Sample Files

The file examples.zip contains sample input and output files to help the user understand the file formats, e.g.

1. The input file should be in PHYLIP format. Also, the species name and the corresponding sequence should be on the same line
2. The tree topologies file should be in Newick format with one tree per line. The trees should not have a semi-colon at the end of a line.

### 3. References

Ababneh, F., L. S. Jermin, C. Ma, and J. Robinson. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinf.* (in press).

Barry D, Hartigan JA. 1987a. Statistical analysis of hominoid molecular evolution. *Stat Sci*, 2:191-210.

Jayaswal, V., L. S. Jermin, and J. Robinson. 2005. Estimation of phylogeny using a general Markov model. *Evol. Bioinf. Online*. 1:62-80.