Manual for DHMM

DHMM is written in two parts. The first script, DHMM_prepare.pl, is implemented in PERL and used to estimate the potentialities of mutations for sites in the core genome. The second script, DHMM.R, is implemented in R and used to estimate the maximum likelihood model for given numbers of states.

DHMM_prepare.pl can be run from the command line as:

```
perl DHMM_prepare.pl <core genome alignment> <gene annotation> <variations to be considered> > <prepared file>
```

NOTE:

<core genome alignment>: a FASTA format file that contains the core genome alignments of multiple sequences. No gap is allowed in the first sequence, which is treated as the reference genome.

<gene annotation>: An annotation file for the reference (first sequence in the core genome alignmenet) to identify coding regions in the core genome alignment.

<variations to be considered >: All the SNPs/indels that are going to include in the DHMM analysis. These SNPs/indels will affect the potentialities of mutations for each site as well.

DHMM_prepare.pl will generate the potentialities of sites in the core genome to **standard output**, which need to be re-directed to a file. It will also generate a summary to the **standard error**.

DHMM.R can be run from the command line as:

```
Rscript DHMM.R <site file> <output file> <no. state / model file> <no. of random starts (0 for sites only) > <no. of iterations>
```

It can also run by copying the script into the interactive command interface of R to run it in a window, after you replaced the following variables in the script with desired values.

Interactive interface variables	Command line parameters	Default value	Description
file.site	site file	DHMM.site	The file prepared by DHMM_prepare.pl
file.output	output file	DHMM.output	The output file
states	no. state / model file	3	Either the no. of states (k) in the HMM model, or a file with pre-calculated model to infer per-site likelihood
randomStart	no. of random starts (0 for sites only)	100	No of random initiations, if this variable is set to be 0, only the per-site likelihood will be calculated
maxIteration	no. of iterations	100	Stop the E-M algorithm after *maxIteration* times even if the E-M algorithm does not converge

The format of output file:

```
Transition:            # transition matrix for each state
0.0981850357217655 0.329429484694875 0.572385479583359
0.00139907792329797 0.9954573445152 0.00314357756150191
1.15291070181028e-05 5.68062925570113e-05 0.999931664600425
Emission:              # emission matrix for each state. Four columns are dN, dS, dNC, dSTOP, repectively
1.36944576564138 3.09804930430475e-40 0.482051219632032 1.07896687635167
0.00572600549319933 0.00242603718686788 0.00544268151102585 0.00127238812059269
0.000939327251861482 0.00116678299452793 0.00136833690372673 6.95794168021289e-05
Initial:               # the assignments of the first site in  the core genome
4.67205042371185e-175 0.999999999679858 4.98886133415103e-187
LL, Diff, AIC, BIC:    # the natural logarithms of likelihood, difference between successive iterations, AIC and BIC values
-40520.0782376111 9.9481112556532e-06 81076.1564752221 81314.1106657795
0 0 0 0 0 0 0 0 0 0            # The following are per-site likelihood of the core genome. Each site is shown in one row.
0 0 0 0 0 0 0 0 0 0            # Non-core sites are shown as 0s.
0 0 0 0 0 0 0 0 0 0            # The first three columns are: potentialities for NS, S and NC changes.
0 0 0 0 0 0 0 0 0 0            # Columns 4-7 are numbers of NS, S, NC mutations and indels in each site
0 0 0 0 0 0 0 0 0 0            # Columns start from 8 are per-site likelihood of assignments to states 1, 2, 3 …
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 3.77299234161563e-176 1.00000000473665 3.80131556790969e-188
0 0 1 0 0 0 0 0.000113327669929423 0.999648575545766 0.000238101521810552
0 0 1 0 0 0 0 0.00011551816222393 0.999398220497543 0.000486266072061992
0 0 1 0 0 0 0 0.000115408449926724 0.999152679695925 0.000731916589730233
0 0 1 0 0 0 0 0.000115252683869612 0.998909934480721 0.00097481756670253
0 0 1 0 0 0 0 0.000115097536764847 0.998669920636881 0.00121498656093297
0 0 1 0 0 0 0 0.000114943960990596 0.998432614698869 0.00145244607284233
0 0 1 0 0 0 0 0.000114791962116525 0.998197994318653 0.00168721845453997
0 0 1 0 0 0 0 0.000114641526247776 0.997966037399771 0.0019193258072959
0 0 1 0 0 0 0 0.000114492639233424 0.997736722118295 0.00214878998161834
0 0 1 0 0 0 0 0.00011434528705505 0.99751002687846 0.0023756325791216
0 0 1 0 0 0 0 0.000114199455841708 0.997285930333668 0.00259987495463304
…
```

Algorithm for DHMM

Models for the core genome and emission function

Let A be the non-recombinant, non-repetitive core genome of Paratyphi A with a total length of L. We assume C be all potentialities of random mutations/indels in A. For $i \in [1, \dots, L]$, Either synonymous ($c_{i,S}$) or non-synonymous ($c_{i,NS}$) mutations can happen if a_i is in a CDS, or else be a mutation in a non-coding region ($c_{i,NC}$). On top of mutations, we assumed that every site potentially corresponds to an indel rate of $c_{i,indel}$ independently of mutation rates. Thus $c_{i,*}$ represents all potential mutations/indels in a_i and is associated with four attributes $c_{i,*} = c_{i,NC} + c_{i,S} + c_{i,NS} + c_{i,indel}$. All possible changes in core genome is presented as $C = \begin{pmatrix} c_{1,NC} & \dots, & c_{L,NC} \\ c_{1,S} & \dots, & c_{L,S} \\ c_{1,NS} & \dots, & c_{L,NS} \\ c_{1,indel} & \dots, & c_{L,indel} \end{pmatrix}$, in which $c_{i,NC}$, $c_{i,S}$ and $c_{i,NS}$ were calculated by the approximate method with a GTR substitution model that is summarized from all SNPs. Finally, we reached a total number of 2.44MB non-synonymous sites, 1.15 MB synonymous sites, 0.48 MB non-coding sites and 4.07MB indel sites.

For $V \in \{NC, S, NS, indel\}$, the occurrences of SNPs/indels in the non-recombinant, non-repetitive core genome can be presented as a matrix $O = \begin{pmatrix} o_{1,NC} & \dots, & o_{L,NC} \\ o_{1,S} & \dots, & o_{L,S} \\ o_{1,NS} & \dots, & o_{L,NS} \\ o_{1,indel} & \dots, & o_{L,indel} \end{pmatrix}$, in which $o_{i,V}$ equals number of SNPs/indels in the ith site, which is >1 for a homoplastic site and =0 for an invariant site. We assumed that each a_i was sampled a hidden state h_i, which is one of the K number of distinct evolutionary scenarios $s_1, \dots, s_K$. We also assumed that the occurrence of SNPs/indels in the kth scenario s_k is Poisson distributed with a mean value of $p_{k,V}$. The probabilities of observation of $o_{i,V}$ evolving under scenario s_k is

$$Pr(o_{i,V} |\, h_i = s_k) = \frac{e^{-(p_{k,V}/c_{i,V})} \times (p_{k,V}/c_{i,V})^{o_{i,V}}}{(o_{i,V})!}$$

And the summation of probability in the ith site is:

$$Pr(o_{i,*} |\, h_i = s_k) = \prod_{V\in\{NC,S,NS,indel\}} Pr(o_{i,V} |\, h_i = s_k).$$

Transition matrix and initial state probabilities

The transition matrix between different evolutionary scenarios is $T = \begin{pmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,K} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,K} \\ \cdots & \cdots & \cdots & \cdots \\ t_{K,1} & t_{K,2} & \cdots & t_{K,K} \end{pmatrix}$, in which $t_{k,l}$ indicates the transition rate from s_k to s_l. The initial state probabilities for the first base in the core genome is $U = (u_{s_1}, \ldots, u_{s_K})$, in which $\sum u = 1$.

E-M algorithm

The probability of the model $Pr(D) \equiv Pr(H|s_1, \ldots, s_K; T, P)$ can be calculated directly using the forward algorithm (*17*) and the local maximum values of all the parameters can be estimated using the E-M algorithm (*18*). Let $\alpha_{i,s_k} = Pr(h_1, \ldots, h_i|h_i = s_k)$ and $\beta_{i,s_k} = Pr(h_{i+1}, \ldots, h_L|h_i = s_k)$. Then,

$$p_V = \frac{\sum_{i=1}^{L} \left(\alpha_{i,s_k} \times \beta_{i,s_k} \times o_{i,V}/\left(Pr(D) \times c_{i,V}\right)\right)}{L};$$

$$u_{s_i} = \alpha_{1,s_i} \times \beta_{1,s_i};$$

And

$$t_{k,l} = \frac{\sum_{i=1}^{L-1} \left(\alpha_{i,s_k} \times \beta_{i+1,s_l} \times Pr(o_{i+1,*}|h_{i+1} = s_l)/Pr(D)\right)}{L-1};$$

The E-M algorithm iterates until the difference of probabilities between successive iterations is <0.00001.