STARRInIGHTS mini problem set
DPWG Bioinformatic Workshop
Jesse Shapiro jesse1@mit.edu
March 11, 2011

You have sampled 8 bacterial strains from 4 healthy (H) individuals and 4 sick (S) individuals affected by a disease of interest with a suspected microbial component. The strains are known to be closely related. You sequence the entire genome of each strain and define a ‘core’ genome of 2500 aligned base pairs. Your goal is to determine if:

- there is any evidence for homologous recombination among the strains
- the H and S strains represent separate gene pools
- there are any disease-associated regions of the core genome

1. Download the genome data for one of the 3 biological scenarios:

```
/scenario1/STARRI/
/scenario2/STARRI/
/scenario3/STARRI/
```

Save the files locally in a directory called

```
/STARRI/
```

which contains the data files:

	description:
genome1.1.subseqs.txt	list of subsequences (i,j) of informative SNPs
genome1.poly.variants.key.txt	genomic positions of informative SNPs
lk.txt	log-likelihoods and trees for each subseq (i,j)
strain.list	the names of all 8 strains
healthy.txt	the 4 strains isolated from healthy individuals
sick.txt	the 4 strains isolated from sick patients
genome1.fa	aligned core genome sequence for 8 strains
key.txt	summary of the above files
pB.list	a range of recombination breakpoint penalties (log-likelihoods of introducing a breakpoint)

and perl scripts:

dp.pl	dynamic prog. to combine genomic subseqs
dp+em.pl	dp w/ breakpoint penalty inferred by E-M
findML_noEM.pl	compares likelihoods across values of pB
findML.pl	compares likes; shows initial and E-M-inferred values of pB

2. Maximum-likelihood (ML) trees have been pre-computed based on every possible subsequence (i,j) of informative SNPs in the core genome. You can view these trees and their associated log-likelihoods in the file **lk.txt**.

Do the trees for different subsequences appear to be the same or different?

Combine the subsequences using dynamic programming and a range of recombination breakpoint penalties (found in the file pB.list) by running:

```
perl dp.pl /STARRI/ pB.list > dp.screenout
```

You can print a summary of the results using different breakpoint penalties by doing:

```
perl findML_noEM.pl starri.init.pB_-
```

This will print 4 columns of data to the screen:
(1) the initial set value of pB
(2) the final inferred value of pB (used for E-M only, below; n/a for no-E-M model)
(3) the log likelihood of the data under the current model and value of pB
(4) nb, the number of inferred recombination breakpoints in the core genome

How do different pB settings affect the results?
What is the most likely number of recombination breakpoints? number of events?

Repeat, but letting E-M converge on an optimal (maximum-likelihood) value of pB from different starting values:

```
perl dp+em.pl /STARRI/ pB.list > dp+em.screenout
```

```
perl findML.pl starriEM.init.pB_
```

Does the E-M converge? What is the likeliest number of breakpoints in the genome? Which value of pB provides the best (maximum likelihood) inferences of recombination breakpoints?

3. Using the value of pBinit that converges on the best pB (call this `[best_pBinit]`, let’s visualize the data and see which parts of the genome, if any, are associated with disease.

```
perl parseBlockParsTrees.pl /STARRI/key.txt /STARRI/ [best pBinit]
STARRI.2.0.EM
```

This produces output files:

	description
`pars.snp.pB_[best pBinit].txt`	`list of common parsimonious SNPs`
`homoplasic.snp.pB_-2.30.txt`	`list of common homoplasic SNPs`
`trees.pB_[best_pBinit].txt`	`list of blocks supporting diff. trees`
`1.break.plot.txt`	`locations of breakpoints btw. blocks`

Now we will use this output to make a final summary of our results and plot it visually:

```
perl parseBlockIngroupStats.pl /STARRI/key.txt /STARRI/ [best pBinit]
STARRI.2.0.EM pars.snp.pB_[best pBinit].txt healthy.txt sick.txt
```

This produces our final results file:

```
trees.pB_[best_pBinit]+stats.ingr.healthy.txt.txt
```

and the directory containing R code snippets to make plots:

```
output.pB_[best pBinit]/
```

containing files:

[#].Nsup.txt	to plot heatmap of # SNPs in a block supporting each tree partition
[#].support.txt	to plot heatmap of fraction of SNPs supporting each tree partition
[#].reject.txt	to plot heatmap of # SNPs in a block inconsistent w/ each partition
[#].Ykey.txt	key to the Y axis; a list of tree partitions ranked by their prevalence genomewide

You can plot a heatmap of support for different tree partitions across the genome by pasting `[#].support.txt` into an R command window. Brighter shades or white/gray indicate stretches of the genome supporting different phylogenetic partitioning of strains (shown on the Y-axis; see `[#].Ykey.txt`).

Is there any support for association between microbes and disease state (healthy/sick)? Is association localized to certain genomic regions? If so, what downstream analyses would you perform on these regions? If not, can you suggest another strategy to pinpoint genes/mutations associated with disease?