an assignment: gene prediction

We want to predict genes from the bacterial genomic sequences by using HMM. As a simple toy example, we will train HMM by using only one strain of *Staphylococcus aureus* and predict genes in another strain of *Staphylococcus epidermidis*. You can choose any programming language such as java, python, and C for implementation.

At first, you will consider the followings in order to design your HMM

- states (for example, gene region, intergenic region, start, stop)
- symbols (for example, A, C, G, T, N....)
- emission probabilities (for example, 1st order, second order,...)
- transition probabilities
- initial probabilities

an assignment: gene prediction

For training (estimating model parameters),

- your program (train.c) gets two input files for the gene region and the intergenic region.
  - gene region: train_gene.fa (from the portal)
  - intergenic region: train_non.fa (from the portal)
- your program (train.c) can give them predefined emission probability for some states such as start and stop.
- your program (train.c) outputs model parameters in a file, which will be used for gene prediction. Thus, the output of your training program (train.c) is a file that includes model parameters (model.txt). You can save your parameters in any format. But you should specify the format in detail in your document.

an assignment: gene prediction

For testing (predicting genes),

- your program (test.c) gets two input files: one input file for a set of DNA sequences for testing and one input file for the parameters estimated
    - sequence: test.fa (from the portal)
    - model parameters: model.txt (from your file)
- your program (test.c) includes mainly Viterbi algorithm including backtracking procedure.
- your program (test.c) outputs the prediction result(result.txt) in your own format in a file. Your program should clearly specify which region is gene region.

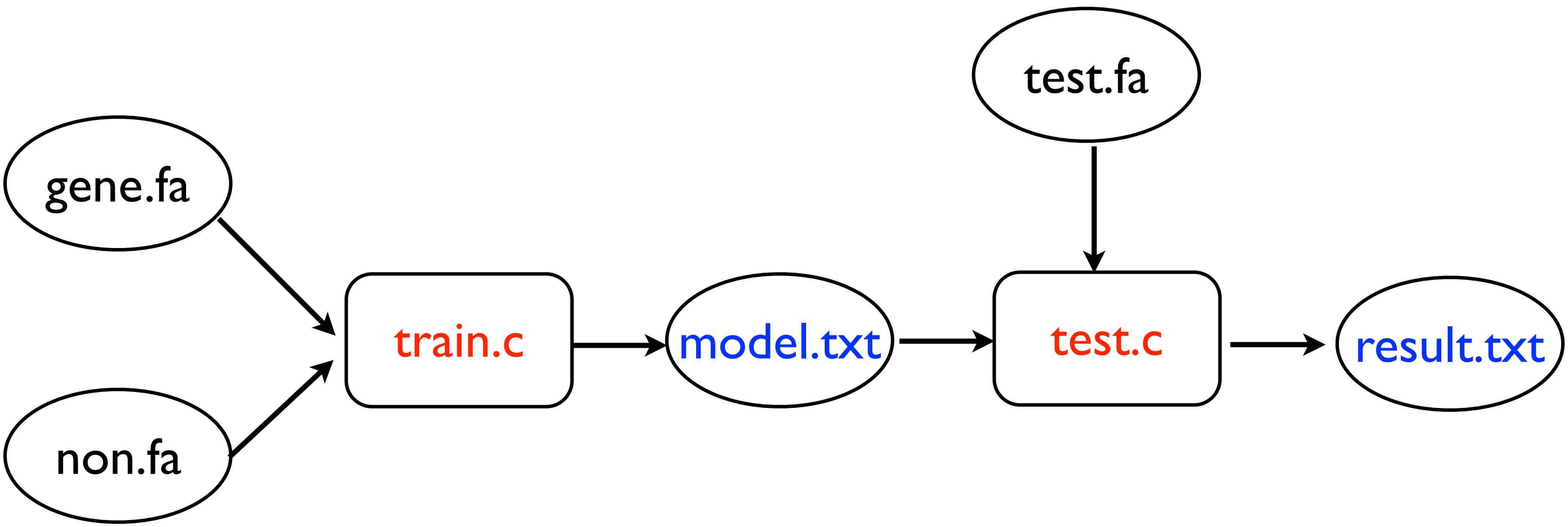
an assignment: gene prediction

Submission

- due date: 2015/11/28(Sat) 11:59 PM
- you should submit a zip file (your_srudent_id.zip) in the portal
- you should include train.c, test.c, manual.txt, and report.pdf
- late submission policy: 10% of your assignment score is subtracted per day

an assignment: gene prediction

an assignment: gene prediction



an assignment: gene prediction

train_gene.fa

```
>1
ATGTCGGAAAAAGAAATTTGGGAAAAAGTGCTTGAAATTGCTCAAGAAAAATTATCAGCTGTAAGTTACT
CAACTTTCCTAAAAGATACTGAGCTTTACACGATTAAAGATGGTGAAGCTATCGTATTATCGAGTATTCC
TTTTAATGCAAATTGGTTAAATCAACAATATGCTGAAATTATCCAAGCAATCTTATTTGATGTTGTAGGC
TATGAAGTTAAACCTCACTTTATTACTACTGAAGAATTAGCAAATTATAGTAATAATGAAACTGCTACTC
CAAAAGAAACAACAAAACCTTCTACTGAAACAACTGAGGATAATCATGTGCTTGGTAGAGAGCAATTCAA
TGCCCATAACACATTTGACACTTTTGTAATCGGACCCGGTAACCGCTTTCCACATGCAGCGAGTTTAGCT
GTGGCCGAAGCACCAGCCAAAGCGTACAATCCATTATTTATCTATGGAGGTGTTGGTTTAGGAAAAACCC
ATTTAATGCATGCCATTGGTCATCATGTTTTAGATAATAATCCAGATGCCAAAGTGATTTACACATCAAG
TGAAAAATTCACAAATGAATTTATTAAATCAATTCGTGATAACGAAGGTGAAGCTTTCAGAGAAAGATAT
CGTAATATCGACGTCTTATTAATCGATGATATTCAGTTCATACAAAACAAGGTACAAACACAAGAAGAAT
TTTTCTATACTTTTAATGAATTGCATCAGAATAACAAGCAAATAGTTATTTCGAGTGATCGACCACCAAA
GGAAATTGCACAATTAGAAGACCGATTACGTTCACGCTTTGAATGGGGGCTAATTGTTGATATTACGCCA
CCAGATTATGAAACTCGAATGGCAATTTTGCAGAAGAAAATTGAAGAAGAAAAATTAGATATTCCACCAG
AAGCTTTAAATTATATAGCAAATCAAATTCAATCTAATATTCGTGAATTAGAAGGTGCATTAACACGTTT
ACTTGCATATTCACAATTATTAGGAAAACCAATTACAACTGAATTAACTGCTGAAGCTTTAAAAGATATC
ATTCAAGCACCAAAATCTAAAAAGATTACCATCCAAGATATTCAAAAAATTGTAGGCCAGTACTATAATG
TTAGAATTGAAGATTTCAGTGCAAAAAAACGTACAAAGTCAATTGCATATCCGCGTCAAATAGCTATGTA
CTTGTCTAGAGAGCTTACAGATTTCTCATTACCTAAAATTGGTGAAGAATTTGGTGGGCGTGATCATACG
ACCGTCATTCATGCTCATGAAAAAATATCTAAAGATTTAAAAGAAGATCCTATTTTTAAACAAGAAGTAG
>2
ATGATGGAATTCACTATTAAAAGAGATTATTTTATTACACAATTAAATGACACATTAAAAGCTATTTCAC
CAAGAACAACATTACCTATATTAACTGGTATCAAAATCGATGCGAAAGAACATGAAGTTATATTAACTGG
TTCAGACTCTGAAATTTCAATAGAAATCACTATTCCTAAAACTGTAGATGGCGAAGATATTGTCAATATT
TCAGAAACAGGCTCAGTAGTACTTCCTGGACGATTCTTTGTTGATATTATAAAAAAATTACCTGGTAAAG
ATGTTAAATTATCTACAAATGAACAATTCCAGACATTAATTACATCAGGTCATTCTGAATTTAATTTAAG
TGGCTTAGATCCAGATCAATATCCTTTATTACCTCAAGTTTCTAGAGATGACGCAATTCAATTGTCGGTA
AAAGTGCTTAAAAACGTGATTGCACAAACAAATTTTGCAGTGTCCACCTCAGAAACACGCCCAGTACTAA
CTGGTGTGAACTGGCTTATACAAGAAAATGAATTAATATGCACAGCGACTGACTCACACCGCTTGGCTGT
AAGAAAGTTGCAGTTAGAAGATGTTTCTGAAAACAAAAATGTCATCATTCCAGGTAAGGCTTTAGCTGAA
TTAAATAAAATTATGTCTGACAATGAAGAAGACATTGATATCTTCTTTGCTTCAAACCAAGTTTTATTTA
AAGTTGGAAATGTGAACTTTATTTCTCGATTATTAGAAGGACATTATCCTGATACAACACGTTTATTCCC
TGAAAACTATGAAATTAAATTAAGTATAGACAATGGGGAGTTTTATCATGCGATTGATCGTGCCTCTTTA
TTACCCCCTCAACCTCCTAATAACCTTATTAAATTAACTACACCTCATCACCTTCTTCAATTCTCTTCTA
```

an assignment: gene prediction

test.fa

```
>test
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCC
AGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTG
AAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTT
GCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGC
TGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGT
TACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCT
GAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCG
CCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGC
TGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGT
CAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCG
CTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATAC
CGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGC
ATTTCCAATCTGAATAACATGGCAATGTTCAGCGTTTCTGGTCCGGGGATGAAAGGGATGGTCGGCATGG
CGGCGCGCGTCTTTGCAGCGATGTCACGCGCCCGTATTTCCGTGGTGCTGATTACGCAATCATCTTCCGA
ATACAGCATCAGTTTCTGCGTTCCACAAAGCGACTGTGTGCGAGCTGAACGGGCAATGCAGGAAGAGTTC
TACCTGGAACTGAAAGAAGGCTTACTGGAGCCGCTGGCAGTGACGGAACGGCTGGCCATTATCTCGGTGG
TAGGTGATGGTATGCGCACCTTGCGTGGGATCTCGGCGAAATTCTTTGCCGCACTGGCCCGCGCCAATAT
CAACATTGTCGCCATTGCTCAGGGATCTTCTGAACGCTCAATCTCTGTCGTGGTAAATAACGATGATGCG
ACCACTGGCGTGCGCGTTACTCATCAGATGCTGTTCAATACCGATCAGGTTATCGAAGTGTTTGTGATTG
GCGTCGGTGGCGTTGGCGGTGCGCTGCTGGAGCAACTGAAGCGTCAGCAAAGCTGGCTGAAGAATAAACA
TATCGACTTACGTGTCTGCGGTGTTGCCAACTCGAAGGCTCTGCTCACCAATGTACATGGCCTTAATCTG
GAAAACTGGCAGGAAGAACTGGCGCAAGCCAAAGAGCCGTTTAATCTCGGGCGCTTAATTCGCCTCGTGA
AAGAATATCATCTGCTGAACCCGGTCATTGTTGACTGCACTTCCAGCCAGGCAGTGGCGGATCAATATGC
CGACTTCCTGCGCGAAGGTTTCCACGTTGTCACGCCGAACAAAAAGGCCAACACCTCGTCGATGGATTAC
TACCATCAGTTGCGTTATGCGGCGGAAAAATCGCGGCGTAAATTCCTCTATGACACCAACGTTGGGGCTG
GATTACCGGTTATTGAGAACCTGCAAAATCTGCTCAATGCAGGTGATGAATTGATGAAGTTCTCCGGCAT
TCTTTCTGGTTCGCTTTCTTATATCTTCGGCAAGTTAGACGAAGGCATGAGTTTCTCCGAGGCGACCACG
CTGGCGCGGGAAATGGGTTATACCGAACCGGACCCGCGAGATGATCTTTCTGGTATGGATGTGGCGCGTA
AACTATTGATTCTCGCTCGTGAAACGGGACGTGAACTGGAGCTGGCGGATATTGAAATTGAACCTGTGCT
GCCCGCAGAGTTTAACGCCGAGGGTGATGTTGCCGCTTTTATGGCGAATCTGTCACAACTCGACGATCTC
```

an assignment: gene prediction

model.txt

```
Transition=
GG      0.9990
GE      0.0010
ER      0.9965
ES      0.0030
ES1     0.0005
    :
```

Emission_gene=

```
0.3711  0.1152  0.3455  0.1682
0.3279  0.1773  0.2255  0.2693
0.3166  0.2363  0.2278  0.2193
0.3403  0.1270  0.3192  0.2135
0.3715  0.1084  0.3509  0.1692
0.3901  0.1444  0.1983  0.2672
0.3984  0.1748  0.2546  0.1721
0.3293  0.1116  0.3209  0.2381
0.3920  0.1266  0.3009  0.1805
0.2963  0.1438  0.2594  0.3005
0.3596  0.1489  0.2848  0.2067
0.3167  0.1256  0.3263  0.2314
0.3651  0.0965  0.3507  0.1876
0.3770  0.1496  0.2184  0.2549
0.3866  0.1536  0.2784  0.1814
0.3256  0.0923  0.3556  0.2266
```

an assignment: gene prediction

result.txt

34	310	ACCCCCCGGTGTCGACGTCA....
400	520	CGGTGTCGACGTCAAAGCTGCAC....
600	750	TCGACGTCAAAGCTGCAC....
990	1100	AAAGCTGCACCGCTGCGTGCGACGACTCAG....

an assignment: gene prediction