

Algorithms in Bioinformatics I, WS2006/07

Assignment sheet # 15

Christian Rausch

February 5, 2007

1 Implementation of Markov chains (4 points)

Please write a java class `albi.hmm.MarkovChain`. This class should support the following methods:

`MarkovChain()` - constructor
`read(Reader r)` - read in a Markov chain
`write(Writer w)` - write
`get/setNumberOfStates(int n)` -get or set the number of states
`get/setStates(String states)` -get or set the state labels (single letters)
`get/setTransitionMatrix(double[][] trans)` - get or set the transition probs.

`trainTransitionMatrix(String data)` - set the transition matrix from a string of training data

`double getLogProbability (String str)` - get the log of the probability associated with a string of states

Please use the following file format to describe an HMM:

```
# Number of states:
6
# State labels:
A C G T * +
# Transition matrix:
.2995 .2045 .2845 .2095 0 .002
.3215 .2975 .0775 .3015 0 .002
.2475 .2455 .2975 .2075 0 .002
.1765 .2385 .2915 .2915 0 .002
.2495 .2495 .2495 .2495 0 .002
0 0 0 0 0 1
```

2 Constructing Markov chains from training data (3 points)

Using the class `albi.hmm.MarkovChain`, write a Java program `albi.hmm.MakeDNAMarkovChain` that reads as input a string of DNA from a file and produces as output a corresponding Markov chain.

Usage: `java albi.hmm.MakeDNAMarkovChain infile outfile`.

To do this, assume a uniform probability of starting with an A, C, G or T. Also, assume that the probability of transitioning into the end state is 0.002 for each state A, C, G or T. Apply this program to the two files `CpG.seq` and `nonCpG.seq` to obtain two new files `CpG.mc` and `nonCpG.mc`. These two files represent data in CpG and non-CpG island sequence, respectively.

3 Classifying sequence using Markov chains (3 points)

Using the class `albi.hmm.MarkovChain`, write a program `albi.hmm.ApplyMarkovChain` that takes as input a Markov chain description and a second file containing a sequence. Output is the log probability $\log P$ that the Markov chain computes for the given sequence.

Usage: `java albi.hmm.ApplyMarkovChain infile outfile`.

Using the Markov chains `CpG.mc` and `nonCpG.mc`, apply the program `albi.hmm.ApplyMarkovChain` to all files `data01.seq` – `data20.seq`. For each file, report the probability that the contained sequence is a CpG island or not, respectively.

All data files and necessary Java files are contained in:

<http://www-ab.informatik.uni-tuebingen.de/teaching/ws06/albi1/assign/java/assign15.zip>.

Assignments due: **Monday, February 12, 10am**