Hidden Markov Models and some applications

Oleg Makhnin
New Mexico Tech
Dept. of Mathematics

November 11, 2011

HMM description

Application to genetic analysis

Applications to weather and climate modeling

Discussion

HMM description

Hidden Markov Models (HMM)

a) Markov Chain describing the system state (hidden)

Markov Property: the value of X_{t+1} depends only on immediate past X_t

b) Emission probabilities describing how the hidden state affects the observable outcomes.

Observed states: Y_t depends only on X_t

HMM past and present

- • Originated in speech processing ($\sim$ 1970)
- • Popular in
  - - Fraud detection (Scott 2002)
  - - Language and music analysis
  - - Bioinformatics (analysis of nucleic acid and protein sequences)
  - - ... ...

A toy example

Suppose a coin can be switched from "fair" coin (Head to Tail ratio 1:1) to "biased" coin (Head to Tail ratio 3:1). It cannot be switched too often because the people will notice. It also cannot stay biased for very long, or the people will become suspicious. Thus, we will have the *transition matrix* **P** and the *emission matrix* **G**

$$\mathbf{P} = \begin{matrix} \\ \texttt{fair} \\ \texttt{biased} \end{matrix} \begin{bmatrix} \texttt{next fair} & \texttt{next biased} \\ 0.95 & 0.05 \\ 0.1 & 0.9 \end{bmatrix}$$

$$\mathbf{G} = \begin{matrix} \\ \texttt{fair} \\ \texttt{biased} \end{matrix} \begin{bmatrix} \texttt{Heads} & \texttt{Tails} \\ 0.5 & 0.5 \\ 0.9 & 0.1 \end{bmatrix}$$

Estimation methods

- "Decoding"
- Parameter estimation

Let F = "fair", B = "biased", H = "Heads", T = "Tails"

Suppose that P and G matrices are known. For example, let the sequence $Y_{1:3} = \texttt{HHH}$ be observed. For every possible sequence of $X_{1:3}, \texttt{FFF}$, for example, we can calculate conditional probability

$$P(FFF|HHH) = \frac{P(HHH \mid FFF)P(FFF)}{\sum_{X,Y,Z \in \{F,B\}} P(HHH \mid XYZ)P(XYZ)}$$

(Bayes' Theorem).

Estimation methods

Then, one possible decoding method is to determine the *most likely* sequence of X's to generate the observed Y's.

Another method is based on determining $P(X_t = F \mid Y_{1:T})$, for all $t = 1, ..., T$.

This calculation is very costly (2^T possible values for $X_{1:T}$)

Estimation methods

Exploiting the Markov property of X can be done iteratively

- *Forward pass*:

let $P_t^F = P(X_t = F \mid Y_{1:t})$

$P_1^F = P(X_1 = F \mid Y_1)$ is easy, then

$P(X_t = F \text{ and } X_{t+1} = F \mid Y_{1:(t+1)})$ is obtained using Markov Property and then

$P(X_{t+1} = F \mid Y_{1:(t+1)}) \equiv P_{t+1}^F$ is obtained etc.

Eventually we get $P(X_T = F \mid Y_{1:T})$.

Estimation methods

- *Backward pass*:

Based on $P(X_t = F \mid Y_{1:T})$, apply Markov Property again to get $P(X_{t-1} = F \mid Y_{1:T})$, etc all the way down to X_1.

A similar version of backward pass lets you find the most likely sequence, or to sample the values of X_t (conditional on all data) one by one.

Estimation methods

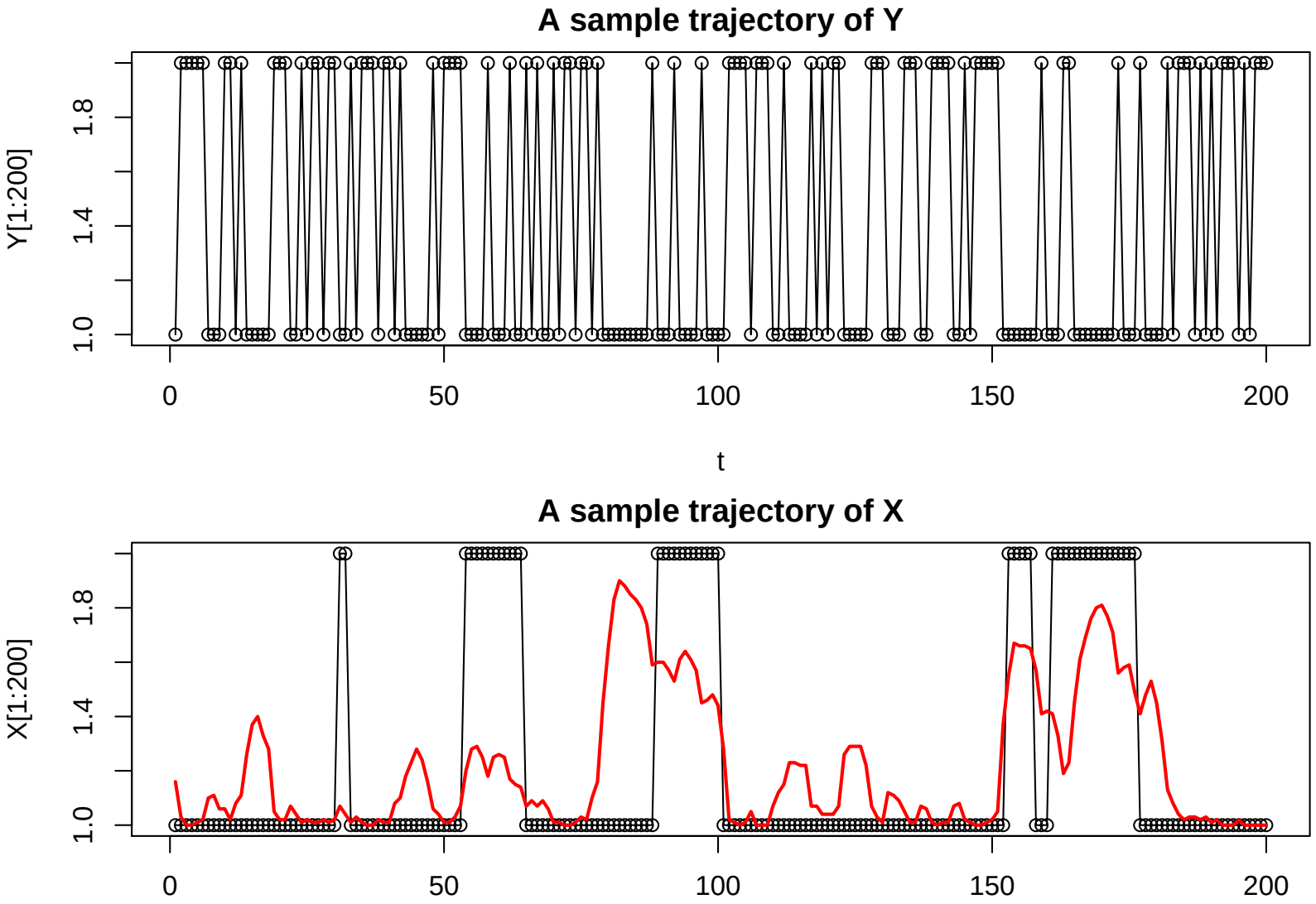
If matrices **P**, **G** and other parameters are not known, then we can use an iterative sampling method within the Gibbs sampler:

- sample the sequence $X_{1:T}$ given the current values of **P**, **G**, then
- update **P**, **G** based on current sequence $X_{1:T}$

  etc ... etc ...

until we get a long enough sample of X's. We can use it to estimate the probability that $P(X_t = F \mid \texttt{all data})$ for all t

Output: observed Y, simulated X and the decoded probabilities (red line) of $X_t = B$.



HMMs and bioinformatics

Genetic code: `TTGGTTATCGTTTTTCAC...` four "letters" (bases) A, C, G, T (adenine, cytosine, guanine, thymine).

Some sections of DNA code for proteins ("coding regions", about 1.5% of human genome), most do not ("junk DNA") but might have some function.

First uses of HMM in bioinformatics: late 80's (Churchill)

Applications to gene-finding, protein sequence alignment etc etc

CpG islands

CpG islands *(Irizarry et al, 2009)*:

Areas in the genome with a higher concentration of C and G bases, along with a higher than usual occurrence of CG-pairs. Have some important, still not well understood biological functions, e.g. in regulating gene expression.
Original definition of CpG island: a region with at least 200 bp, based on GC content higher than 50% and OE ratio above 0.6

note: OE (observed/expected ratio) is given by

```
p(CG) = OE*p(C)*P(G)
```

This definition misses some regions with the function similar to known CpG islands. Also, it might differ for different species.

Need a more flexible definition.

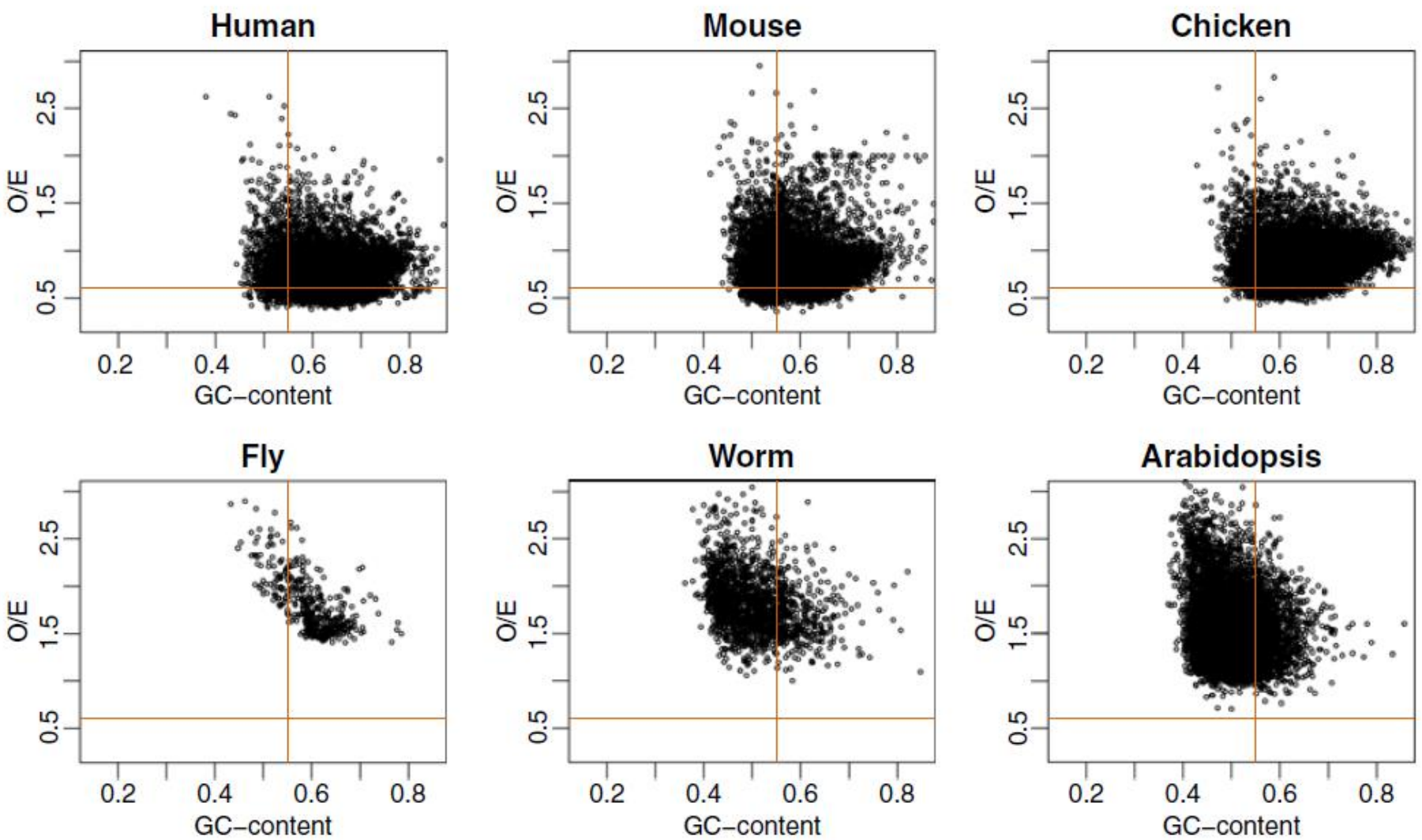


Fig. 1 The O/E and GC content were computed for all islands found by our HMM procedure. O/E is plotted against GC content for six
The *vertical and horizontal lines* represent the cutoffs used by the Gardiner-Garden and Frommer CGI definition

A little problem: CpG appearance cannot be modeled by an HMM (since Y_t is conditionally independent of Y_{t-1}, given the hidden state X_t).

Solution: run HMM not on single bases, but on the segments of, say, 20 bp.
The emission probabilities are assumed Poisson with the means

$$a_i \times 20 \times p_C \times p_G$$

where $i = 1$ is baseline, and $i = 2$ is a CpG island state.

For Arabidopsis (below), $a_1 \approx 0.5$, $a_2 \approx 0.9$, and $p_C \approx p_G \approx 0.2$.

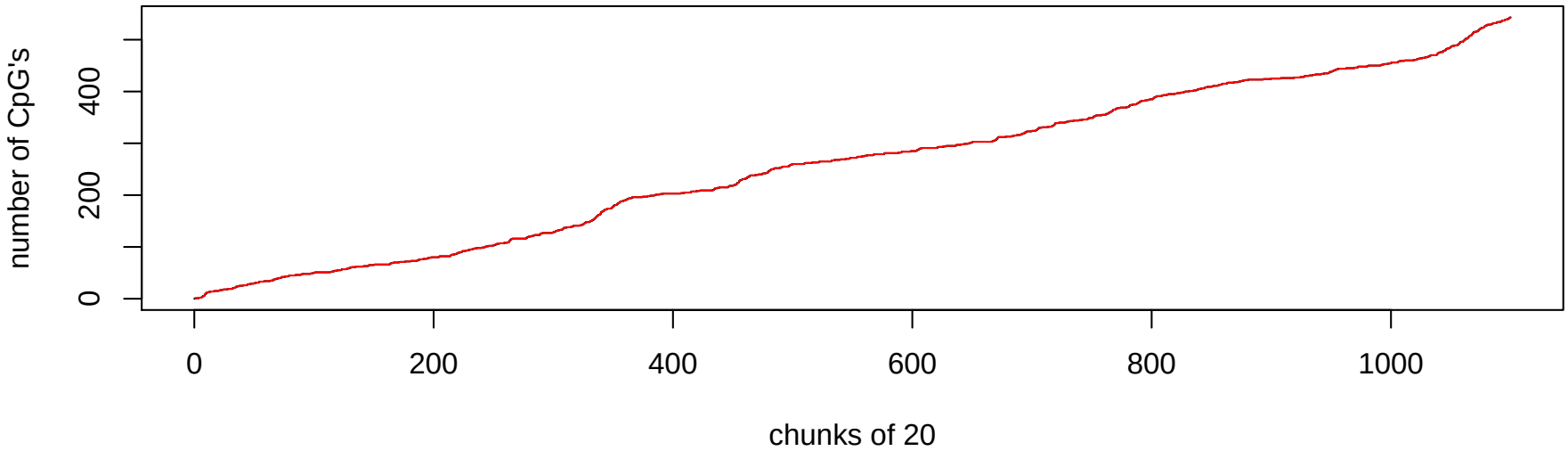
Genetic example: Arabidopsis

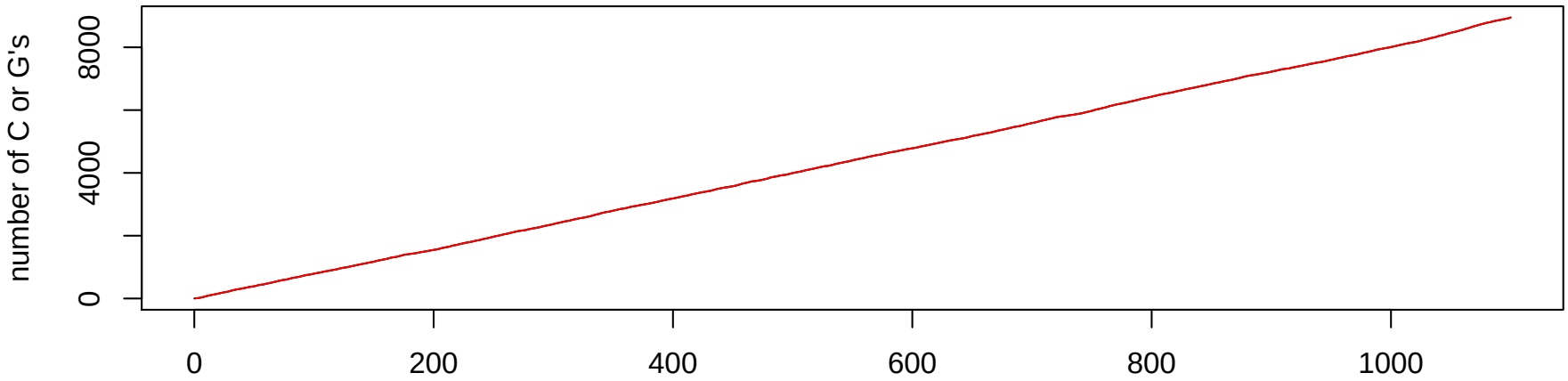
- • has a small genome, 157 megabase pairs (Mbp) [human genome ∼2900 Mbp]
- • popular in plant genetic studies, first plant genome to be sequenced

`TTGGTTATCGTTTTTCAC...`

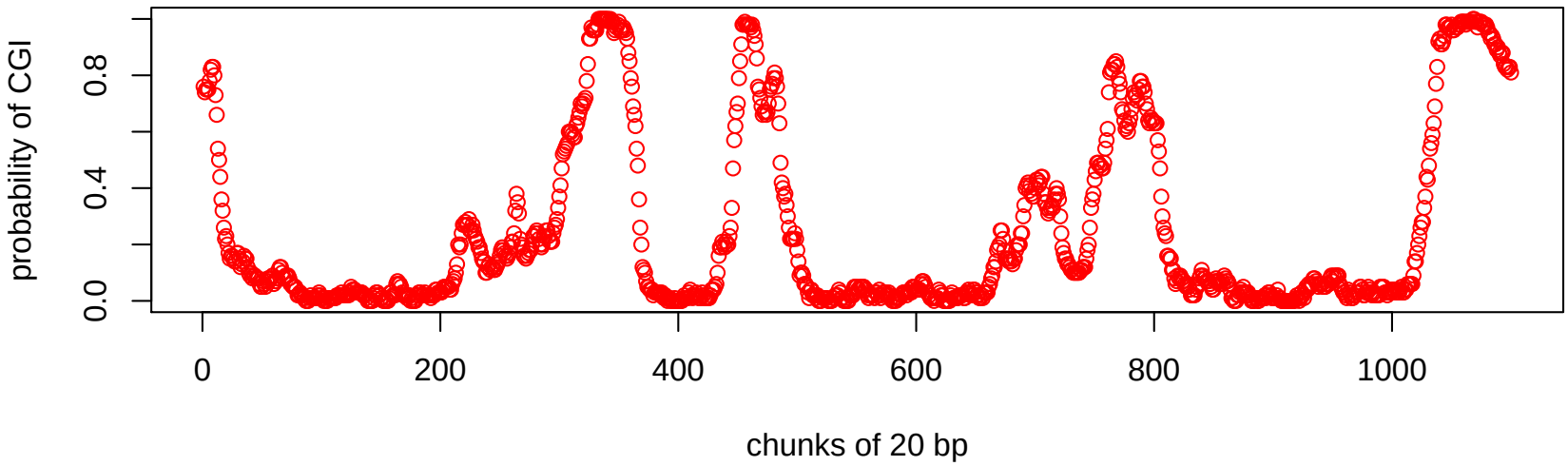
count of CG's in the chunks of 20 bp:

1 0 0 1 0 0 2 1 4 2 1 1 0 1 0 0 1 0 0 0 .....



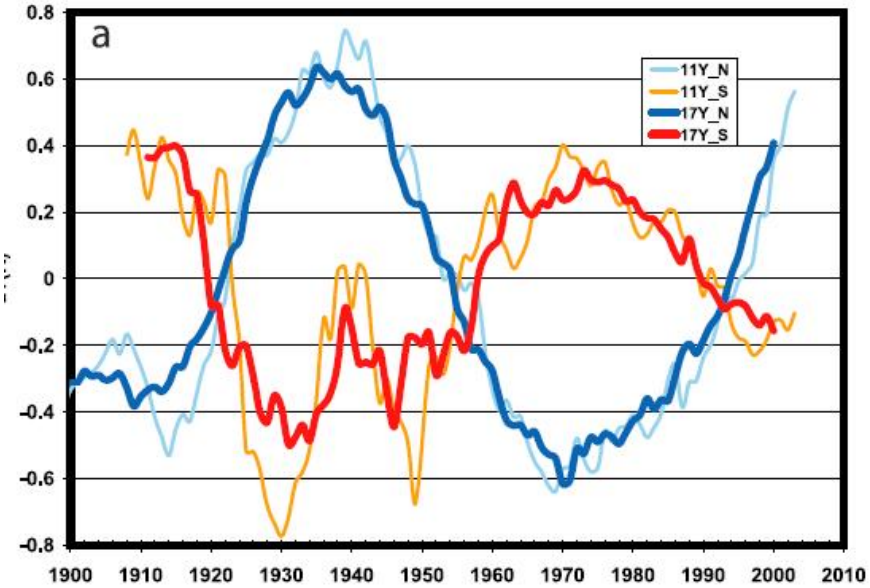


HMM decoding results:



A possible appilcation to climate modeling:

Recently, Chylek et al established a connection between Arctic and Antarctic climate:



likely due to the Atlantic Meridian Overturning Circulation (AMOC).
HMM's can be used to track similar "seesaw" behavior in the past (based on ice core data), when AMOC turns on/off.

Bibliography

- *A species-generalized probabilistic model-based definition of CpG islands*, Irizarry, Wu & Feinberg, Mamm Genome (2009) 20:674-680

- *Bayesian methods for Hidden Markov Models*, S.L. Scott, Journal of Amer. Stat. Assoc. (2002) **97**, 337-351

- *Twentieth century bipolar seesaw of the Arctic and Antarctic surface air temperatures*, Chylek, Folland, Lesins and Dubey, GRL (2010)

QUESTIONS?

THANK YOU!

see www.nmt.edu/~olegm/talks/HMM