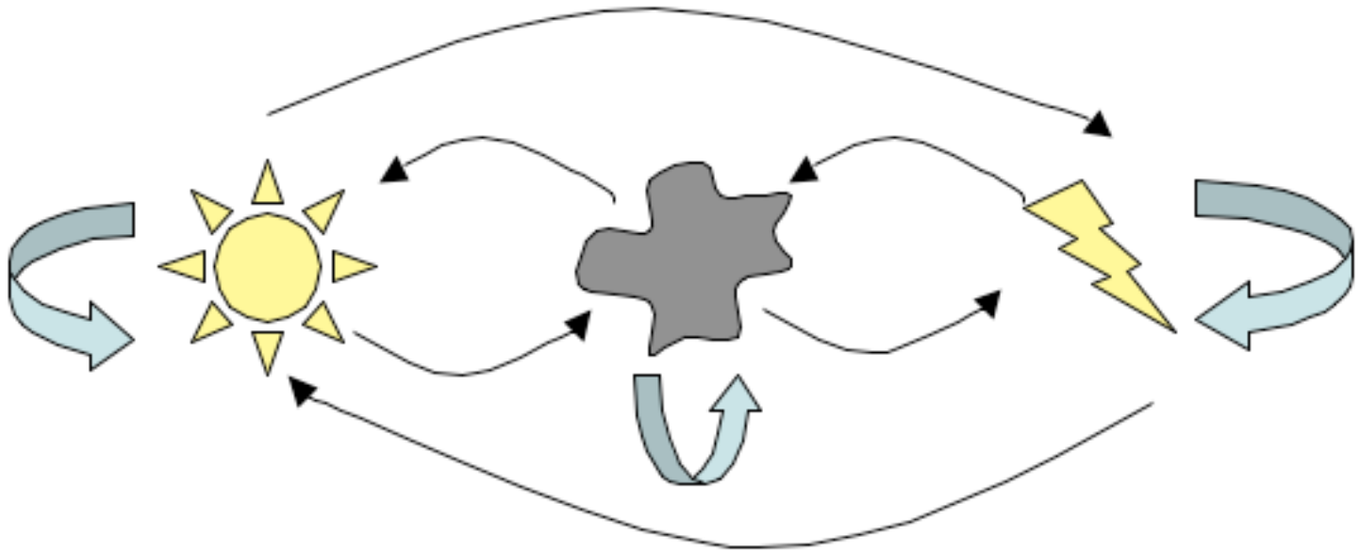
Baum-Welch and HMM applications

December 4, 2018

Markov chains

3 states of weather: sunny, cloudy, rainy
Observed once a day at the same time



All transitions are possible, with some probability
Each state depends only on the previous state

Hidden Markov Models

Weather yesterday \ Weather today	Sunny	Cloudy	Rainy
Sunny	0.50	0.20	0.30
Cloudy	0.10	0.60	0.30
Rainy	0.20	0.40	0.40

	Dog in	Dog out	Dog porch
Sun	0.2	0.7	0.1
Cloud	0.4	0.4	0.2
Rain	0.7	0.1	0.2

All we observe is the dog:

`IOOOIPIIIOOOOOPPIIIIIPI`

What's the underlying weather (the hidden states)?

How likely is this sequence, given our model of how the dog works?

What portion of the sequence was generated by each state?

Hidden Markov Models: the three questions

Evaluation

Given a HMM, M, and a sequence of observations, x

Find P(x|M)

Decoding

Given a HMM, M, and a sequence of observations, x

Find the sequence Q of hidden states that maximizes P(x, Q|M)

Learning

Given an unknown HMM, M, and a sequence of observations, x

Find parameters θ that maximize P(x|θ, M)

review

x are observations $\in A$, $q_1 \ldots q_n$ are hidden states $\in S$

$$\alpha(t,i) = p(x_1 x_2 \ldots x_t, q_t = S_i)$$

$$\beta(t,i) = p(x_T x_{T-1} \ldots x_t | q_t = S_i)$$

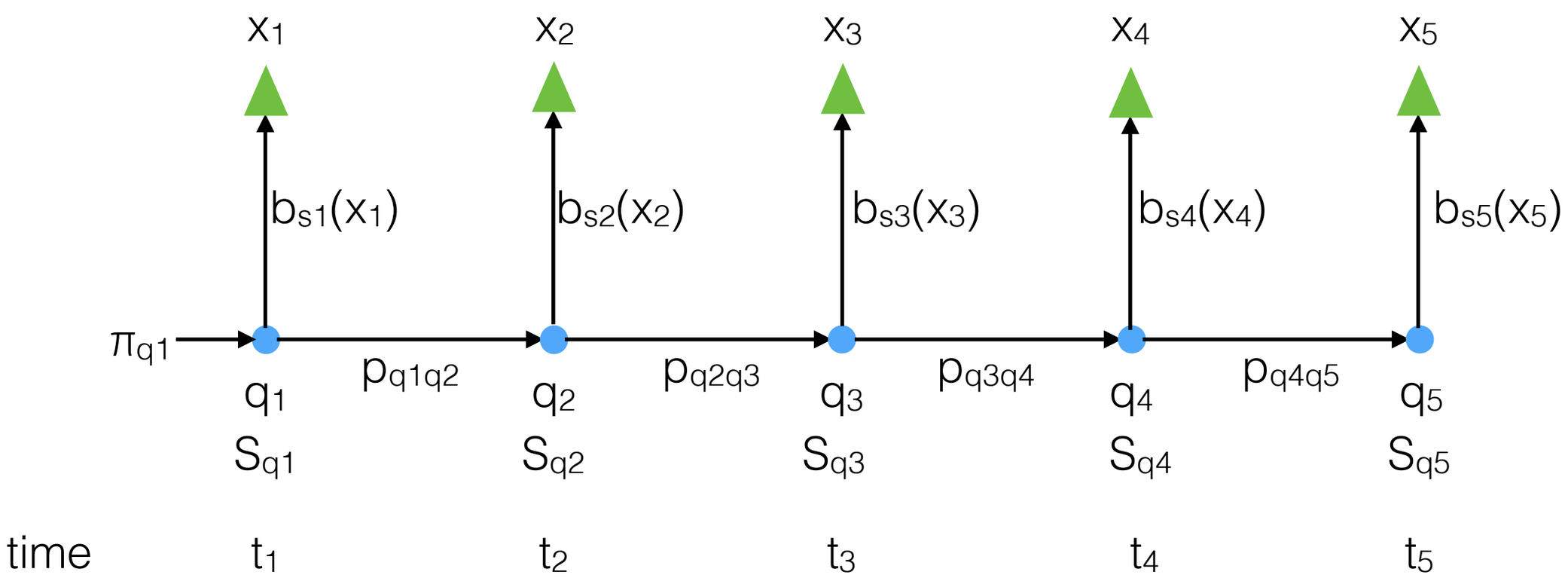
$$p(x, q_t = S_i | M) = \alpha(t,i)\beta(t,i)$$

$$p(x|M) = \sum_{i=1}^{N} \alpha(T,i) \qquad p(x|M) = \sum_{i=1}^{N} \beta(1,i)$$

$$\alpha(t,i) = p(x_1 x_2 \dots x_t, q_t = S_i)$$

$$\beta(t,i) = p(x_T x_{T-1} \dots x_t | q_t = S_i)$$

$$p(x, q_t = S_i | M) = \alpha(t,i)\beta(t,i)$$



Baum-Welch expectation maximization algorithm

You have: observed data

You want: parameters of the HMM that generated that data

Problem: the calculation space is too big for exact calculation -> use heuristic method (even though it's a partial solution it's very useful!)

We are finding locally optimal parameters.

Baum-Welch expectation maximization algorithm

Assume: data come from some random process that we can fit to a HMM

Assumption #1: alphabet A and the number of states, N, are fixed. Transition, emission and initial distribution probabilities are all unknown.

Baum-Welch expectation maximization algorithm

Assumption #2: data are a set of observed sequences $\{x^{(d)}\}$ each of which has a hidden state sequence Q^d

Assumption #3: we can set all parameters/probabilities to some initial values

Can choose from some uniform distribution

Can choose to incorporate some prior knowledge

Can just be random

Cannot be flat

Baum-Welch expectation maximization algorithm

$$\alpha(t,i) = p(x_1 x_2 \ldots x_t, q_t = S_i)$$

$$\beta(t,i) = p(x_T x_{T-1} \ldots x_t | q_t = S_i)$$

$$p(x, q_t = S_i | M) = \alpha(t,i)\beta(t,i)$$

first step: make up some probabilities.
need vector of initial values
emission matrix
transition matrix

Baum-Welch expectation maximization algorithm

testing and refining the probabilities:

$$P(q_t = S_i, q_{t+1} = S_l | x, \theta) = \frac{\alpha(t,i) p_{il} b_l(x_{t+1}^d) \beta(t+1,l)}{P(x)}$$

$$x_1 x_2 x_3 \cdots x_t x_{t+1} \cdots x_{T-1} x_T$$

$$q_1 q_2 q_3 \cdots q_t q_{t+1} \cdots q_{T-1} q_T$$

$$\cdots S_i S_l \cdots$$

Baum-Welch expectation maximization algorithm

testing and refining the probabilities: transition matrix

$$p'_{il} = \sum_d \frac{\sum_t \alpha^d(t,i) p_{il} b_l(x^d_{t+1}) \beta^d(t+1,l)}{P(x^d)}$$

How to figure out the probability of a transition from hidden state *i* to ℓ:

1) postulate that transition at every single spot in every single observed sequence (separately)

2) see how those probabilities compare to the best probabilities for those observed sequences

3) use that ratio for the updated $p_{i\ell}$ transition probability

Baum-Welch expectation maximization algorithm

testing and refining the probabilities: emission matrix

$$b'_l(a) = \sum_d \frac{\sum_{t|x_t^d=a} \alpha^d(t,l)\beta^d(t,l)}{P(x^d)}$$

Figure out the probability of an emission of symbol *a* from hidden state ℓ

1) postulate that hidden state under every symbol *a* in every single observed sequence (separately)

2) see how those probabilities compare to the best probabilities for those observed sequences

3) use that ratio for the updated b'$_\ell$(*a*) transition probability

Baum-Welch expectation maximization algorithm

Then recalculate $P(x^d|M, \theta)$ for all observed data in the learning set (use Forward, Backward, or Forward/Backward to do this)

Rinse & repeat . . .

Successive iterations increase P(data) and we stop when the probability stops increasing significantly (usually measured as log-likelihood ratios).

Baum-Welch example

I observe dog #2 at noon every day. Sometimes he's inside, sometimes he's outside.

I guess that since he can't open the door by himself (yet) that there is another factor, hidden from me, that determines his behavior

Since I am lazy I will guess that there are only two hidden states

Baum-Welch example

- guessing two hidden states. I need to invent a transition matrix and an emission matrix.

today

yesterday

	S1	S2
S1	0.5	0.5
S2	0.4	0.6

	in	out
S1	0.2	0.8
S2	0.9	0.1

initial: p(S1) = 0.3, p(S2) = 0.7

Baum-Welch example

one set of observations: `II, II, II, II, IO, OO, OI, II, II`

today

yesterday	S1	S2
S1	0.5	0.5
S2	0.4	0.6

	in	out
S1	0.2	0.8
S2	0.9	0.1

initial: p(S1) = 0.3, p(S2) = 0.7

Baum-Welch example

guess: if `II` came from S1•S2 the probability is

0.3 * 0.2 * 0.5 * 0.9 = 0.027

today

yesterday

	S1	S2
S1	0.5	0.5
S2	0.4	0.6

	in	out
S1	0.2	0.8
S2	0.9	0.1

initial: p(S1) = 0.3, p(S2) = 0.7

Baum-Welch example

estimating the transition matrix:

Seq	P(Seq) if S1•S2	Best P(seq)
II	0.027	0.3403 S2•S2
II	0.027	0.3403 S2•S2
II	0.027	0.3403 S2•S2
II	0.027	0.3403 S2•S2
IO	0.003	0.2016 S2•S1
OO	0.012	0.096 S1•S1
OI	0.108	0.108 S1•S2
II	0.027	0.3403 S2•S2
II	0.027	0.3403 S2•S2
Total	0.285	2.4474

Our estimate for the S1->S2 transition probability is now 0.285/2.4474 = 0.116. Calculate the S2->S1, S2->S2, S1->S1 as well and normalize so they add up to 1 as needed, to update the transition matrix.

Baum-Welch example

estimating the emission matrix:

Seq	Best P(Seq) if O came from S1	Best P(seq)
IO	0.2016 (S2•S1)	0.2016 (S2•S1)
OO	0.096 (S1•S1)	0.096 (S1•S1)
OI	0.108 (S1•S2)	0.108 (S1•S2)

Baum-Welch example

estimating initial probabilities:

1. assume all sequences start with hidden state S1, calculate best probability
2. assume all sequences start with hidden state S2, calculate best probability
3. normalize to 1

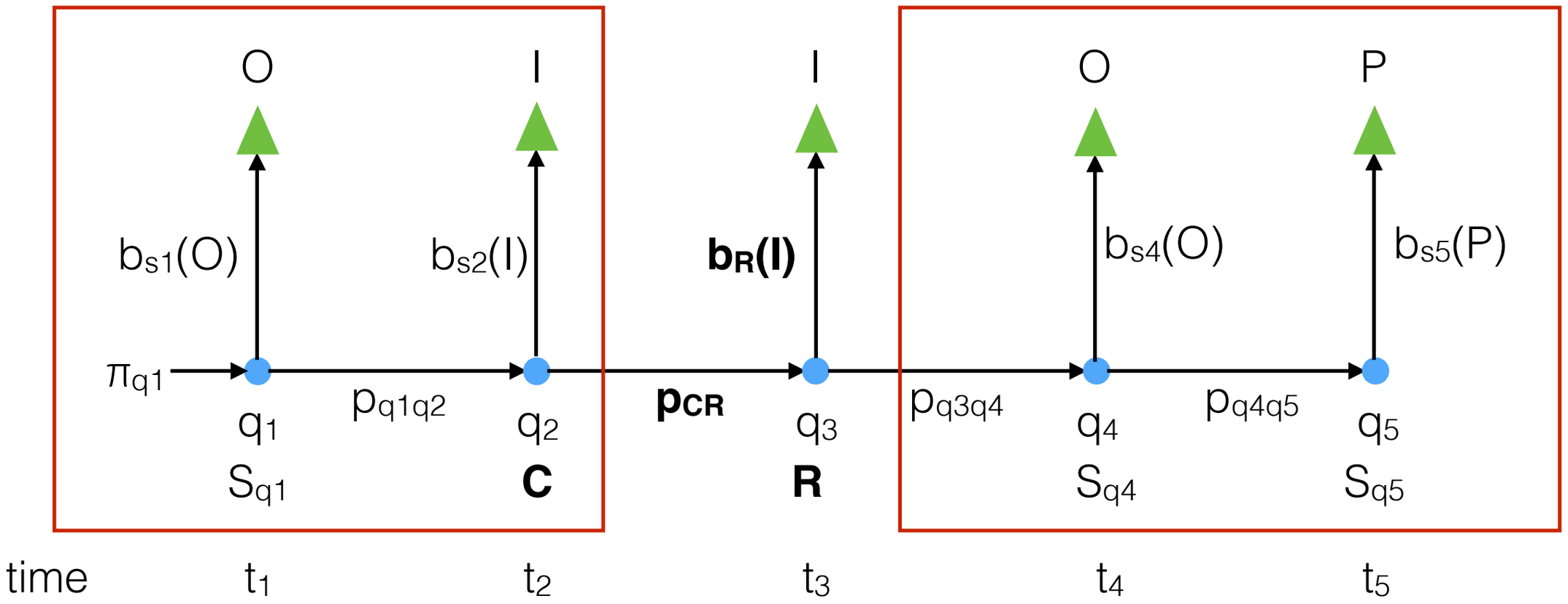
Baum-Welch example

Now we have generated updated transition, emission, and initial probabilities. Repeat this method until those probabilities converge.

If you have guessed the wrong number of hidden states, it will be clear, though it's a very bad strategy to go through a huge range of possible hidden states to find the best model – you will over-optimize.

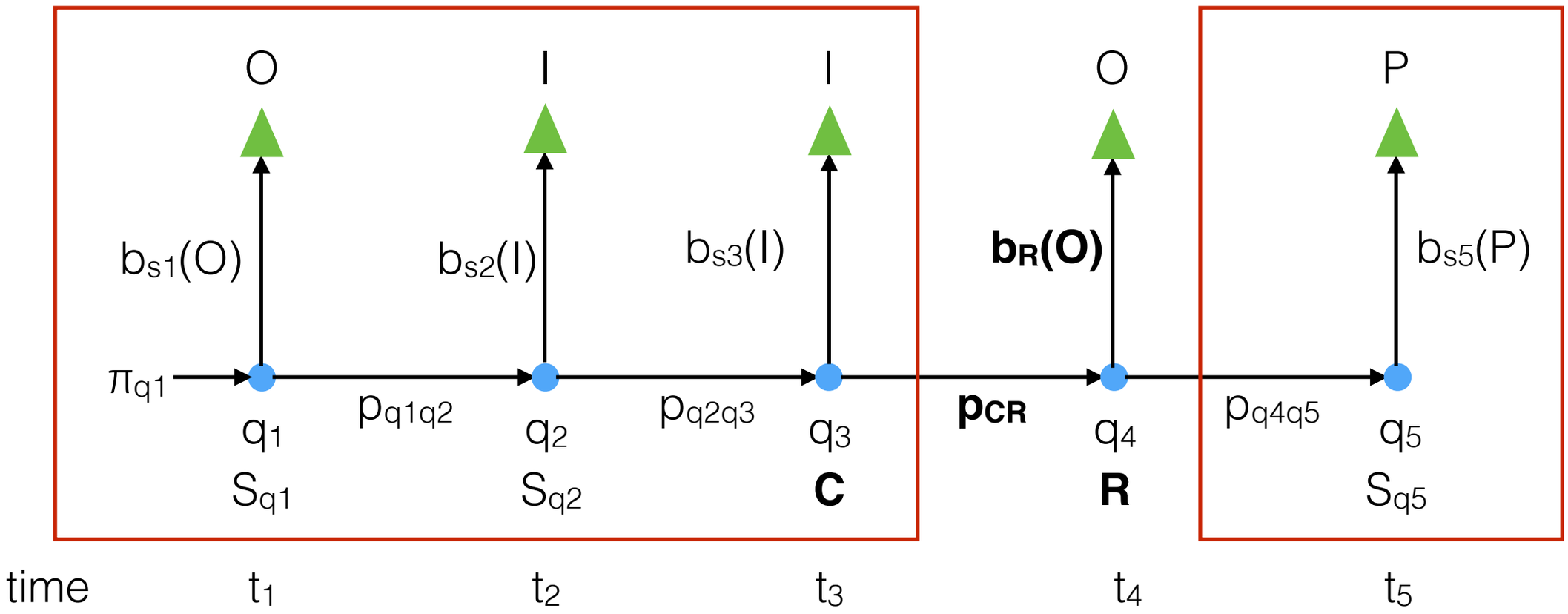
$$P(q_t = S_i, q_{t+1} = S_l|x, \theta) = \alpha(t, i)p_{il}b_l(x^d_{t+1})\beta(t + 1, l)$$

$$P(q_2 = C, q_3 = R|x, \theta) = \alpha(2, C)p_{CR}b_R(I)\beta(3, R)$$



$$P(q_t = S_i, q_{t+1} = S_l|x, \theta) = \alpha(t, i)p_{il}b_l(x_{t+1}^d)\beta(t+1, l)$$

$$P(q_3 = C, q_4 = R|x, \theta) = \alpha(3, C)p_{CR}b_R(O)\beta(4, R)$$



$$P(q_t = S_i, q_{t+1} = S_l|x, \theta) = \alpha(t, i)p_{il}b_l(x_{t+1}^d)\beta(t+1, l)$$

$$P(q_1 = C, q_2 = R|x, \theta) = \alpha(1, C)p_{CR}b_R(I)\beta(2, R)$$

$$P(q_2 = C, q_3 = R|x, \theta) = \alpha(2, C)p_{CR}b_R(I)\beta(3, R)$$

$$P(q_3 = C, q_4 = R|x, \theta) = \alpha(3, C)p_{CR}b_R(O)\beta(4, R)$$

$$P(q_4 = C, q_5 = R|x, \theta) = \alpha(4, C)p_{CR}b_R(P)\beta(5, R)$$

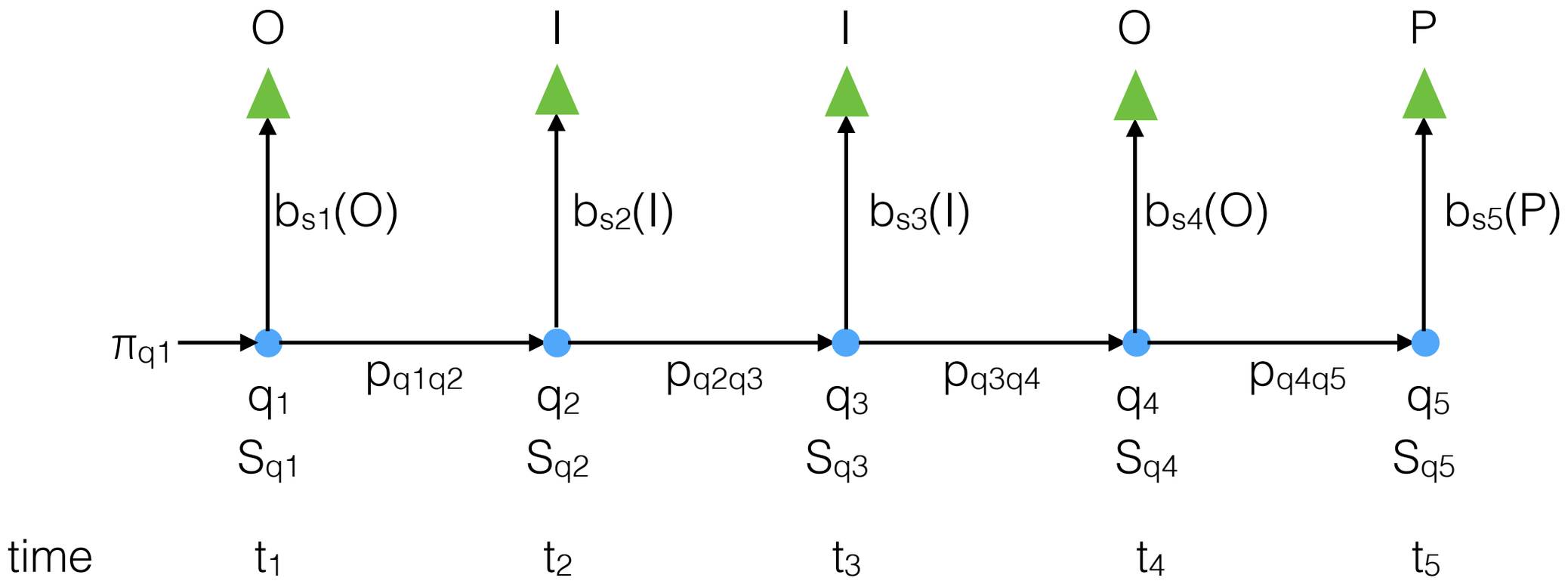
$$p'_{il} = \sum_d \frac{\sum_t \alpha^d(t, i)p_{il}b_l(x_{t+1}^d)\beta^d(t+1, l)}{P(x^d)}$$

$$\alpha(t,i) = p(x_1 x_2 \dots x_t, q_t = S_i)$$

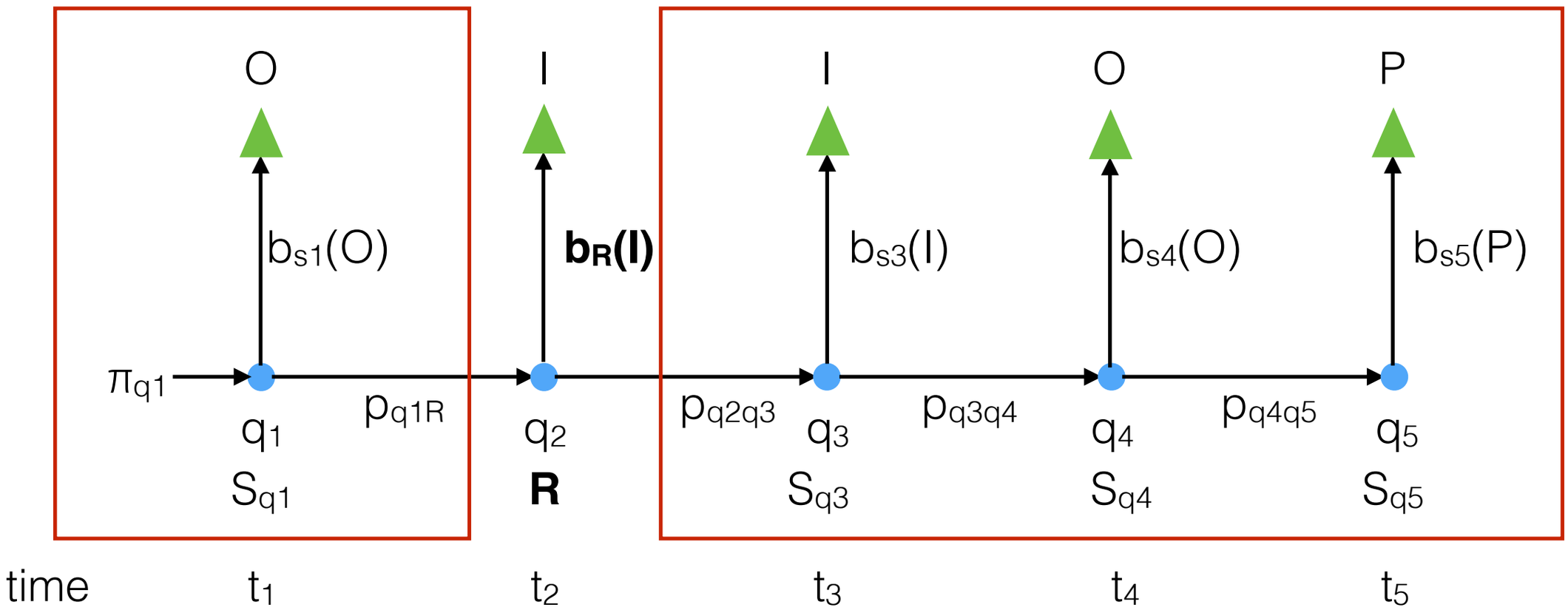
$$\beta(t,i) = p(x_T x_{T-1} \dots x_t | q_t = S_i)$$

$$p(x, q_t = S_i | M) = \alpha(t,i)\beta(t,i)$$

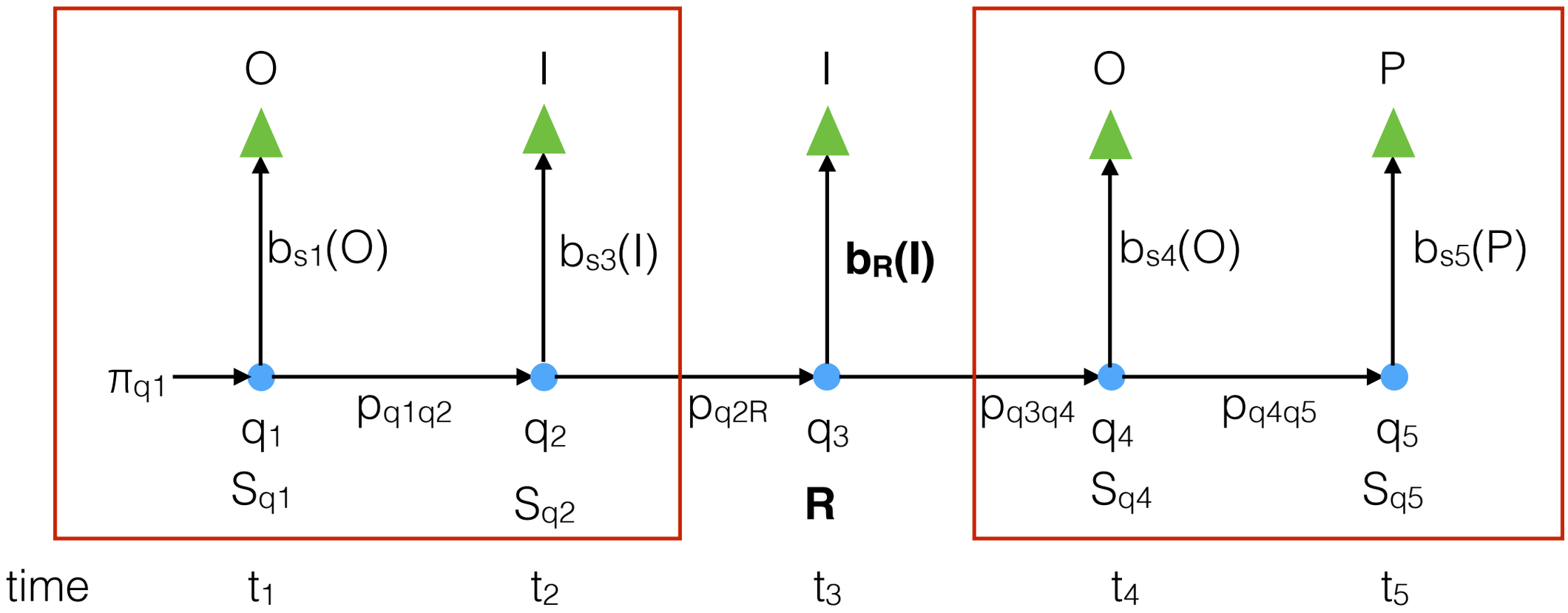
	I	O	P
S	$b_S(I)$	$b_S(O)$	$b_S(P)$
R	$b_R(I)$	$b_R(O)$	$b_R(P)$
C	$b_C(I)$	$b_C(O)$	$b_C(P)$



$$P(x, q_2 = R|M) = \alpha(2, R)\beta(2, R)$$



$$P(x, q_3 = R|M) = \alpha(3, R)\beta(3, R)$$



$$P(x, q_2 = R|M) = \alpha(2, R)\beta(2, R)$$

$$P(x, q_3 = R|M) = \alpha(3, R)\beta(3, R)$$

$$b'_l(a) = \sum_d \frac{\sum_{t|x_t^d=a} \alpha^d(t, l)\beta^d(t, l)}{P(x^d)}$$

Applications of HMMs

note - most of these are implemented as Viterbi (decoding) questions

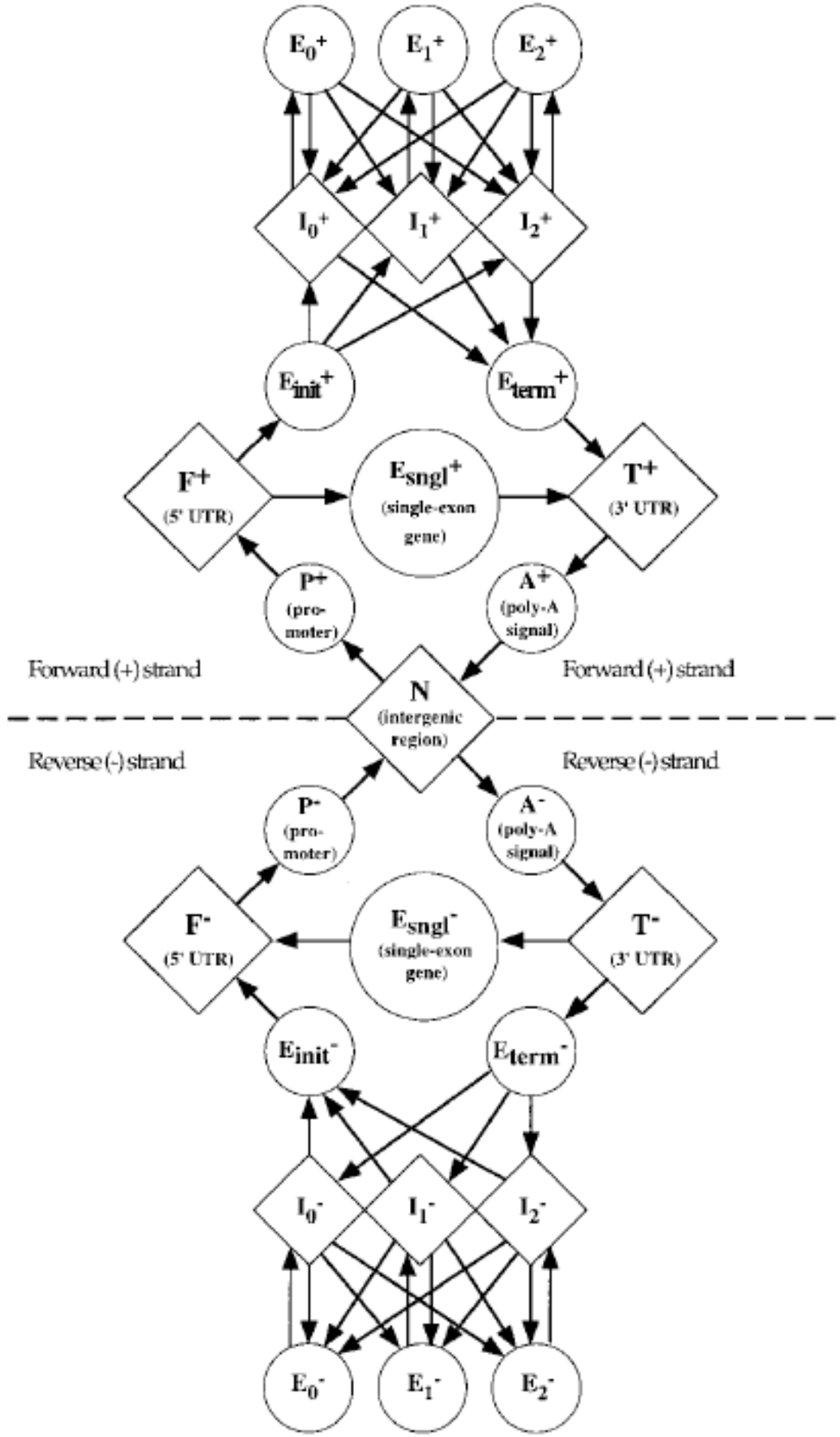
- Exon finding through orthology (Haussler)
- ECG signal analysis (beat segmentation and classification)
- Analysis of microarray data especially tiling arrays
- Sequence feature prediction using homology information
- Sequence alignments, pairwise and multiple
- Analyzing ChIP-chip on tiling arrays

Finding genes

- The first gene finders were for prokaryotes
  - No introns
  - Distinct and known signals
- GLIMMER (1998, Salzberg et al.) was an early gene-finding program and was very successful
  - Only for prokaryotes (first version)
  - Tested on relatively short sequences

GENSCAN (1997)

- GENSCAN (Burge and Karlin) was a huge breakthrough in eukaryotic gene-finding, and is still used
- How is it different?
  - Assumes that the input sequence can have no genes, one gene, multiple genes, or parts of genes
  - Models all known aspects of a eukaryotic gene
  - Uses general 3-periodic inhomogeneous fifth-order Markov model of coding regions
  - Does not use specific models of protein structure or database homology


E0+
E1+
E2+
I0+
I1+
I2+
Einit+
Eterm+
F+
(5' UTR)
Esngl+
(single-exon
gene)
T+
(3' UTR)
P+
(pro-
moter)
A+
(poly-A
signal)
Forward (+) strand
Forward (+) strand
N
(intergenic
region)
Reverse (-) strand
Reverse (-) strand
P-
(pro-
moter)
A-
(poly-A
signal)
F-
(5' UTR)
Esngl-
(single-exon
gene)
T-
(3' UTR)
Einit-
Eterm-
I0-
I1-
I2-
E0-
E1-
E2-

GENSCAN — MDD

- Maximal Dependence Decomposition
  - Need aligned set of several hundred signal sequences
  - Use conditional probabilities to capture the most significant dependencies between positions
  - Calculate χ^2 for each pair of positions to detect dependencies

Next generation

Three types of de novo predictors

- Single genome sequence (mostly HMMs)
- Two aligned genomes
- Multiple aligned genomes

} infer local rates & patterns of mutation

With good programs can expect 50-70% of the genes correctly predicted, in a compact genome

Next generation

Dual-genome predictors

- Assume that functional regions are more conserved
- SLAM (HMM) - uses joint probability for sequence alignment and gene structure to define types of alignments seen in coding vs noncoding sequence
- More powerful approaches use HMM and dynamic programming
- Problem: in closely related species most of the conserved sequences are noncoding

Next generation

Multi-genome predictors

- More genomes -> stronger evidence
- Hard to get enough species for a good alignment (translocations, deletions, inversions etc destroy alignments)
- Some use phylogenetic trees (phylo-HMMs)

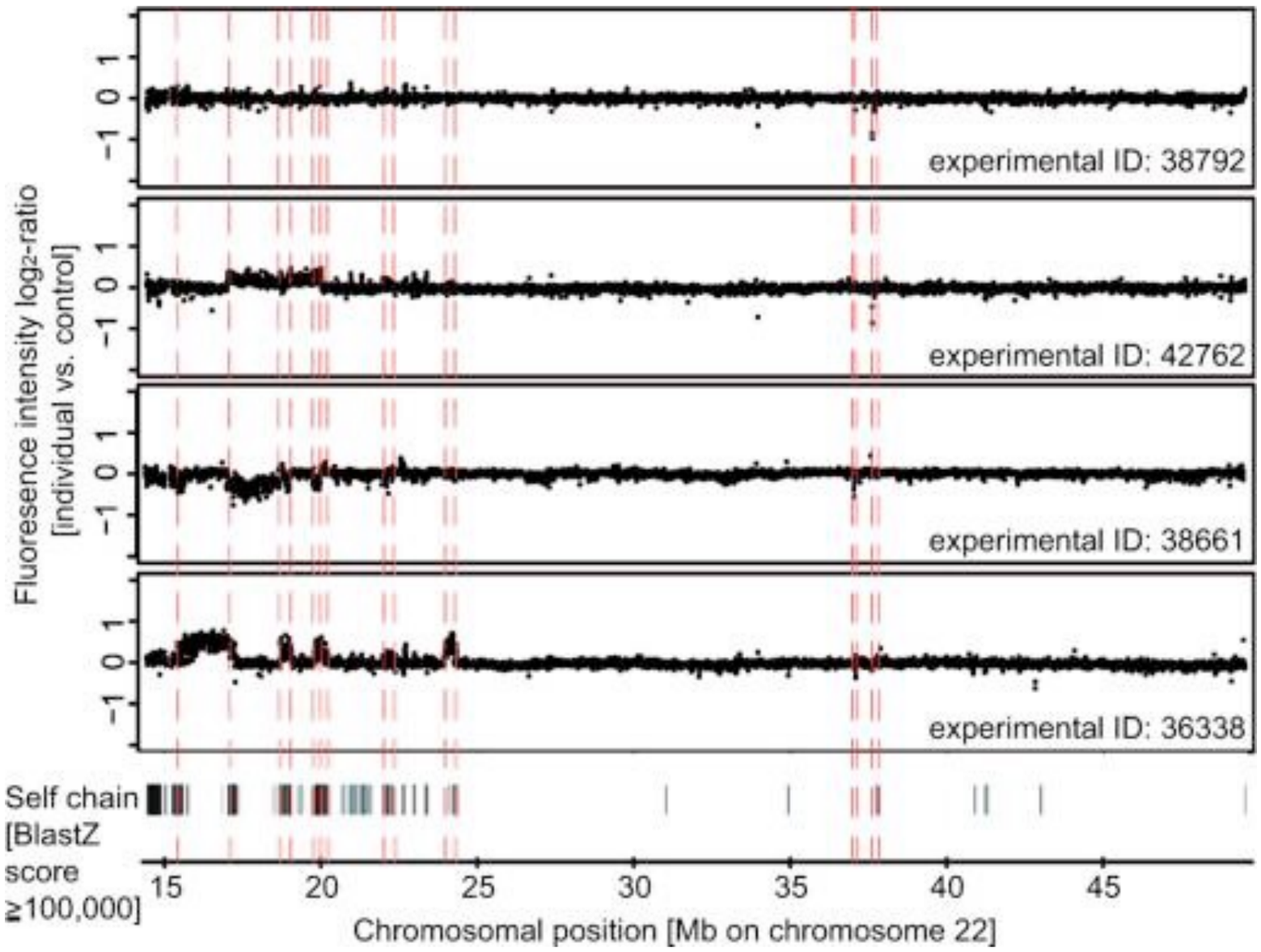
HMM for copy number variation

Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome

Jan O. Korbel*†‡, Alexander Eckehart Urban§¶, Fabian Grubert§, Jiang Du‖, Thomas E. Royce*, Peter Starr*, Guoneng Zhong*, Beverly S. Emanuel, Sherman M. Weissman§, Michael Snyder¶‡, and Mark B. Gerstein*‖‡**

Departments of *Molecular Biophysics and Biochemistry and §Genetics, Yale University School of Medicine, New Haven, CT 06520; †European Molecular Biology Laboratory, 69117 Heidelberg, Germany; Departments of ¶Molecular, Cellular, and Developmental Biology and ‖Computer Science, Yale University, New Haven, CT 06520; and **Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104

HMM for CNV


Fluorescence intensity log2-ratio [individual vs. control]
−1 0 1
experimental ID: 38792
experimental ID: 42762
experimental ID: 38661
experimental ID: 36338
Self chain [BlastZ score ≥100,000]
15
20
25
30
35
40
45
Chromosomal position [Mb on chromosome 22]

detecting pieces of immunoglobulin rearrangements

- infections, neoplasms can stimulate B cell development and antibody production

- antibody production & diversification involves rearranging V(D)J segments of genes

- given an immunoglobulin, what V,D,J segments did it come from?

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 7 2010, pages 867–872
doi:10.1093/bioinformatics/btq056

Sequence analysis

Advance Access publication February 9, 2010

SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements

Supriya Munshaw[1,2] and Thomas B. Kepler[1,3,*]

[1]Center for Computational Immunology, [2]Computational Biology and Bioinformatics Program, Duke University, P.O. Box 90090 and [3]Department of Biostatistics and Bioinformatics, 2424 Erwin Road, Suite 1103, Durham, NC 27705, USA

Associate Editor: Limsoon Wong

CpG islands

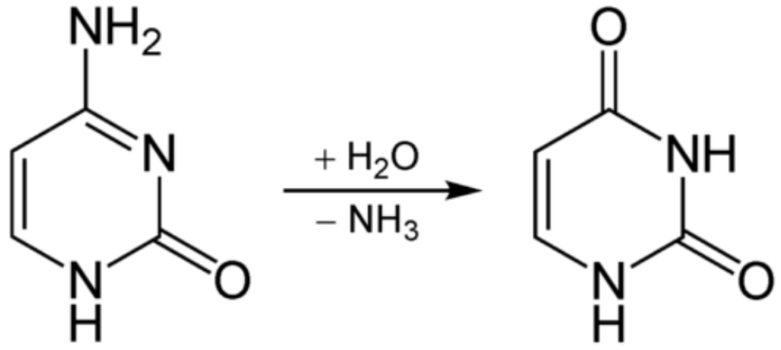
the CG dinucleotide is extraordinarily underrepresented in vertebrate genomes (about 1/5 the expected frequency)

remaining CG dinucleotides cluster in “islands”

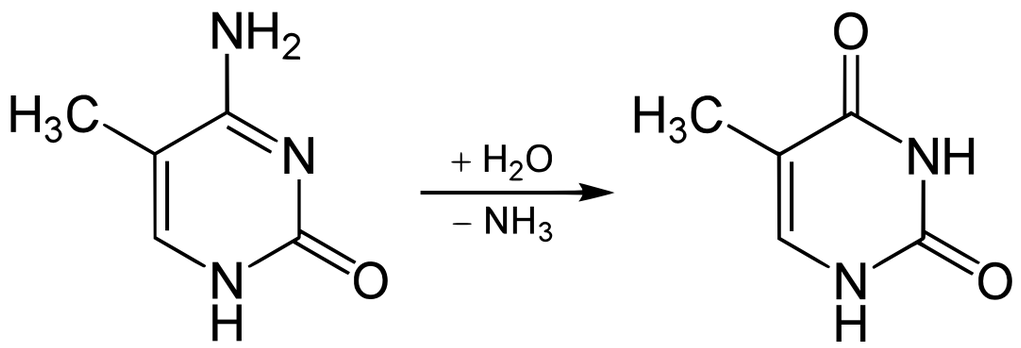
highly regulatory regions in eukaryotic genomes

overall human genome C+G content is ~42%; CpG island is ~65%

fate of cytosines in CpG context



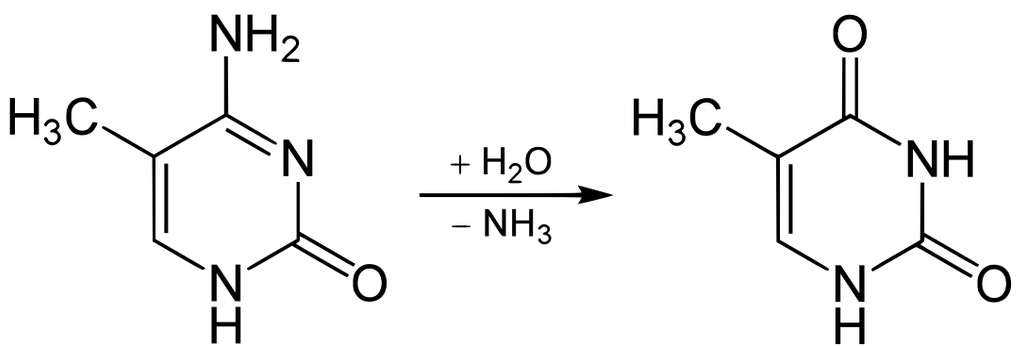
cytosine deamination to uracil, recognized and repaired



methylated cytosine deamination to thymine, not recognized

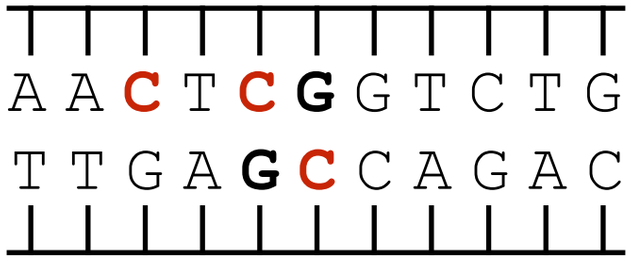
By Yikrazuul (talk) - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=12910574

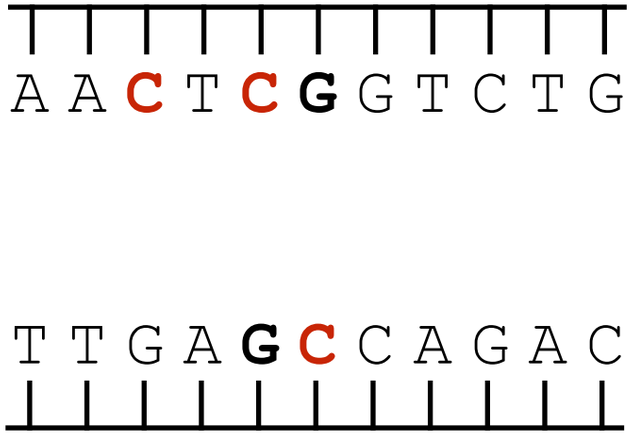
fate of cytosines in CpG context

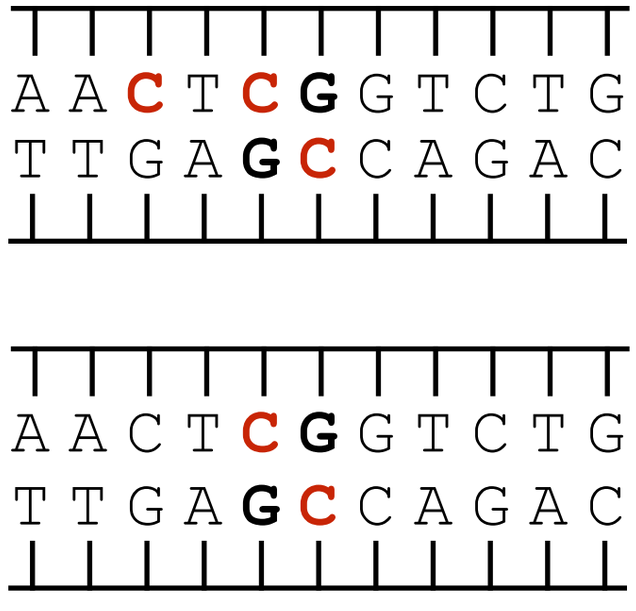


methylated cytosine deamination to thymine, not recognized, persists as a mutation

this is a problem for a cell, as methylated cytosines are an important epigenetic mark!





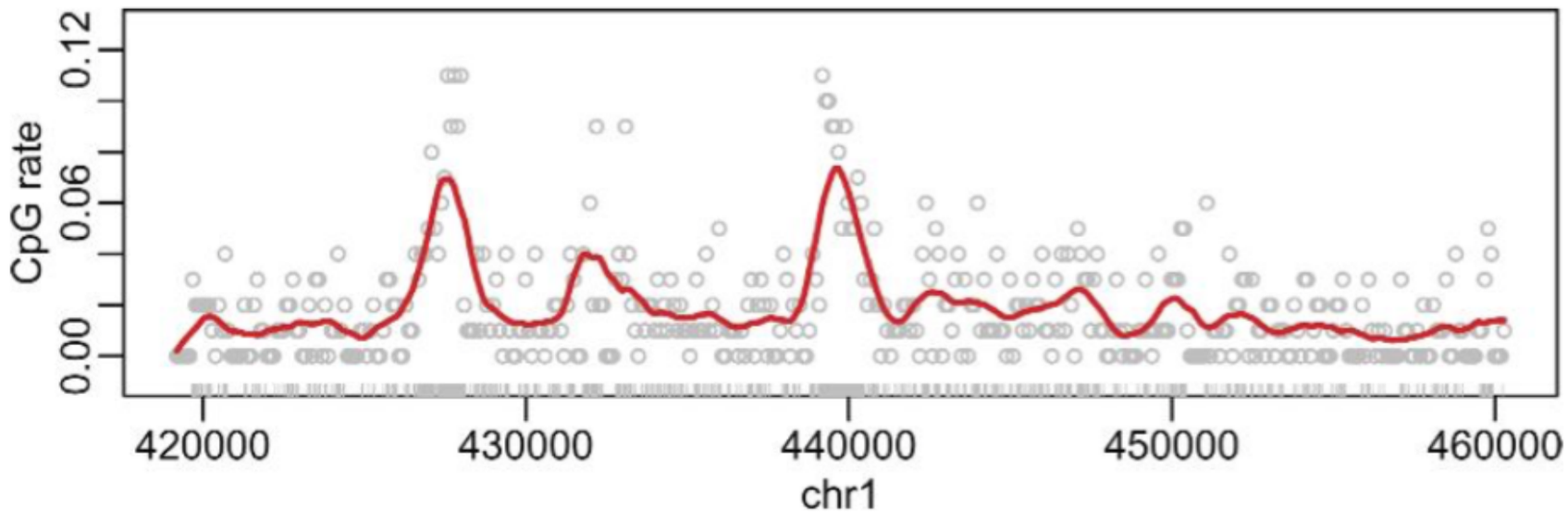


CpG islands

200+ nucleotides long

G+C content > 50%

observed/expected CpG ratio > 0.6



Biostatistics (2010), **11**, 3, *pp*. 499–514
doi:10.1093/biostatistics/kxq005
Advance Access publication on March 8, 2010

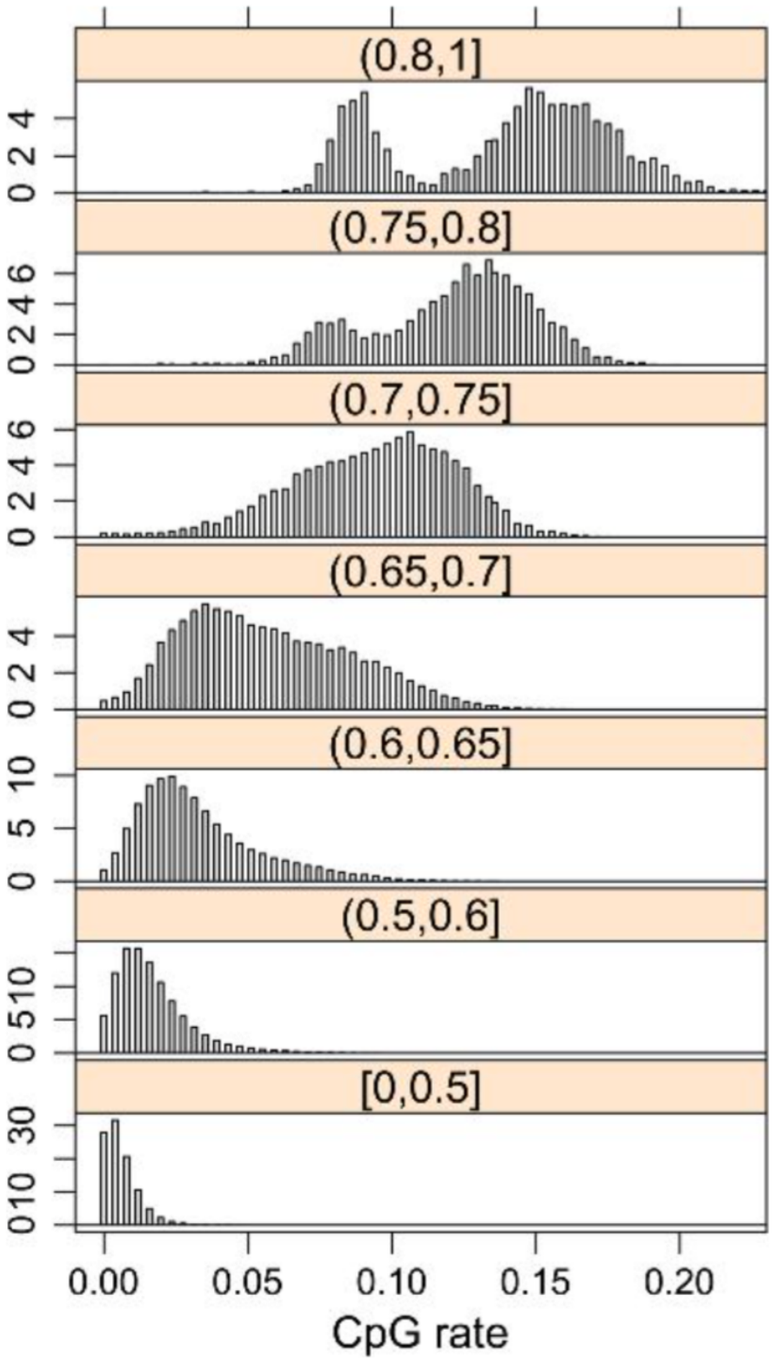
Redefining CpG islands using hidden Markov models

HAO WU, BRIAN CAFFO, HARRIS A. JAFFEE, RAFAEL A. IRIZARRY*

Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA
ririzarr@jhsph.edu

ANDREW P. FEINBERG

Department of Medicine and Center for Epigenetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA



at high GC content, there are two populations of regions, by CpG rate

Fitting a 2-state HMM allows segmentation of DNA sequence into CpG islands and non-CpG islands

Fig. 4. Histogram of CpG rates in nonoverlapping genomic segments of length 256 bases, stratifed by GC content