7.91 / 7.36 / BE.490

Lecture #5

Mar. 9, 2004

Markov Models
&
DNA Sequence Evolution

Chris Burge

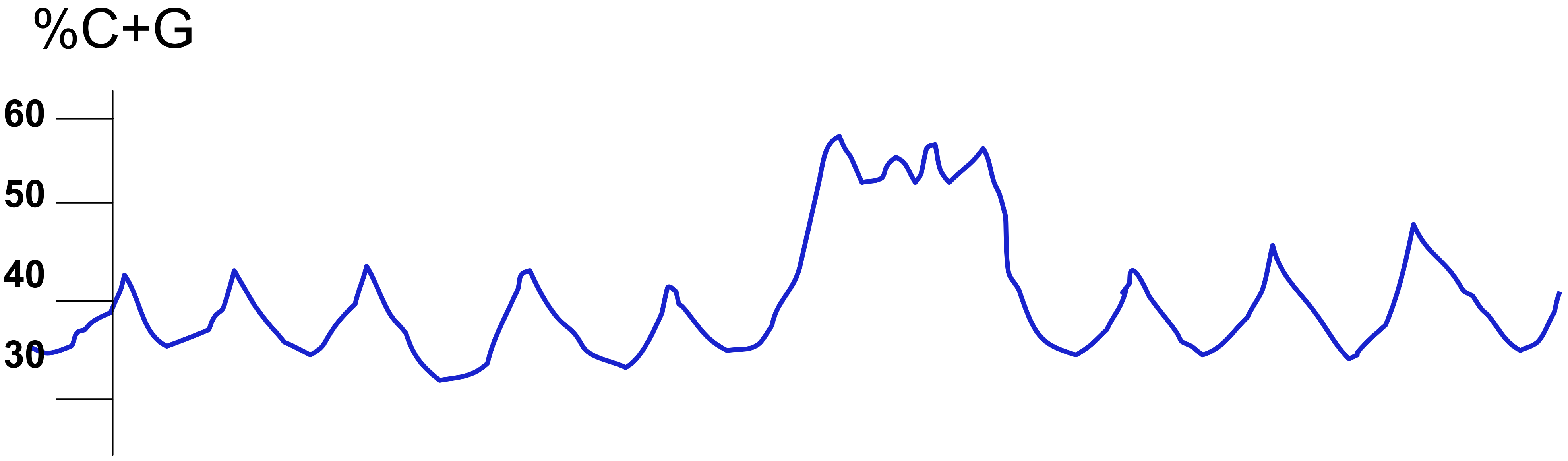
Review of Markov & HMM Models for DNA

- Markov Models for splice sites
- Hidden Markov Models

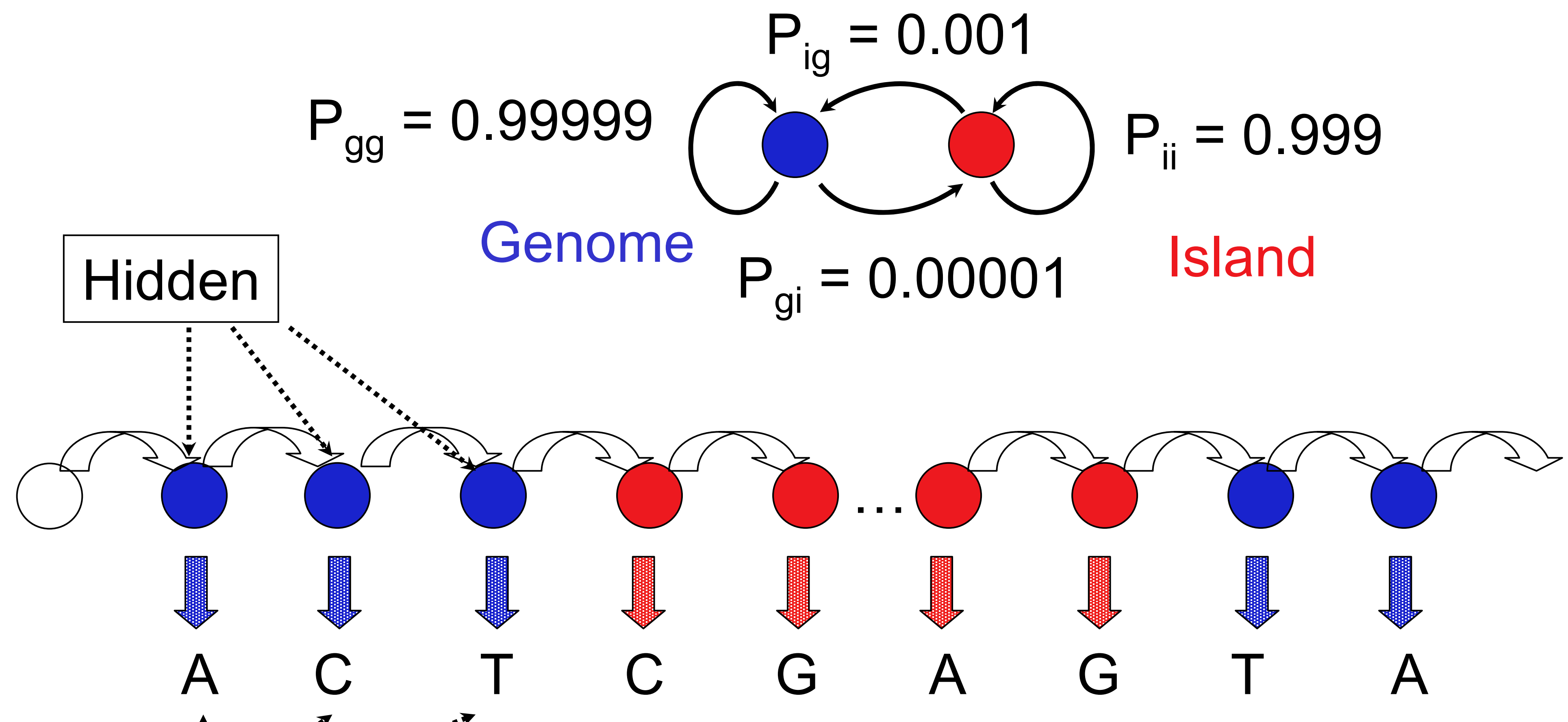
  - looking under the hood
- The Viterbi Algorithm
- Real World HMMs

Ch. 4 of Mount

CpG Islands



CpG Island Hidden Markov Model



$P_{ig} = 0.001$

$P_{gg} = 0.99999$

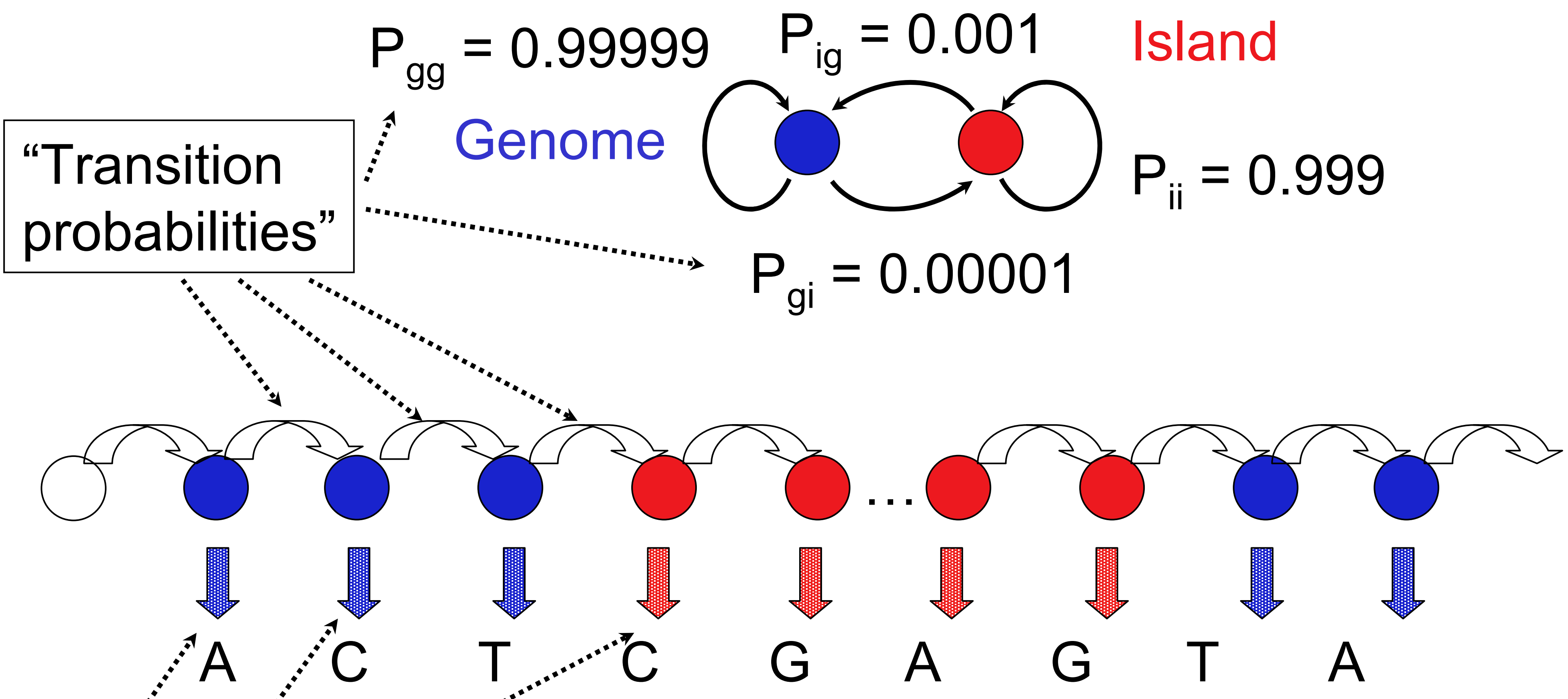
$P_{ii} = 0.999$

$P_{gi} = 0.00001$

Observable

	C	G	A	T
CpG Island:	0.3	0.3	0.2	0.2
Genome:	0.2	0.2	0.3	0.3

CpG Island HMM II



$P_{gg} = 0.99999$

$P_{ig} = 0.001$

Island

Genome

“Transition probabilities”

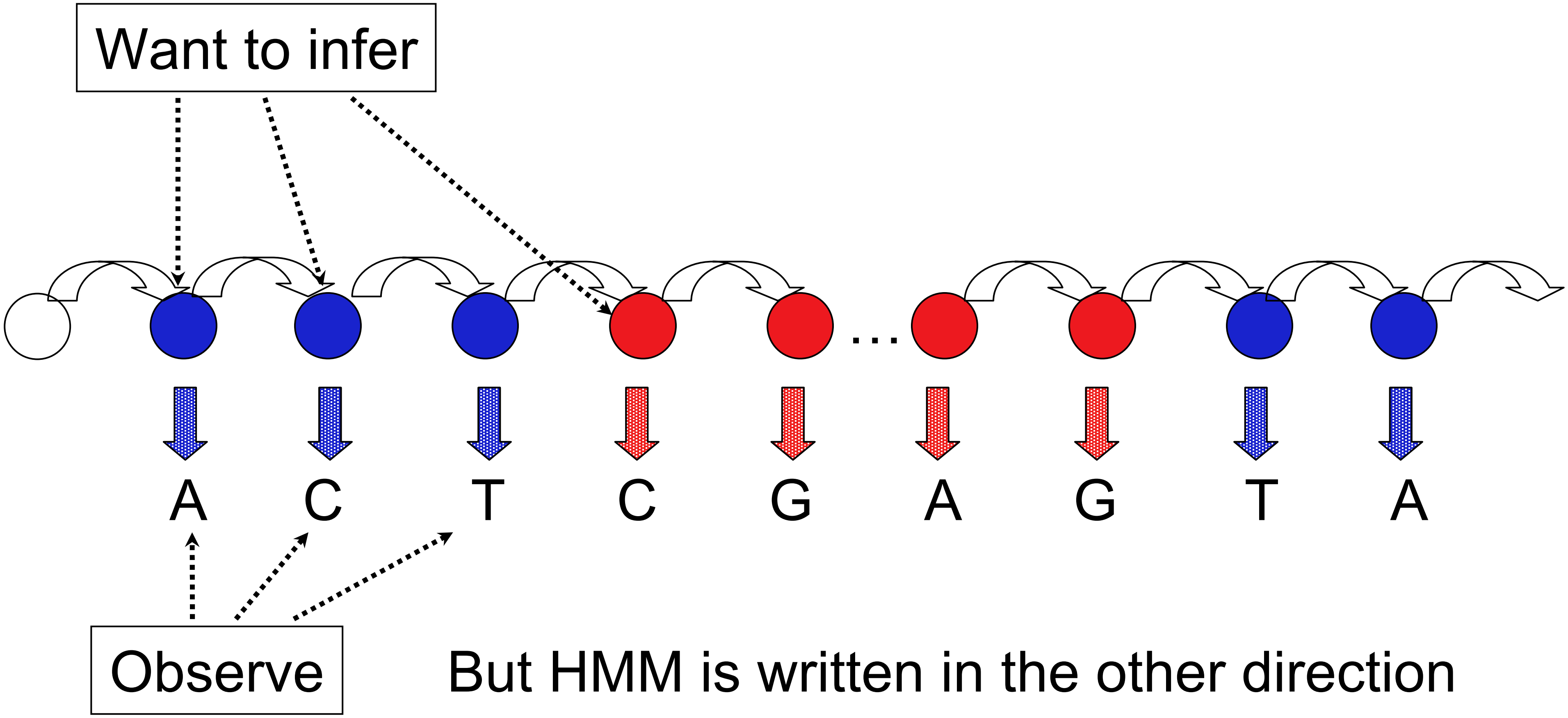
$P_{ii} = 0.999$

$P_{gi} = 0.00001$

“Emission Probabilities”

	C	G	A	T
CpG Island:	0.3	0.3	0.2	0.2
Genome:	0.2	0.2	0.3	0.3

CpG Island HMM III



But HMM is written in the other direction
(observable depends on hidden)

Inferring the Hidden from the Observable (Bayes' Rule)

Conditional Prob:
P(A|B) = P(A,B)/P(B)

$$P(H = h_1, h_2, \ldots, h_n \mid O = o_1, o_2, \ldots, o_n)$$

$$= \frac{P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n)}{P(O = o_1, \ldots, o_n)}$$

$$= \frac{P(H = h_1, \ldots, h_n) P(O = o_1, \ldots, o_n \mid H = h_1, \ldots, h_n)}{P(O = o_1, \ldots, o_n)}$$

$P(O = o_1, \ldots, o_n)$ somewhat difficult to calculate

But notice:

$$P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n) > P(H = h'_1, \ldots, h'_n, O = o_1, \ldots, o_n)$$

implies $P(H = h_1, \ldots, h_n \mid O = o_1, \ldots, o_n) > P(H = h'_1, \ldots, h'_n \mid O = o_1, \ldots, o_n)$

so can treat $P(O = o_1, \ldots, o_n)$ as a constant

Finding the Optimal "Parse" (Viterbi Algorithm)

Want to find sequence of hidden states $H^{opt} = h_1^{opt}, h_2^{opt}, h_3^{opt}, \ldots$

which maximizes joint probability: $P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n)$

(optimal "parse" of sequence)

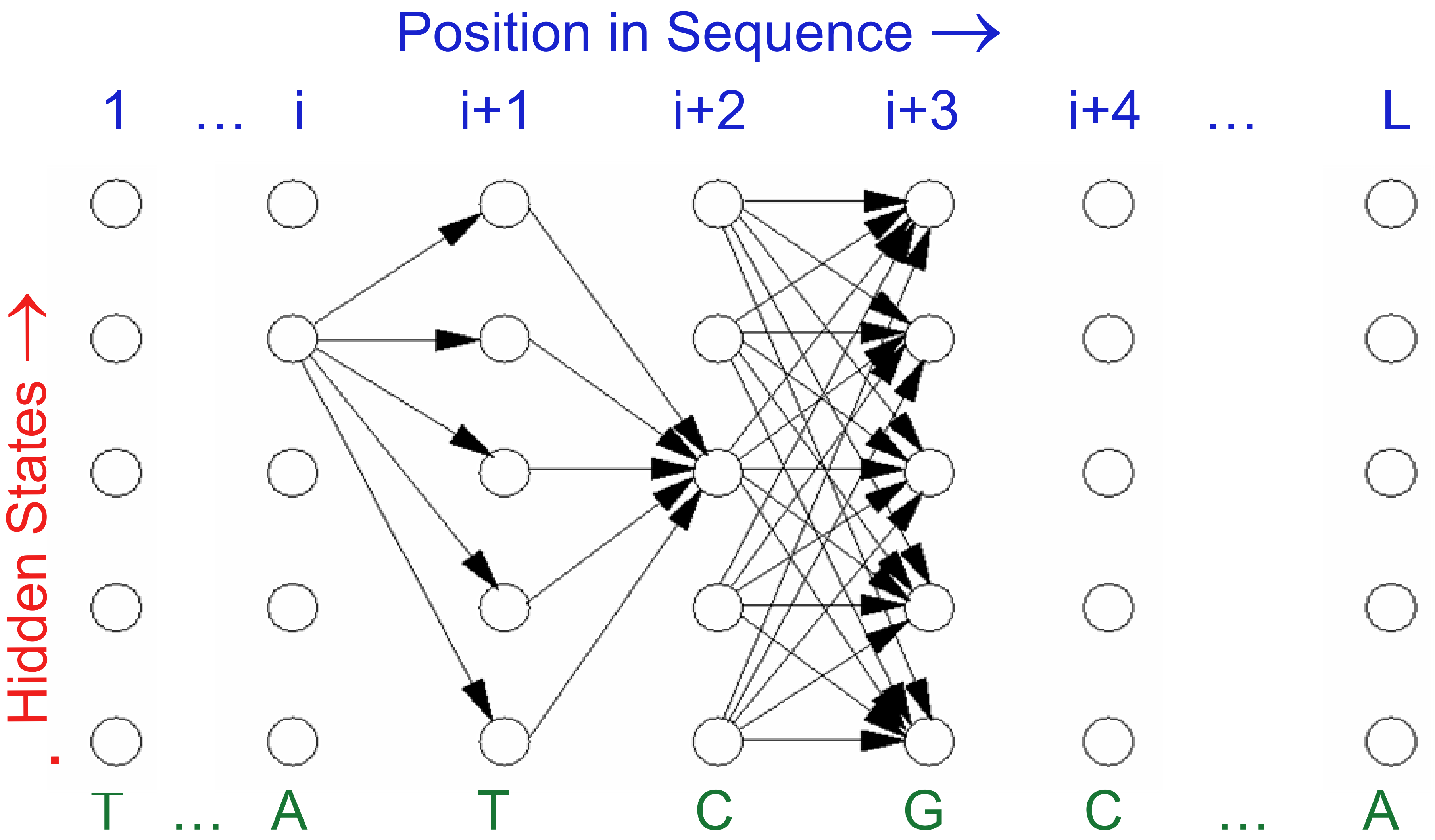
Solution:

Define $R_i^{(h)}$= probability of optimal parse of the subsequence 1..i ending in state h

Solve recursively, i.e. determine $R_2^{(h)}$ in terms of $R_1^{(h)}$, etc.

A. Viterbi, an MIT BS/MEng student in E.E. - founder of Qualcomm

“Trellis” Diagram for Viterbi Algorithm



Run time for k-state HMM on sequence of length L?

Viterbi Algorithm Examples

What is the optimal parse of the sequence:

- $(ACGT)_{10000}$

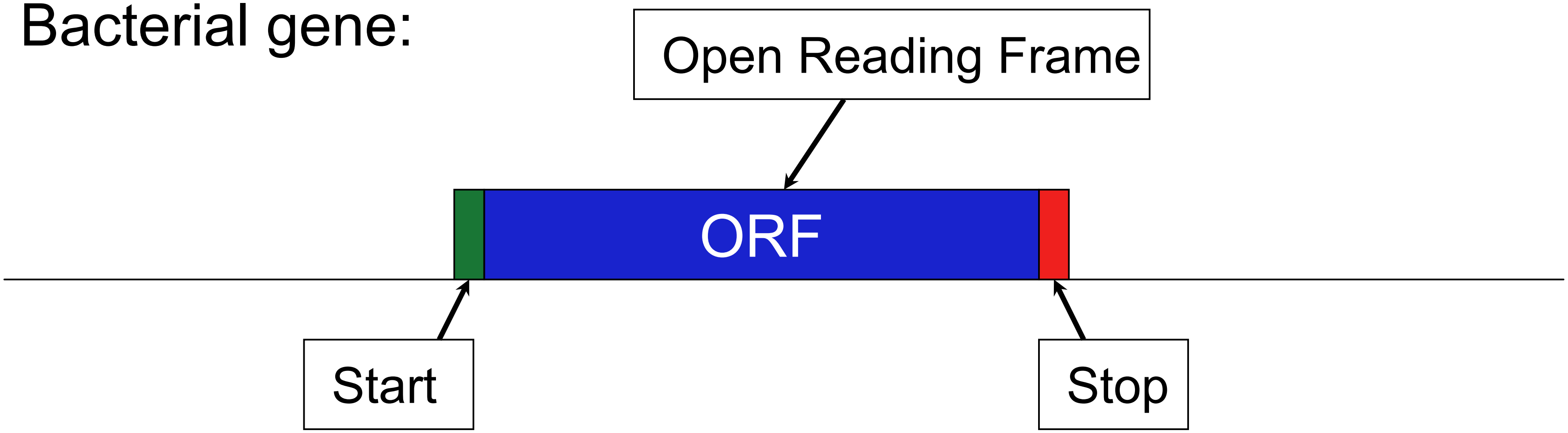
- $A_{1000}C_{80}T_{1000}C_{40}A_{1000}G_{60}T_{1000}$

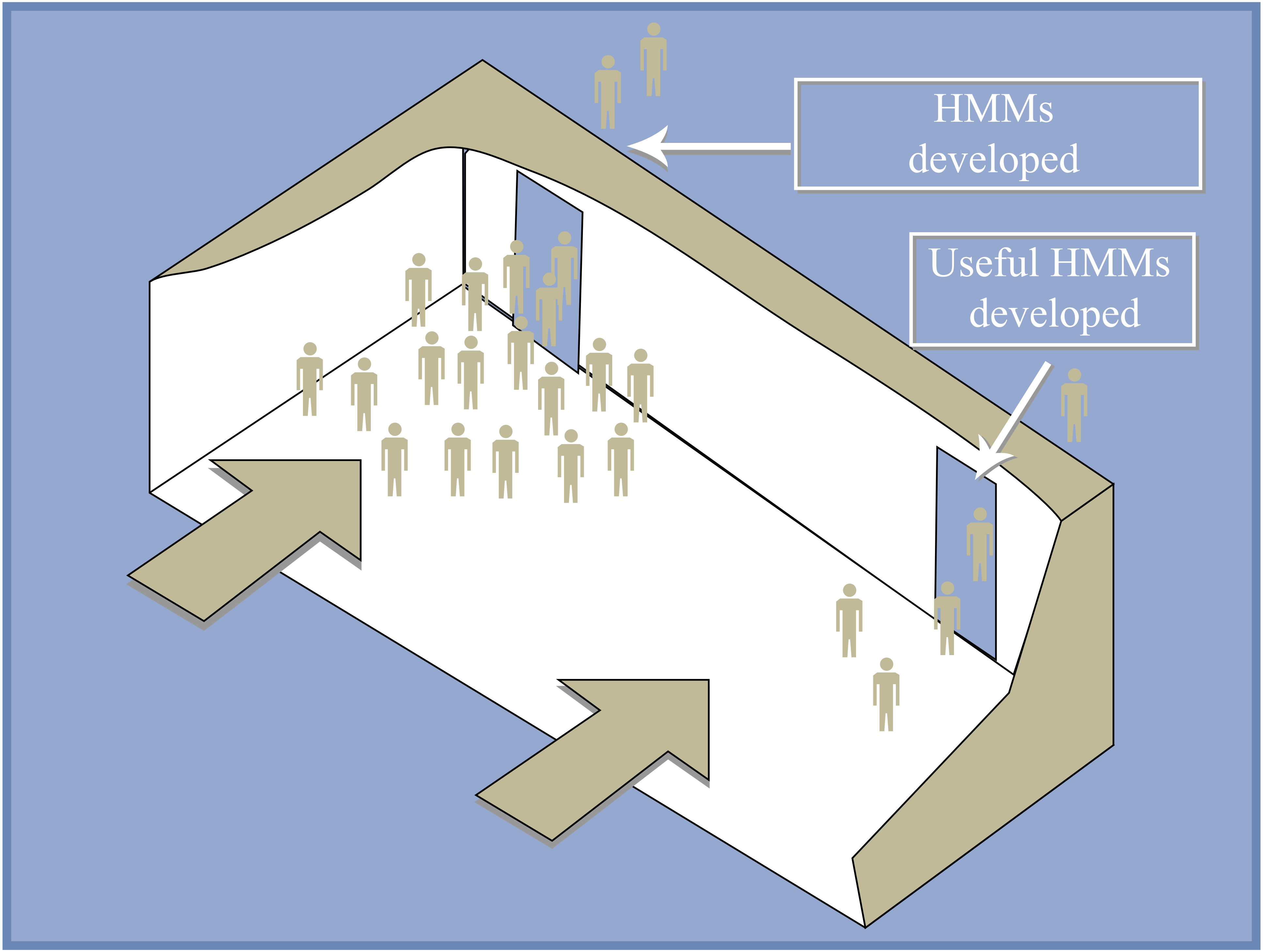
Powers of 1.5:

N =	20	40	60	80
$(1.5)^N$ =	$3x10^3$	$1x10^7$	$3x10^{10}$	$1x10^{14}$

What else can you model with HMMs?

Bacterial gene:





HMMs developed

Useful HMMs developed

Parameter Estimation for HMMs

How many parameters for a k-state HMM over an alphabet of size 4?

Initial probabilities:

Transition probabilities:

Emission probabilities:

Pseudocounts

Courtesy of M. Yaffe

•If the number of sequences in the training set is both large and diverse, then the sequences in the training set represent a good statistical sampling of the motif….*if not, then we have a sampling error!*

Correct for this by adding pseudocounts. How many to add?

→ ***Too many pseudocounts dominate the frequencies… and the resulting matrix won't work!***

→ ***Too few pseudocounts then we'll miss many amino acid variations, and matrix will only find sequences that produced the motif!***

Add few pseudocounts if sampling is good (robust), and add more pseudocounts if sampling is sparse

One reasonable approach is to add √N pseudocounts, where N is the number of sequences…

As N increases, the influence of pseusocounts decreases since N increases faster than √N, but doesn't add enough at low N

Dealing With Small Training Sets

Position:	1	2	3	4	5
A	8				
C	1				
G	1				
T	0				

If the true frequency of T at pos. 1 was 10%, what's the probability we wouldn't see any Ts in a sample of 10 seqs?

$P(N=0) = (10!/0!10!)(0.1)^0(0.9)^{10} = \sim 35\%$

So we should add pseudocounts

Training Set

```
ACCTG
AGCTG
ACCCG
ACCTG
ACCCA
GACTG
ACGTA
ACCTG
CCCCG
ACATC
```

Pseudocounts (Ψ_{counts})

Nt	Count	Ψ_{count}	$\text{Bayes}_{\text{count}}$	ML est.	Bayes est.
A	8	+ 1	9	0.80	0.64
C	1	+ 1	2	0.10	0.14
G	1	+ 1	2	0.10	0.14
T	0	+ 1	1	0.00	0.07
	10		14	1.0	1.0

The ‘add 1 to each observed count’ rule can be derived analytically from the Bayesian posterior distribution under a Dirichlet prior - see Appendix A of statistics primer for details.

Real World HMMs

Please see the following Web site: http://www.cbs.dtu.dk/services/TMHMM/

Reference for TMHMM: Krogh, A, B Larsson, G von Heijne, and EL Sonnhammer. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes." *J Mol Biol.* 305, no. 3 (19 January 2001): 567-80.

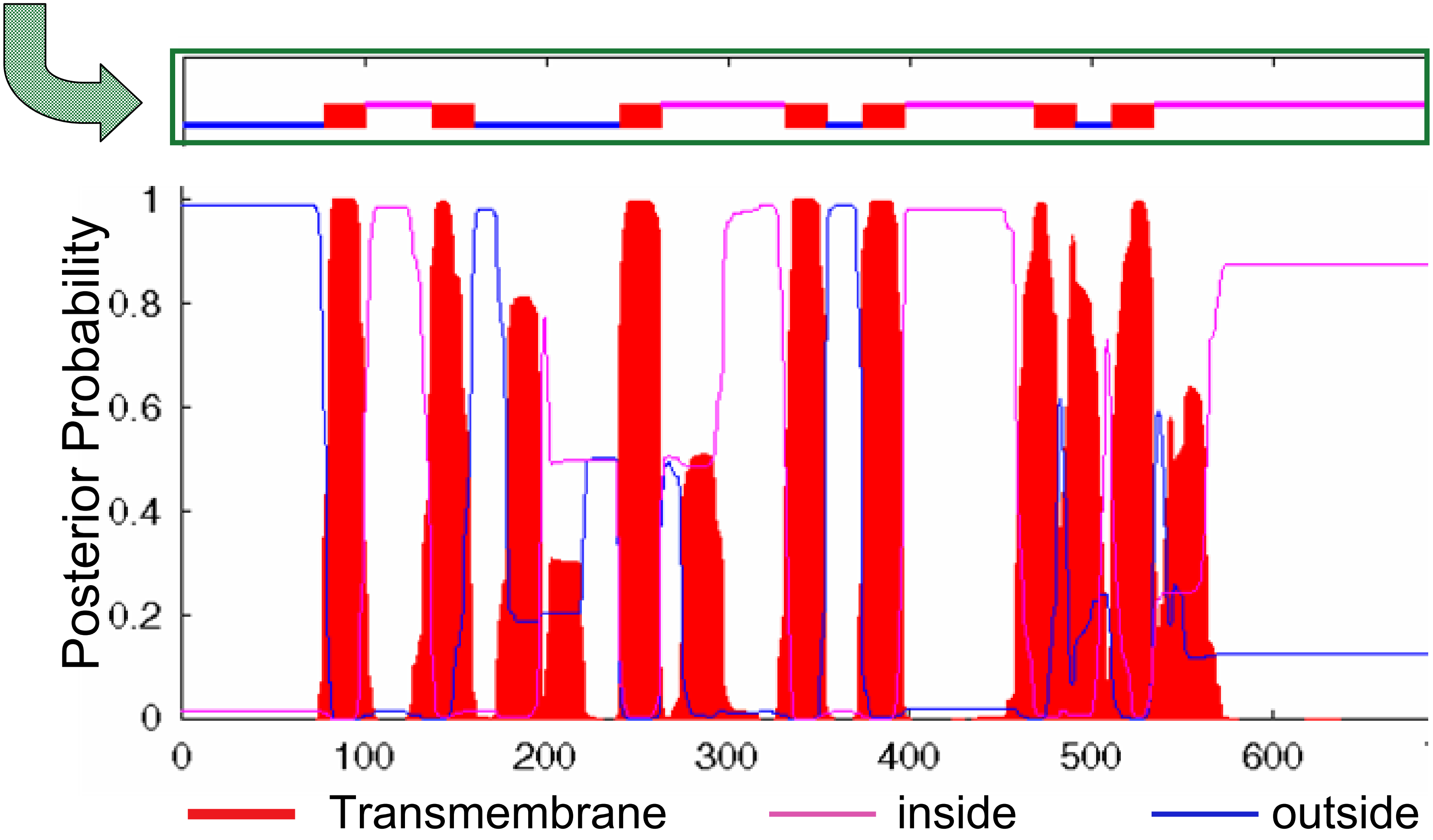
Architecture of TMHMM

Please see figures 1a and 1c of:

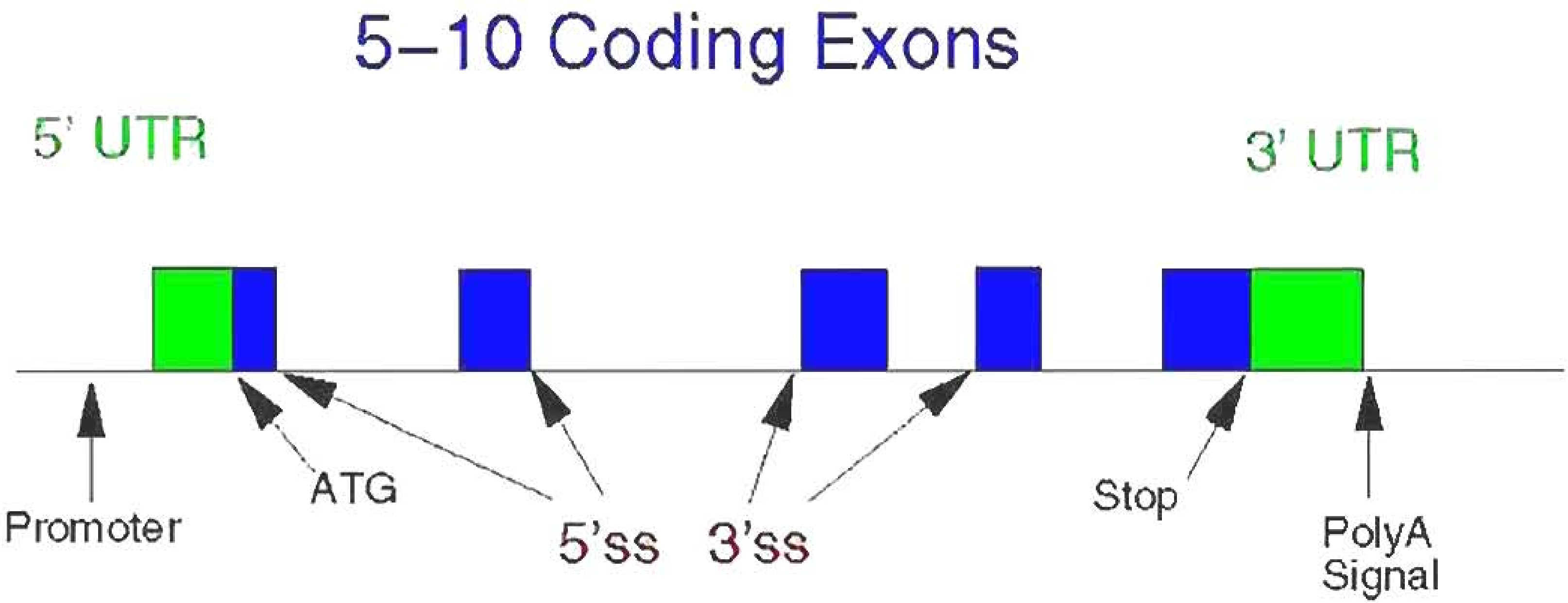
Krogh, A, B Larsson, G von Heijne, and EL Sonnhammer. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes." *J Mol Biol.* 305, no. 3 (19 January 2001): 567-80.

TMHMM Output for Mouse Chloride Channel CLC6

Optimal Parse



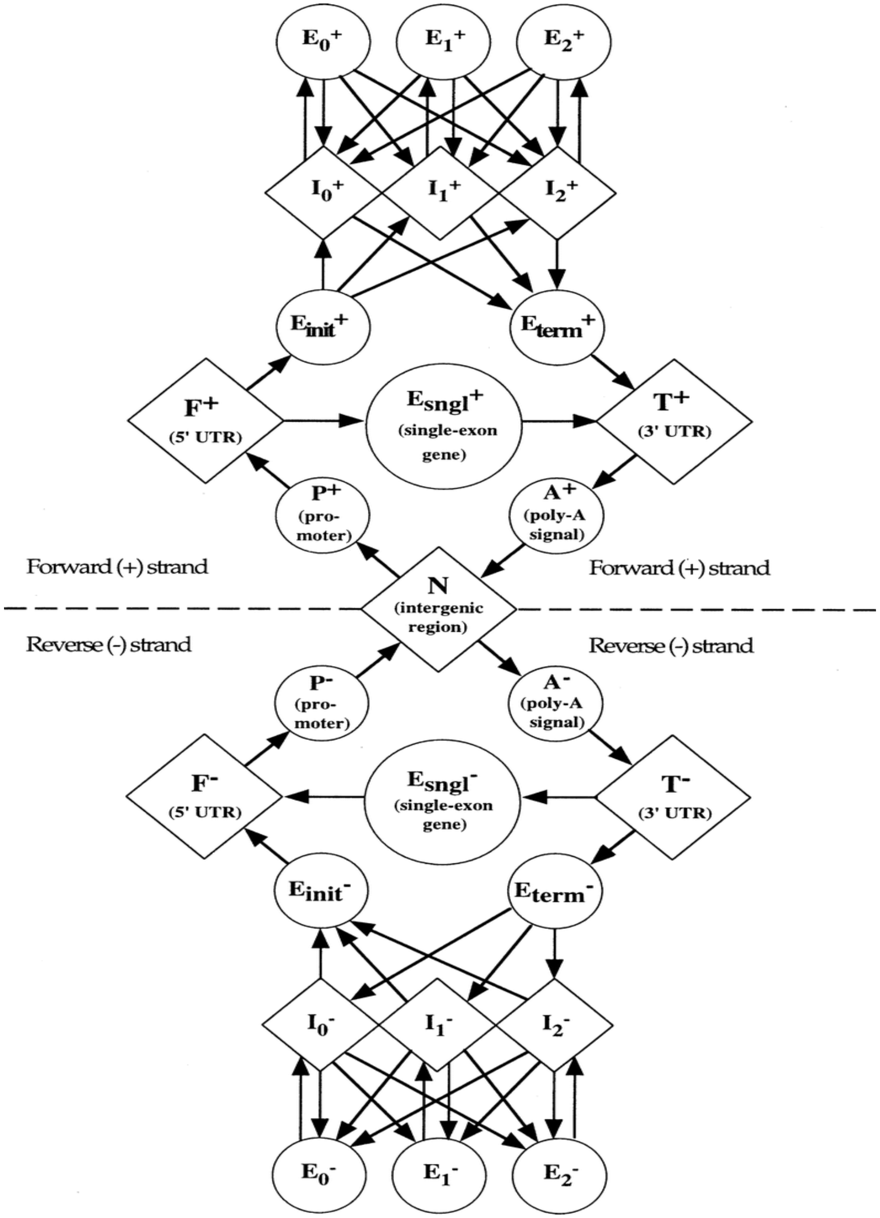
Structure of a Typical Human Gene



Genscan Model

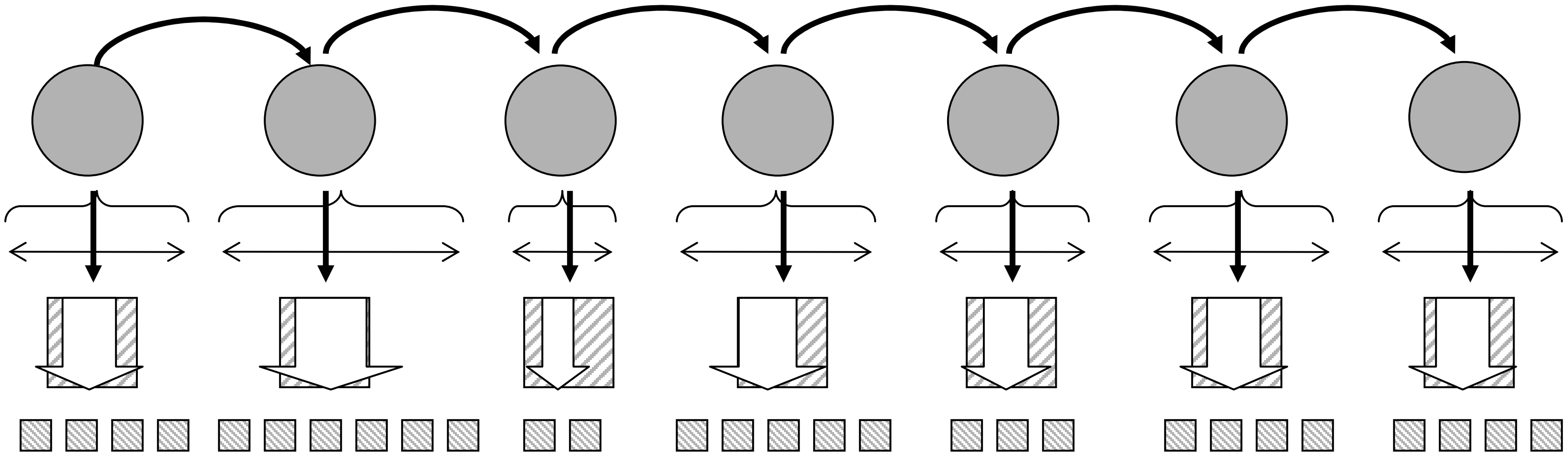
Incorporates:

Transcriptional signals
Splicing signals
Translational signals
Composition of exons
Composition of introns
Other gene features



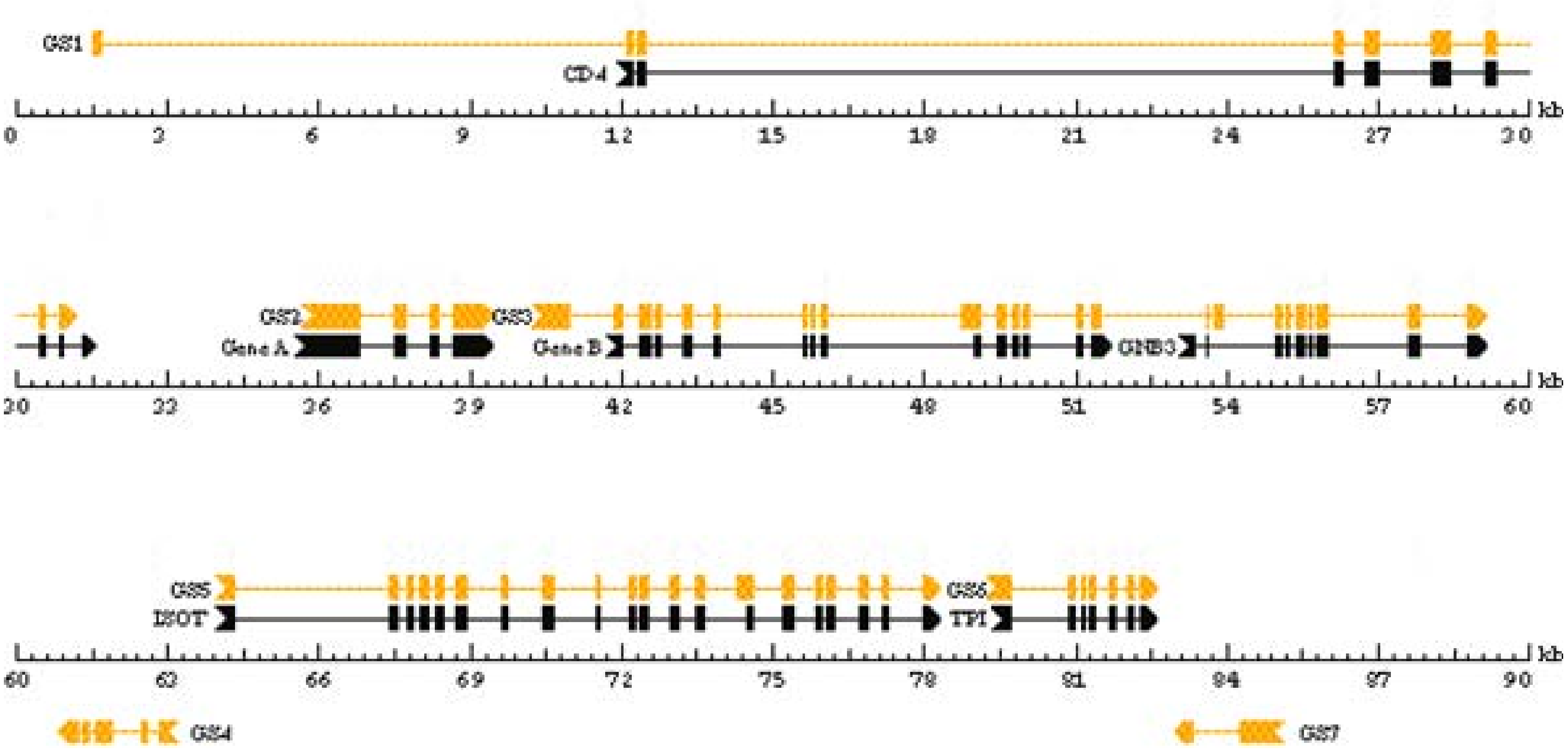
Burge & Karlin, J Mol Biol 1997

Semi-Markov HMM Model



Genscan predictions in human CD4 gene region

■ Annotated exons ▨ Genscan predicted exons



Overall: ~75% of exons exactly correct

Burge and Karlin *J. Mol. Biol. 1997*

Genscan, GenomeScan Predictions in Human BRCA1 Region

Please see figures 1 of

Yeh, RF, LP Lim, and CB Burge. "Computational Inference of Homologous Gene Structures in the Human Genome." *Genome Res.* 11, no. 5 (May 2001): 803-16.

DNA Sequence Evolution

Generation *n-1* (grandparent)

```
5’ TGGCATGCACCCTGTAAGTCAATATAAATGGCTACGCCTAGCCCATGCGA 3’
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3’ ACCGTACGTGGGACATTCAGTTATATTTACCGATGCGGATCGGGTACGCT 5’
```

Generation *n* (parent)

```
5’ TGGCATGCACCCTGTAAGTCAATATAAATGGCTATGCCTAGCCCATGCGA 3’
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3’ ACCGTACGTGGGACATTCAGTTATATTTACCGATACGGATCGGGTACGCT 5’
```

Generation *n+1* (child)

```
5’ TGGCATGCACCCTGTAAGTCAATATAAATGGCTATGCCTAGCCCGTGCGA 3’
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3’ ACCGTACGTGGGACATTCAGTTATATTTACCGATACGGATCGGGCACGCT 5’
```

What is a *Markov* Model (aka *Markov* Chain)?

Classical Definition

A discrete stochastic process $X_1, X_2, X_3, \ldots$
which has the Markov property:

$$P(X_{n+1} = j \mid X_1=x_1, X_2=x_2, \ldots X_n=x_n) = P(X_{n+1} = j \mid X_n=x_n)$$

(for all x_i, all j, all n)

In words:

A random process which has the property that the future (next state) is conditionally independent of the past given the present (current state)

Markov - a Russian mathematician, ca. 1922

DNA Sequence Evolution is a Markov Process

No selection case

S_n = base at generation *n*

$$P = \begin{pmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{pmatrix}$$

$$P_{ij} = P(S_{n+1} = j \,|\, S_n = i)$$

$\vec{q}^{\,n} = (q_A, q_C, q_G, q_T)$ = vector of prob's of bases at gen. *n*

Handy relations: $\vec{q}^{\,n+1} = \vec{q}^{\,n} P$ $\quad \vec{q}^{\,n+k} = \vec{q}^{\,n} P^k$

Limit Theorem for Markov Chains

S_n = base at generation *n* $\quad P_{ij} = P(S_{n+1} = j \mid S_n = i)$

If $P_{ij} > 0$ for all *i,j* (and $\sum_j P_{ij} = 1$ for all *i*)

then there is a unique vector $\vec{r}$ such that

$\vec{r} = \vec{r}P$ and $\lim_{n \to \infty} \vec{q}\,P^n = \vec{r}$ (for any prob. vector $\vec{q}$)

$\vec{r}$ is called the “stationary” or “limiting” distribution of P

See Ch. 4, Taylor & Karlin, An Introduction to Stochastic Modeling, 1984 for details

Stationary Distribution Examples

2-letter alphabet: R = purine, Y = pyrimidine

Stationary distributions for:

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$ 0 < *p* < 1

$$P' = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$ 0 < *p* < 1, 0 < *q* < 1

How does entropy change when a Markov transition matrix is applied?

If limiting distribution is uniform, then entropy increases

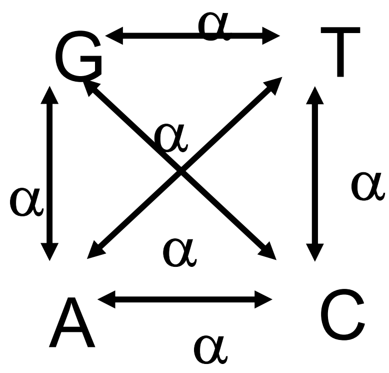
(analogous to 2nd Law of Thermodynamics)

However, this is not true in general (why not?)

How rapidly is the stationary distribution approached?

Jukes-Cantor Model

Courtesy of M. Yaffe



Assume each nucleotide equally likely
to change into any other nt,
with rate of change=α.
Overall rate of substitution = 3α
…so if G at t=0, at t=1, $P_{G(1)}=1-3\alpha$

and $P_{G(2)}=(1-3\alpha)P_{G(1)} + \alpha\,[1-P_{G(1)}]$

Expanding this gives $P_{G(t)}=1/4 + (3/4)e^{-4\alpha t}$

Can show that this gives K = -3/4 ln[1-(4/3)(p)]

K = true number of substitutions that have occurred,
P = fraction of nt that differ by a simple count.

Captures general behaviour…

Literature Discussion Tues. 3/16

Paper #1:

Kellis, M, N Patterson, M Endrizzi, B Birren, and ES Lander. "Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements." *Nature* 423, no. 6937 (15 May 2003): 241-54.

Part 1 - Finding Genes, etc., pp. 241-247
Part 2 - Regulatory Elements, pp. 247-254

Paper #2:

Rivas, E, RJ Klein, TA Jones, and SR Eddy. "Computational Identification of Noncoding RNAs in E. coli by Comparative Genomics." *Curr Biol* . 11, no. 17 (4 September 2001): 1369-73.