Some applications of Markov Chains & HMMs in molecular sequence analysis

Consider two sets of sequences; one set consist of putative CpG islands and the other non-CpG island sequences.

The following table (from Durbin *et al.*) show ML estimates of transition probabilities for the CpG (+) and non-CpG (-) sequences

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

$$S(x) = log\frac{P(x|\text{model}+)}{P(x|\text{model}-)} = \sum_{i=1}^{L} \log\frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}}$$

$$= \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

Where the *β's* are log likelihood ratios of the corresponding transitions under the alternative models

Table of scores

β	A	C	G	T
A	−0.740	0.419	0.580	−0.803
C	−0.913	0.302	1.812	−0.685
G	−0.624	0.461	0.331	−0.730
T	−1.169	0.573	0.393	−0.679

Probability P($S(x > d \mid -)$) < 2^{-d} (Milosavljević & Jurka, 1993)

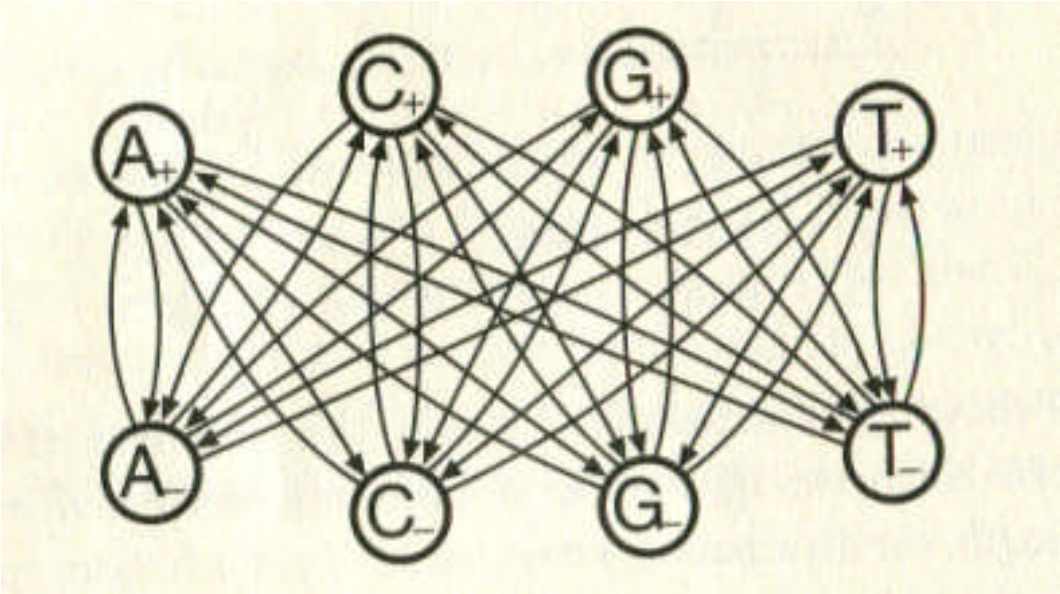
Given the table of scores, below, do you consider that the following sequence provides a significantly better fit to the CpG+ model than to the CpG- model?

A T C C T G G

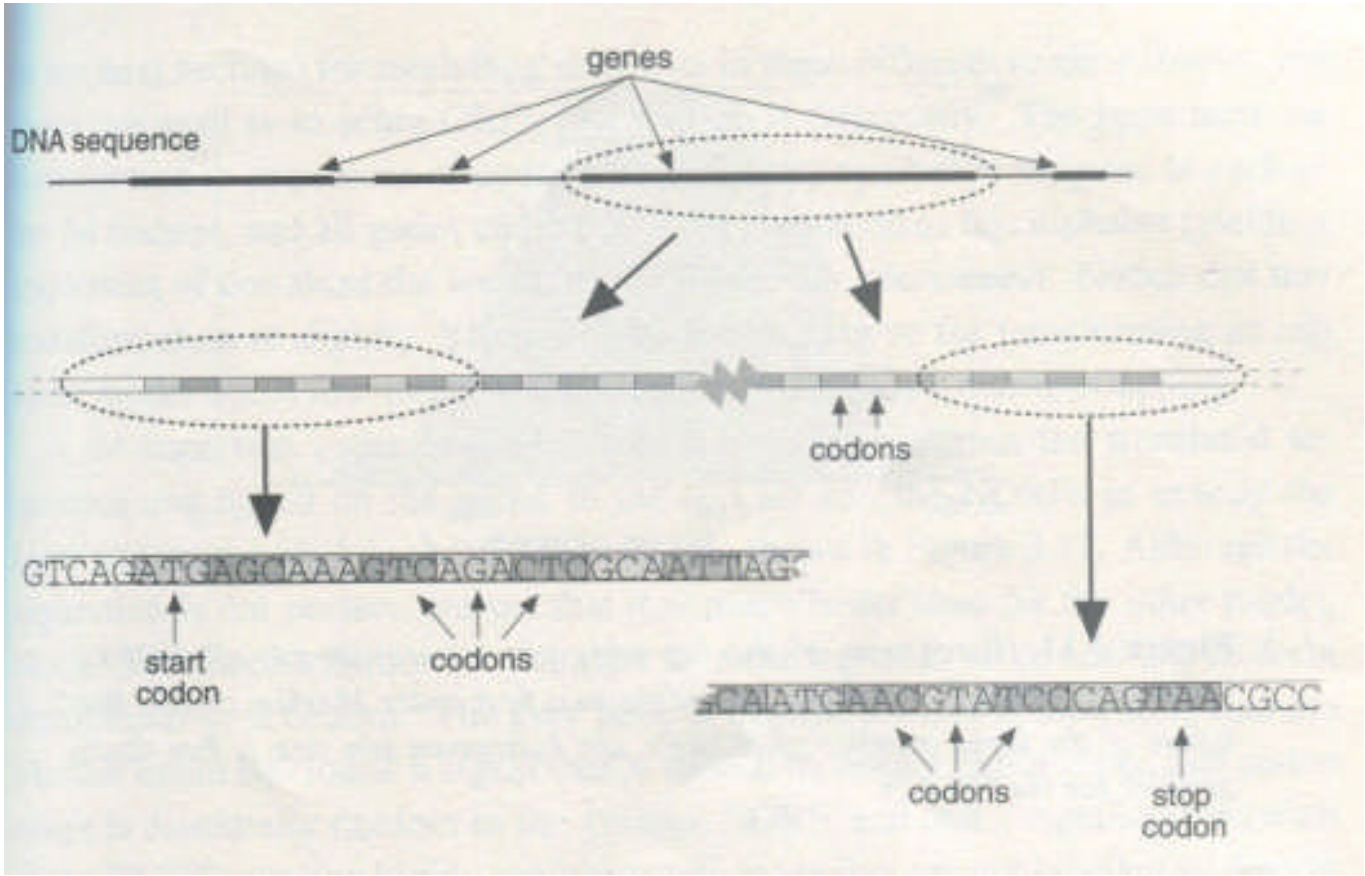
Table of scores

β	A	C	G	T
A	−0.740	0.419	0.580	−0.803
C	−0.913	0.302	1.812	−0.685
G	−0.624	0.461	0.331	−0.730
T	−1.169	0.573	0.393	−0.679

Hidden Markov model for finding CpG islands (Durbin *et al.*)



Finding prokaryotic genes



Prokaryote gene structure (Durbin *et al.* Fig 3.10)

Prokaryote gene structure

Prokaryote protein-coding genes (usually) contain a stretch of DNA obeying the following rules:

1. starts with a start codon (ATG)
2. followed by some number of non-stop codons
3. ends with a stop codon (TAG, TAA, TGA)

Stretches of DNA obeying these rules are called Open Reading Frames (ORFs)

Finding prokaryote genes (continued)

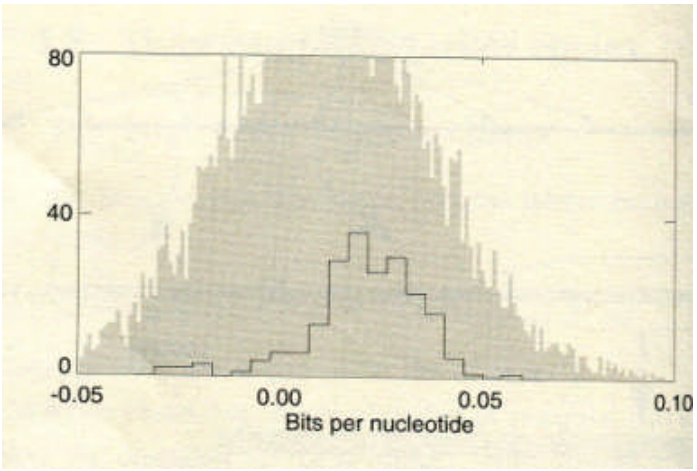
Obtained a set of 1,100 genes (min length 100bp), divided into

- training set (900 genes)
- test set (100 genes)

Identified all ORFs – defined NORFs as ORFs that do not overlap with known genes

Compared scores of genes versus NORFs based on a first-order Markov chain

Little power to discriminate genes from NORFs



Durbin *et al.* Fig 3.11

Shift to higher-order Markov Chain?

For an order n Markov chain

$$P(x_i|x_{i-1},\ldots,x_1) = P(x_i|x_{i-1},x_{i-2},\ldots,x_{i-n})$$

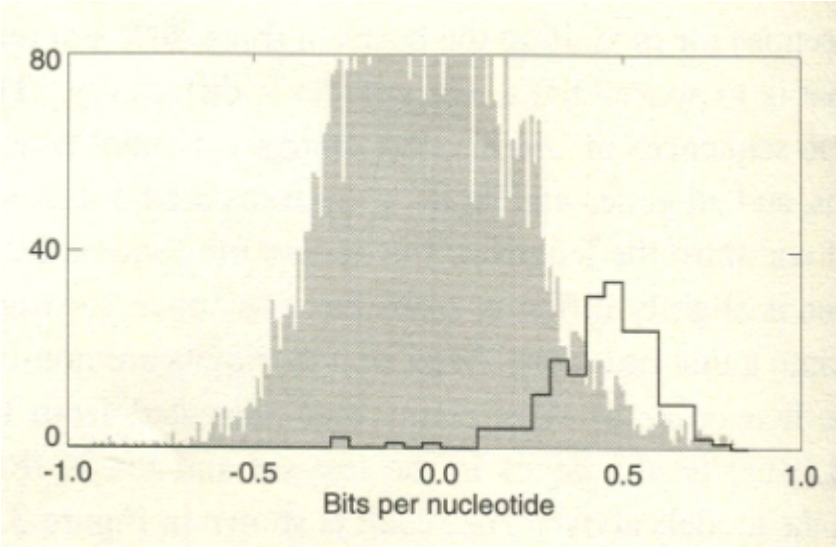
An order n Markov chain is the same as a first-order chain over the alphabet $\mathcal{A}^n$ of n-tuples, since

$$P(x_i|x_{i-1},\ldots,x_{i-n}) = P(x_i,x_{i-1},\ldots,x_{i-n+1}|x_{i-1},\ldots,x_{i-n})$$

Alternatively: consider codons instead of nucleotides

The order *n* Markov chain allows dependencies of the state on the previous *n* states to be modelled, but neglects translation frame

Instead consider chain over the alphabet of 61 ‘sense’ codons



Durbin *et al.* Fig 3.12

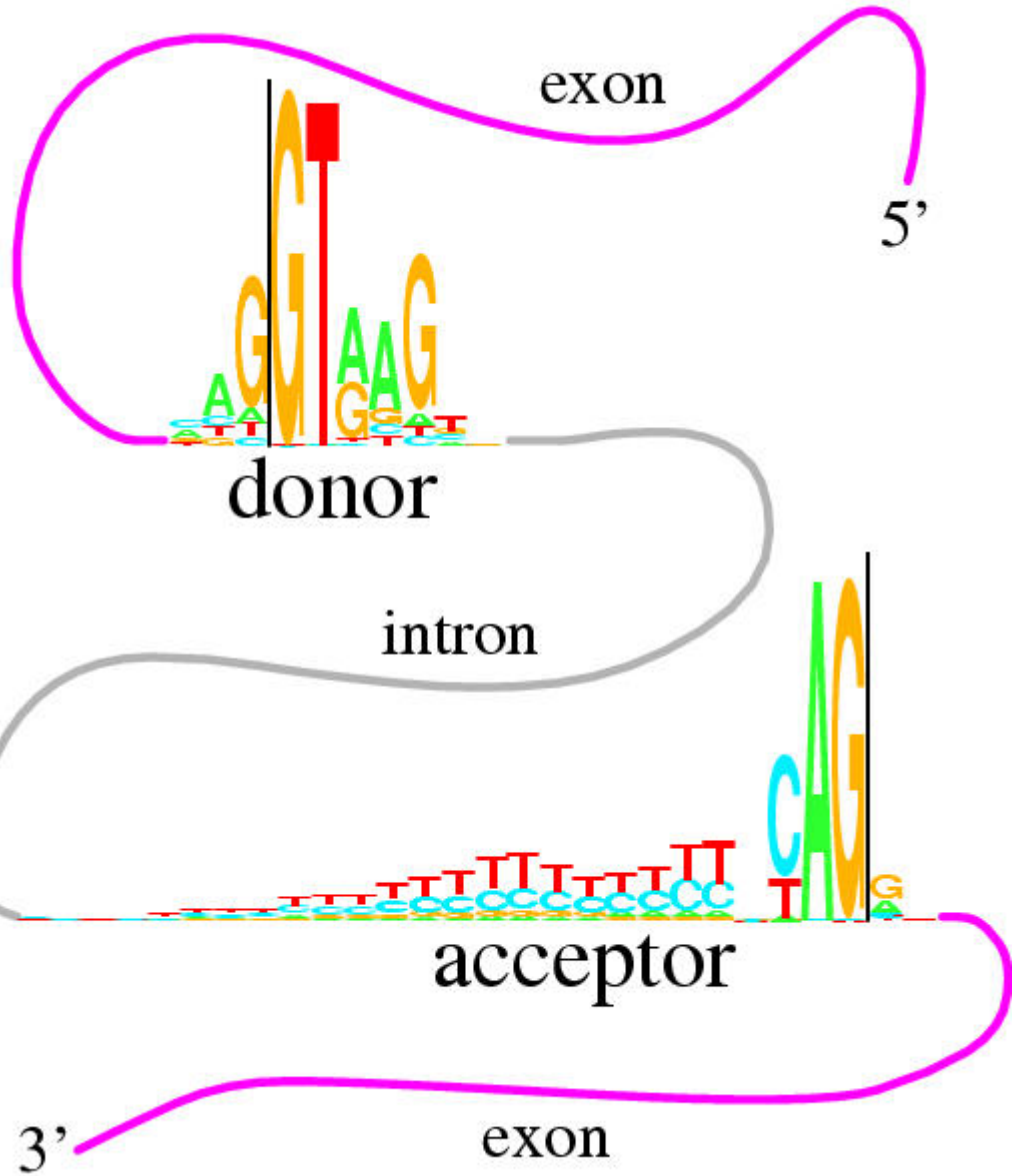
Predicting Eukaryotic genes

More complicated gene structure than prokaryotes (introns)

Typically much lower gene density than prokaryotes (e.g. ~3% coding regions in human genome)

Features of Eukaryotic genes

1. Complex and highly variable promoter region with binding sites for proteins called ‘transcription factors’
2. Transcription start site... first exon
3. 5' Untranslated Region (UTR)
4. Start codon (typically ATG)
5. Splice donor site (at end of exon/start of intron)
6. Splice acceptor site (at intron end/exon start)
7. Polyadenylation signal (typically AAUAAA)



From Stephens & Schneider, 1992

Gene prediction software packages

mGene
SNAP
GeneZilla (formerly TIGRscan)
GlimmerHMM
TWINSCAN
ChemGenome
TWAIN
ExoniPhy
Shadower
JIGSAW (formerly "Combiner")
GenomeThreader
EvoGene
ExonHunter
GlimmerM
Unveil
Glimmer
CRITICA
SGP2
Phat
geneid
SLAM
EuGene
GENSCAN
AUGUSTUS
SGP1
VEIL
MORGAN

GRAIL
GrailEXP
Genie
GeneMark(tm)
FGENESH
ORPHEUS
PROCRUSTES
Diogenes
WebGene
GeneSplicer
ORF Finder
SplicePredictor
NetPlantGene
NetGene2
ELPH
RBSfinder
TransTerm
OC1
TAIL
wotd
tigr++
PROSCAN/
PromoterScan
GenomeScan
DoubleScan
HMMgene

Prediction of Complete Gene Structures in Human Genomic DNA

Burge & Karlin, *Journal of Molecular Biology,* 1997