Markov Chains in Molecular Biology

J.K. Lindsey
Biostatistics, Limburgs University,
Diepenbeek, Belgium

1. Introduction to Molecular Biology Sequences

1.1 DNA Sequence Analysis

The double-stranded helical form of deoxyribonucleic acid (DNA) is well known.

Each strand of DNA consists of a linear sequence of the four nucleic acid bases, adenine (`A`), cytosine (`C`), guanine (`G`), and thymine (`T`).

Opposite strands contain complementary pairs: `A` with `T` and `C` with `G` so that only one of the strands need by studied.

In a protein-coding gene, consecutive, non-overlapping triplets of bases code corresponding sequences consisting of the 20 different amino acids that make up a protein.

This is called an open reading frame (ORF).

A coding region is read by messenger ribonucleic acid (mRNA) and translated by ribosomes into a polypeptide.

There are 64 possible combinations of the bases.

Thus, the code is redundant, particularly in the third base, with several triplets often coding the same amino acid.

Special three-base codes also signal the initiation (`ATG`) and termination (`TAG`, `TGA`, `TAA`) of a coding sequence.

A promoter and enhancer signal region, containing so-called promoter boxes (for example, `TATA`, `CCAAT`), generally occurs somewhat before the first exon in a protein-coding section.

Some other regions are genes coding for ribosomal (rRNA) or transfer (tRNA) ribonucleic acids.

Thus, most bases in a DNA sequence do not code for proteins.

Only selective sections of the strands are actually active.

In addition, the bases coding a given protein are not necessarily all consecutive but may be split into several sections.

These are called the *exons* of the gene whereas the non-coding sections in between are called *introns*.

Because the set of exons define a protein, they are subject to natural selection; one may expect the bases in the introns to be more random.

A mutation in an exon sequence will often result in a code for a non-viable or inappropriate protein, whereas a mutation in an intron does not have this harmful effect.

1.2 Sequencing Methods

A chromosome is first divided in some ordered way into smaller pieces.

DNA molecules are digested by restriction endonuclease, cutting them into small fragments.

Each specific endonuclease has a target site of cutting defined by a unique sequence of four to eight base pairs.

For example, the enzyme *Not*I recognizes the eight base pair sequence, `GCGGCCGC`.

Such sequences are not distributed randomly and the four nucleotide bases do not all appear equally frequently in the genome.

Thus, the length of the fragments produced depends of the target cutting sequence.

These fragments are separated by size using electrophoresis in agarose.

They are multiplied for mapping and sequencing to be possible.

Bacteriophage λ, bacteria containing cosmid recombinants, or yeast artificial chromosomes (YACs) can be used to clone the fragments and generate a library.

Then, the cloned fragments must be positioned in the same linear order as in the chromosome by detecting overlaps.

This produces a physical map of the chromosome.

One possibility for ordering the fragments is chromosome walking:

a clone is chosen and used as a probe to detect other clones with which it will hybridize; these should overlap with it.

This is repeated many times, providing a series of steps.

Other techniques such as restriction enzyme fingerprinting, marker sequences, and hybridization assays are also used.

The chain terminator or dideoxy method for DNA sequencing developed by Sanger uses two important properties of these molecules:

the ability to synthesize a complementary copy from a single strand of DNA and the possibility of using dideoxynucleotides as chain terminators.

DNA is synthesized in the presence of the four deoxynucleoside triphosphate bases, one of which is labelled with ^{32}P.

Four batches each contain a low concentration of one of the different dideoxynucleotides.

Because of the difference in termination, each batch will contain partially synthesized radioactive DNA molecules of different length.

A high-resolution sequencing gel fractionates denatured (single-strand) DNA fragments according to size by electrophoresis.

It is capable of distinguishing fragments differing in length by only one base pair.

The labelled DNA bands can be examined manually to determine the sequence after autoradiography on X-ray film.

The maximum length of DNA that can be sequenced at one pass is between 300 and 500 bases.

However, for the process to be automated, the radioactive tags are replaced by fluorescent ones attached to the terminators.

Each dideoxynucleotide carrying a different fluorophore.

The four bands can be then detected in the same lane of gel and many lanes electronically analyzed simultaneously.

The sequenced fragments can either be reassembled

(1) by previously constructing a physical map of the genome or

(2) by a shotgun approach of matching overlapping ends of fragments to produce the assembly.

During this process, the partial sequences created are known as contigs (contiguous sequences).

The final result of the assembly is a consensus sequence.

Roughly 5000 to 10 000 bases must be analyzed to produce a sequence of 1000 bases.

1.3 Alignment

DNA sequences coding similar proteins must be similar.

This will be true of two proteins in the same organism but also of those in two closely related organisms.

The latter may differ through evolutionary mutations.

On the other hand, the non-coding sequences may differ widely.

Only certain mutations that change an exon, those that still produce a viable protein, are permissible.

Mutations of the introns can be much more random because they do not affect the protein.

In order to compare such sequences, the DNA must be aligned.

Then, one can decide if such an alignment would likely to have arisen by chance or because the sequences are related.

Several factors must be taken into account:

- what alignments should be allowed;
- how should they be ranked;
- what algorithm should be applied to find an optimal alignment;

- what statistical procedure should be used to evaluate significance of the ranked scores.

Simple procedures only perform pairwise alignment.

Two basic types of mutations can change sequences:

(1) substitutions of one base for another and

(2) insertions or deletions of bases.

Some forms of mutations are observed more frequently than others because natural selection generally removes the nonviable ones.

For example, because of the redundancy in the third base of a triplet, more variability can often be observed there.

At each site, a score is assigned to the pair of bases occurring there.

For DNA bases, there are 16 possible scores but, by symmetry, not all are different.

These form a 4×4 score or substitution matrix.

To align sequences optimally, gaps may have to be left in some of the sequences, corresponding to insertions and deletions.

A penalty is assigned for opening a gap and another (usually smaller) one for widening it.

The total ranked score for an alignment, then, consists of a sum of terms for each aligned pair of bases plus those for the gaps.

Additivity implies that mutations at different sites have occurred independently.

Various algorithms are used to obtain optimal alignment among two or more sequences.

These dynamic programming techniques are guaranteed to find the optimal pairwise alignment.

A number of these programs are publicly available; sequences can also be submitted for alignment over the internet.

Global alignment of complete sequences is generally performed by the Needleman–Wunsch algorithm, whereas location alignment of subsequences uses the Smith–Waterman algorithm.

Multiple sequence alignments are more complex.

Scoring methods must allow for the evolutionary dependence among the sequences, including the fact that some sites may be more conserved than others.

Once a set of scores has been chosen, multidimensional dynamic programming must be applied.

1.4 Finding genes and their exons

Once a section of DNA has been sequenced so that its content is known, one question to be asked is which sections of it are active in coding a protein.

Evidence for the location of genes in a sequence must be derived from a variety of indications.

A protein-coding sequence may have a number of characteristics:

- it should be preceded by known promoter regions such as a `TATA` box;

- it should start with an initiation codon and end with a termination codon;

- it may be sufficiently similar to that for another gene in the genome or to the same gene in another genome to be recognizable;

- it can show codon (triplet) regularity;

- it is unlikely to contain major sections of repeats.

Gene finding is particularly difficult when introns are present.

Many types of software are available on the internet for

- integrated gene identification;

- promoter recognition;

- database searches to find similar gene sequences;

- repeat analysis.

2. Introduction to Log Linear Models

2.1 Data, Models, and Inference

Suppose that each observation, y_i, can take one among a small number of possible values.

For example, the four nucleic acid bases of DNA or RNA, or the 20 amino acids of proteins.

The results can be summarized as a frequency table giving the number of times, n_i, that each value occurs.

For the complete human betaglobin gene, the frequencies are

	A	C	G	T
n_i	360	277	296	491
$\widehat{\pi}_i$	0.25	0.19	0.21	0.34

For the exons, the frequencies are

	A	C	G	T
n_i	88	113	137	106
$\widehat{\pi}_i$	0.20	0.25	0.31	0.24

and for the introns,

	A	C	G	T
n_i	272	164	159	385
$\widehat{\pi}_i$	0.28	0.17	0.16	0.39

If the observations are independent, their joint probability can be written as

$$\Pr(\mathbf{n}) = \binom{n_\bullet}{n_1 \cdots n_I} \prod \pi_i^{n_i}$$

where $n_\bullet = \sum_{i=1}^{I} n_i$.

This is a *multinomial distribution*.

Models are defined by the way in which numbers are assigned to the probabilities, π_i, of the possible observed values.

Inferences are made by studying the probability of the observed data for various such models.

This is called the *likelihood function*, $L(\pi)$.

It is a function of the models, whereas the probability is a function of the data.

Often, it is easier to study the negative log likelihood:

$$-\log[L(\pi)] \propto -\sum n_i \log(\pi_i)$$

for which smaller values indicate better models.

The *maximum likelihood estimate* (mle) is the model that makes the data most probable or the negative log likelihood smallest.

For independence, the mles of π are just the relative frequencies.

2.2 Log Linear Models

Generally, the sequence of observed values is not independent.

It may depend on various factors.

Thus, in the betaglobin gene, the probabilities of the four bases appear to depend on whether they lie in an intron or an exon:

	A	C	G	T
Exon	0.20	0.25	0.31	0.24
Intron	0.28	0.17	0.16	0.39

One way to model this is to set

$$\pi_{ij} = \frac{e^{\mu_i + \alpha_{ij}}}{\sum_i e^{\mu_i + \alpha_{ij}}}$$

where j indexes the location of the base.

Some constraints need to be placed on the parameters, such as $\sum_i \mu_i = 0$ and $\sum_i \alpha_{ij} = \sum_j \alpha_{ij} = 0$.

Then, this can be rewritten as

$$\log\left(\frac{\pi_{ij}}{\dot{\pi}_j}\right) = \mu_i + \alpha_{ij}$$

where $\dot{\pi}_j$ is the geometric mean of the probabilities at location j.

The mles are $\widehat{\mu} = (-0.02, -0.15, -0.07, 0.24)$ and

$$\widehat{\alpha} = \begin{pmatrix} -0.20 & 0.18 & 0.29 & -0.28 \\ 0.20 & -0.18 & -0.29 & 0.28 \end{pmatrix}$$

reflecting the fact that introns have fewer C and G bases.

In the model of independence, the probabilities of the bases are the same in both locations.

That is $\pi_{ij} = \pi_i$ for all i and j.

This can be written

$$\log\left(\frac{\pi_{ij}}{\dot{\pi}_j}\right) = \mu_i$$

The respective negative log likelihoods are 1936.3 for independence and 1900.7 when a difference between introns and exons is allowed.

The model with dependence on location makes the observed data much more probable:

$e^{1936.3-1900.7} = 3.1 \times 10^{15}$ times more probable!

However, the latter model has three extra parameters.

In making inferences, this can be allowed for by penalizing the negative log likelihood by adding the number of estimated parameters.

These are respectively 3 and 6, yielding 1939.3 and 1906.7.

This is penalization is called the Akaike Information Criterion (AIC).

2.3 Software

Most available software does not allow direct modelling of the multinomial distribution.

Generally, only the *Poisson distribution* is available:

$$\Pr(n_i) = \frac{\mathrm{e}^{-\nu_i}\nu_i^{n_i}}{n_i!}$$

Here, ν_i is the theoretical average *number* of events of type i, while π_i was the theoretical *proportion* of events of that type.

Suppose that a set of frequencies, $n_1 \cdots n_I$, has a Poisson distribution with means $\nu_1 \cdots \nu_I$.

Then, their sum, $n_\bullet$, also has a Poisson distribution with mean, $\nu_\bullet$, the sum of the individual means.

Recall that the conditional probability for an event A given an event B is defined by

$$\Pr(A|B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)}$$

Then, if we condition on the total number of events,

$$\begin{aligned}
\Pr(n_1, \ldots, n_I | n_\bullet) &= \frac{\prod_{i=1}^I \mathrm{e}^{-\nu_i} \nu_i^{n_i} / n_i!}{\mathrm{e}^{-\nu_\bullet} \nu_\bullet^{n_\bullet} / n_\bullet!} \\
&= \frac{n_\bullet! \mathrm{e}^{-\nu_\bullet} \prod_{i=1}^I \nu_i^{n_i}}{\prod_{i=1}^I n_i! \mathrm{e}^{-\nu_\bullet} \nu_\bullet^{n_\bullet}} \\
&= \binom{n_\bullet}{n_1 \cdots n_I} \prod_{i=1}^I \left(\frac{\nu_i}{\nu_\bullet}\right)^{n_i}
\end{aligned}$$

which is the multinomial distribution with $\pi_i = \nu_i / \nu_\bullet$.

The two distributions are identical.

Thus, the Poisson distribution can be used for log linear models instead of the multinomial.

For example,

$$\log\left(\frac{\pi_{ij}}{\dot{\pi}_j}\right) = \mu_i + \alpha_{ij}$$

with multinomial probabilities is equivalent to

$$\log\left(\frac{\nu_{ij}}{\dot{\nu}_j}\right) = \mu_i + \alpha_{ij}$$

with Poisson means.

Most common software use a standard notation to communicate models.

Variables are specified by their names.

The model

$$\log(\nu_{ij}) = \log(\dot{\nu}_j) + \mu_i + \alpha_{ij}$$

would correspond to

$$location + base + base \cdot location$$

This can also be written more simply as

$$base * location$$

The independence model is

$$base + location$$

2.4 More Complex Models

Often, we may wish to study how more than one factor influences the probabilities of the observed values.

For example, does the distribution of nucleic acid bases differ among species as well as between exons and introns?

		A	C	G	T
Human	Exon	0.20	0.25	0.31	0.24
	Intron	0.28	0.17	0.16	0.39
Chimp	Exon	0.19	0.25	0.32	0.24
	Intron	0.27	0.17	0.16	0.40
Gorilla	Exon	0.19	0.24	0.32	0.25
	Intron	0.28	0.17	0.16	0.39

We can extend our model to

$$\log\left(\frac{\pi_{ijk}}{\dot{\pi}_{jk}}\right) = \mu_i + \alpha_{ij} + \beta_{ik} + \gamma_{ijk}$$

with with k indexing species.

Constraints on the parameters similar to those above are also required.

β_{ik} will measure the differences among species.

γ_{ijk} will allow for the possibility that the relationship between exons and introns differs among species.

The model with only differences between locations and not among species,

$$\log\left(\frac{\pi_{ijk}}{\dot{\pi}_{jk}}\right) = \mu_i + \alpha_{ij}$$

has an AIC of 4493.6 with six parameters.

This compares to 4590.1 for the independence model with three parameters.

If we also allow for species differences

$$\log\left(\frac{\pi_{ijk}}{\dot{\pi}_{jk}}\right) = \mu_i + \alpha_{ij} + \beta_{ik}$$

the AIC is 4499.4 with 12 parameters.

Finally, the full model has an AIC of 4505.3 with 18 parameters.

These models indicate no significant differences among the three species.

When using the Poisson approach in software, the minimal model can be written

$$R_1 + R_2 + \cdots + E_1 * E_2 * \cdots$$

R_i represents a response variable (here only one, base type).

E_j an explanatory variable (here location and species).

The product indicates all possible combinations of interactions among variables.

This cannot be simplified even if the AIC indicates that some terms are unnecessary.

In our example, independence is specified by

$$base + location * species$$

Dependency of a response on an explanatory variable is introduced as a product:

$$R_1 + R_2 + \cdots + E_1 * E_2 * \cdots + R_1 * E_1$$

as is dependency between responses:

$$R_1 + R_2 + \cdots + E_1 * E_2 * \cdots + R_1 * R_2$$

Thus dependency of base type on location is given by

$$base + location * species + base * location$$

that on species by

$$base + location * species + base * location$$

and on both by

$$base + location * species + base * location + base * species$$

3. Introduction to Markov Chains

3.1 Serial Dependence

A finite number of different types of events, observed in a sequence, defines the *states*, say x, of the process.

Suppose that the individual value, y_t, at a given point, t, in the sequence depends only on the state, y_{t-1}, at the immediately preceding point:

$$\Pr(y_t|y_{t-1}, \ldots, y_1) = \Pr(y_t|y_{t-1})$$

This is the hypothesis of a *first-order* Markov chain.

Because DNA sequences are read in one direction ($5'$ to $3'$), Markov chain theory can be applied.

Then, the probability for a complete sequence is

$$\Pr(y_1, \ldots, y_N) = \Pr(y_1) \prod_{t=2}^{N} \Pr(y_t | y_{t-1})$$

These conditional probabilities can be represented in a square *transition matrix*, $\mathbf{T}$, of each state given the previous one.

If if depends further back, the chain is of higher order.

If the rows correspond to the states at the previous time point and the columns to the present states, then the row probabilities sum to one.

If this matrix is the same for all positions in the sequence, the chain is said to be *homogeneous*.

Pre-multiplying this matrix by the vector, $\mathbf{n}_t$, of frequencies of units in the different states (the marginal frequencies) at a given point, t, will give the vector for the next point, $t+1$:

$$\mathbf{n}_{t+1}^T = \mathbf{n}_t^T \mathbf{T}$$

The marginal *stationary distribution* of the states is the π such that

$$\pi^T = \pi^T \mathbf{T}$$

A Markov chain is said to be *irreducible* if any state can be reached from any other.

Various assumptions about Markov chains, such as order or homogeneity, can be compared by fitting appropriate log linear models

3.2 More Complex Markov Chains

If the present state depends on the k previous states

$$\Pr(y_t|y_{t-1},\ldots,y_1) = \Pr(y_t|y_{t-1},\ldots,y_{t-k})$$

the chain is said to be of order k.

Any such sequence can be written as a first-order Markov chain by changing the state space.

Instead of the states, x, take the states to be all possible combinations of a set of k xs.

For example, with $k = 2$, a sequence CGTCA becomes CG–GT–TC–CA.

Here, some of the transition probabilities must be zero: TC cannot follow CG, etc.

If the transition matrix changes depending on the position in the sequence, the chain is inhomogeneous.

A DNA sequence coding a protein consists of triplets.

The transition matrix may depend on the position in the triplet.

There will be three different matrices, at positions 1, 2, and 3.

Within a gene, the transition matrix may be different between exons and introns.

3.3 Comparing Transition Matrices in Exons and Introns

For the betaglobin data, the transition matrix for the entire gene is

	A	C	G	T
A	0.29	0.18	0.21	0.32
C	0.31	0.23	0.03	0.43
G	0.21	0.22	0.30	0.27
T	0.22	0.17	0.25	0.36

As might be expected, we see that C is very rarely followed by G.

The matrices for the exons and for the introns are respectively

	A	C	G	T
A	0.26	0.28	0.31	0.15
C	0.27	0.31	0.04	0.38
G	0.18	0.25	0.33	0.24
T	0.08	0.19	0.57	0.16

and

	A	C	G	T
A	0.29	0.15	0.18	0.38
C	0.34	0.18	0.02	0.45
G	0.24	0.19	0.27	0.30
T	0.25	0.16	0.17	0.42

We can use log linear models to investigate if there is a difference in transitions between exons and introns.

The contingency tables are

	A	C	G	T
Exons				
A	23	24	27	13
C	30	35	5	43
G	25	34	45	33
T	9	20	60	17
Introns				
A	80	42	48	102
C	56	30	3	74
G	38	30	43	48
T	98	62	65	160

The independence model, where the base at a given position depends neither on the previous base nor on the location (exon or intron),

$$base + location * previous$$

has an AIC of 3663.3 with 3 parameters.

That with dependence only on location,

$$base + location * previous + base * location$$

has 3627.3 with 6 parameters and that for previous base only

$$base + location * previous + base * previous$$

has 3609.8 with 12 parameters.

This latter model assumes that the transition matrix is the same in exons and introns.

However, the model where dependence on the previous base is different in the exon and intron

$$base * location * previous$$

has an AIC of 3555.0 with 24 parameters.

This shows that the relationship between successive bases is different in exons and introns.

The two transition matrices are significantly different.

The sequence over the whole gene is not homogeneous.

4. Introduction to Hidden Markov Models

4.1 Basic Concepts

Suppose that a sequence of responses is discrete-valued, often categories that would *appear* to be the observed states of some Markov chain.

However, dependence cannot adequately be described by the simple Markov property.

In a hidden Markov model, an underlying, *unobserved* sequence of states follows a Markov chain, the hidden state determining the probabilities of the observed states.

Such an approach is widely used in speech processing and in biological sequence analysis of nucleic acids in DNA and of amino acids in proteins.

For a binary time series, each event might be generated by one of two Bernoulli distributions.

The process switches from the one to the other according to the state of the hidden Markov chain, in this way generating state dependence.

Analogous models can be constructed for other discrete distributions, such as the Poisson or multinomial distributions.

The distributions could even, themselves, be Markov chains with different transition matrices.

4.2 The Model

Consider an irreducible homogeneous Markov chain with $M \times M$ transition matrix, $\mathbf{H}$.

This gives the probabilities of changing among the hidden states, with marginal stationary distribution, $\boldsymbol{\pi}$.

The latter can be calculated from the transition matrix and hence does not introduce any new parameters.

Then, the probability of the observed response at position t, $\nu_{mt} = \mathsf{Pr}(y_t|m; \boldsymbol{\kappa}_m)$, will depend on the unobserved state, m, at that position.

ν_{mt} is called the *emission probability*.

The series of responses on a given unit are assumed to be independent, given the hidden state.

Thus, there are $M(M-1)$ unknown parameters in the transition matrix as well as M times the length of $\boldsymbol{\kappa}_m$ in the probability distributions.

Although the probability of the observed data is complex, it can be written in a recursive form over the sequence:

$$f(\mathbf{y};\boldsymbol{\kappa},\mathbf{H}) = \boldsymbol{\pi}^T \prod_{t=1}^{R} (\mathbf{H}\mathbf{F}_t)\mathbf{J}^T$$

$\mathbf{F}_t$ is an $M \times M$ diagonal matrix containing, on the diagonal, the probabilities, ν_{mt}, of the observed data given the various possible states.

To construct the likelihood function from this, first calculate the marginal probability times the observed probability for each state at position 1, say $a_m = \pi_m \Pr(y_1|m;\boldsymbol{\kappa}_m)$.

At the second point, the first step is to calculate the observed probability for each state multiplied by this quantity and by the transition probabilities in the corresponding column of $\mathbf{H}$.

These are summed yielding, say $b_m = \sum_h a_m H_{mh} \Pr(y_2|h;\boldsymbol{\kappa}_h)$.

This is the new vector of forward probabilities, but, to prevent underflow, it is divided by its average, yielding a new vector, $\mathbf{a}$.

This average is also cumulated as a correction to the likelihood function.

These steps are repeated at each successive position.

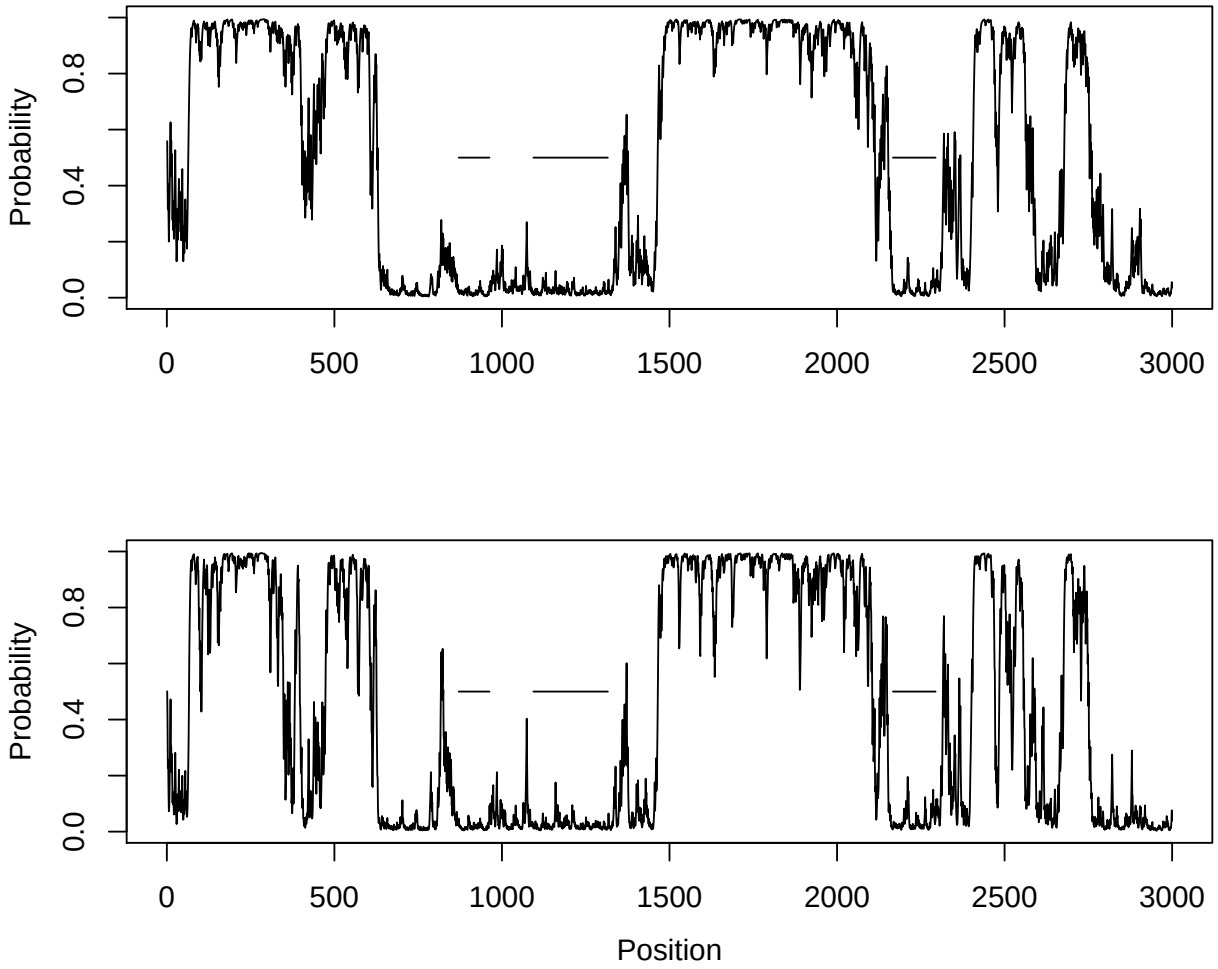
Finally, the sum of these a_m at the last point in the sequence is the likelihood except that the cumulative correction must be added to it.

At each step, the vector, $\mathbf{a}$, divided by its sum gives the (filtered) conditional probabilities of being in the various possible states given the previous observations.

4.3 Locating the Betaglobin Gene

Let us first apply hidden Markov models to the complete sequence of 3007 bases to see if any correspondence can be found between the hidden states and the coding sections.

The model for multinomial independence has an AIC of 4091.8, whereas that with two hidden states has 4044.2.



Filtered conditional probabilities of being in state 1 for the complete β-globin sequence. Top graph: simple two state model; bottom graph: model with two Markov chains.

We see that the three exons are all completely located in one of the states.

The second intron is similar to the sections of the sequence before and after the gene whereas the first intron is indistinguishable from the exons by this method.

The transition matrix is

$$\begin{pmatrix} 0.997 & 0.003 \\ 0.003 & 0.997 \end{pmatrix}$$

with stationary probabilities, 0.481 and 0.519.

In the first state, the probabilities of A, C, G, and T are respectively 0.31, 0.15, 0.14, and 0.40.

In the second, they are 0.23, 0.25, 0.26, and 0.27.

The latter is the state in which the exons occur.

Thus, the noncoding regions are CG poor.

Adding a third state further reduces the AIC to 4023.8 but does not further aid in distinguishing the gene.

Allowing the probability of each type of base to cycle through each of the three positions of triplets along the whole sequence with two hidden states does not improve the model;

the AIC is 4048.6.

On the other hand, if an ordinary Markov chain is used instead of a hidden one, the AIC is reduced to 3997.0.

If the process is allowed to switch between two such Markov chains using a hidden Markov model, the AIC is 3938.6.

The hidden transition matrix for this model is

$$\begin{pmatrix} 0.995 & 0.005 \\ 0.004 & 0.996 \end{pmatrix}$$

and the two 'observed' transition matrices are

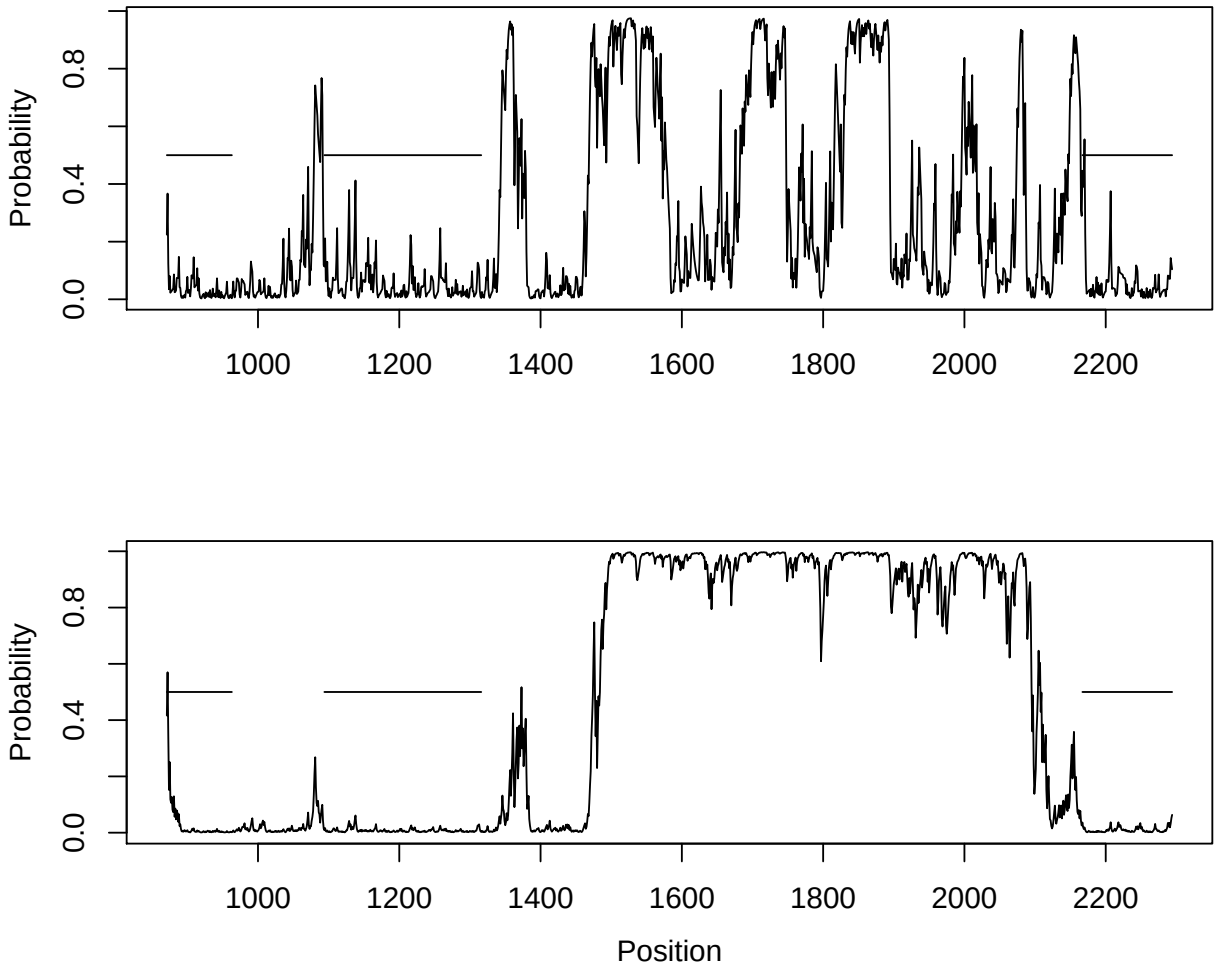
	A	C	G	T
A	0.338	0.140	0.124	0.399
C	0.365	0.284	0.064	0.288
G	0.141	0.317	0.394	0.148
T	0.193	0.175	0.420	0.212
A	0.240	0.223	0.331	0.206
C	0.288	0.300	0.035	0.377
G	0.254	0.212	0.313	0.222
T	0.164	0.242	0.326	0.268

Notice how rarely G follows C in either state.

4.4 Locating the exons

Let us now look more closely at the gene itself, ignoring the noncoding regions on each side.

The multinomial independence model has an AIC of 1939.3 compared to 1913.2 for the two-state model.



Filtered conditional probabilities of being in state 1 for the gene section of the β-globin sequence. Top graph: simple two state model; bottom graph: model with dependence on triplet position.

The complete exons still occur in one hidden state.

However, the second intron is not so clearly distinguished as when the whole sequence is used.

On the other hand, there is some indication of the first intron being similar to the second.

The transition matrix is

$$\begin{pmatrix} 0.976 & 0.024 \\ 0.009 & 0.991 \end{pmatrix}$$

with stationary probabilities, 0.265 and 0.735.

The probabilities of the four bases in state 2, containing the exons, are respectively 0.27, 0.20, 0.26, and 0.27, whereas they are 0.21, 0.18, 0.06, and 0.54 in state 1.

Indeed, 41% of intron 2 consists of T.

If I now allow a different set of probabilities for the four bases at each of the three positions in a triplet, the AIC is reduced to 1912.0.

This is rather surprising as only the second exon has a complete set of triplets and neither of the introns does.

Note that the triplets in the second exon do not correspond to amino acids because the first intron occurs in the middle of a triplet.

Thus, triplets are out of alignment among the three exons.

Nevertheless, the changes of state become much clearer, as can be seen in the lower graph.

The transition matrix is now

$$\begin{pmatrix} 0.998 & 0.002 \\ 0.001 & 0.999 \end{pmatrix}$$

with stationary probabilities, 0.317 and 0.683.

The probabilities of the four bases at the three positions of a triplet in the two states are summarized in the following table:

State	Position	A	C	G	T
1	1	0.29	0.18	0.10	0.43
	2	0.31	0.13	0.13	0.43
	3	0.29	0.14	0.11	0.46
2	1	0.19	0.21	0.29	0.31
	2	0.25	0.25	0.27	0.23
	3	0.22	0.23	0.28	0.27

As for the complete sequence, adding a third state improves the model, with an AIC of 1896.9, but does not further help to locate the coding regions.

This example should not be taken as typical of the success with which coding sections of a sequence can be located.

It happens that intron 2 of this gene is rather special; this greatly helped in locating the areas of interest.

4.5 Extensions

- nonstationary marginal distribution
- inhomogeneous hidden transition matrix
- higher order hidden Markov chain

5. Applications of Hidden Markov Models

5.1 Finding CpG Islands

The dinucleotide, CG (written CpG to distinguish it from the C–G base pair across strands) occurs rarely.

In this combination, C is usually methylated and mutated to T.

In certain short sections of a genome, methylation is suppressed, such as in promoter regions of a gene.

These CpG islands are generally a few hundred to a few thousand bases long.

In a CpG island, the transition matrix will be different than elsewhere in the genome.

The transition probability, C→G will be larger.

In a set of 41 human DNA sequences with 48 known CpG islands, the transition matrices are

	A	C	G	T
A	0.18	0.27	0.43	0.12
C	0.17	0.37	0.27	0.19
G	0.16	0.34	0.38	0.13
T	0.08	0.36	0.38	0.18

for CpG islands and

	A	C	G	T
A	0.30	0.21	0.29	0.21
C	0.32	0.29	0.08	0.30
G	0.25	0.25	0.29	0.21
T	0.18	0.24	0.29	0.29

elsewhere.

The problem is that we do not know at what point the transition matrix changes.

One of the first applications of hidden Markov models in molecular biology was to resolving this problem.

In the above 41 sequences, all but two CpG islands were found but, 121 others were also predicted.

However, the falsely predicted ones were quite short compared to the real ones.

Predictions less than 500 bases apart can be concatenated and those shorter than 500 bases ignored.

This reduces the false predictions to 67.

5.2 Pairwise Alignment

Consider a short section of the human betaglobin sequence, TGTACATATACACATATATATATATATTT as aligned with that of a chimpanzee, GTATATATACATACATATATATATATATATATATAT:

```
TG.....TACATATACACATATATATATATAT..TT
GTATATATACATACATATATATATATATATATATAT
```

After optimal alignment, the observed states in the two sequences may be

- identical nucleotides,
- different nucleotides,
- a gap in one sequence and a nucleotide in the other.

In aligning two sequences, we can have three possible hidden states:

1. the bases in the two sequences are aligned (M),

2. the first sequence requires an insert opposite a gap in sequence 2 (X_1),

3. the second sequence requires an insert opposite a gap in sequence 1 (X_2),

Then, the hidden transition matrix will be

	M	X_1	X_2
M	$1-2\delta$	δ	δ
X_1	$1-\epsilon$	ϵ	0
X_2	$1-\epsilon$	0	ϵ

δ is the probability of opening a gap.

ϵ is the probability of widening an existing gap.

$1 - \epsilon$ is the corresponding probability of closing a gap

There will be 16 emission probabilities in state M corresponding to all possible combinations of pairs of nucleotides

and four emission probabilities in each of states X_1 and X_2 corresponding to the possible nucleotide insertions.

Using hidden Markov models for alignment instead of dynamic programming algorithms provides

- likelihood measures of reliability of the alignment obtained,

- comparison of suboptimal alignments.

Generally, there will be several alternative alignments with almost the same likelihood.

Some will differ only in a few positions from the optimal alignment.

If there are repeats in one or both sequences, suboptimal alignments may differ substantially or completely from the optimal alignment.

5.3 Multiple Alignments

Aligning simultaneously several sequences is much more complex.

Usually they are sequences of DNA for similar proteins (α-, β-, and γ-globin) or sequences for the same protein from different species.

For different species, they are used to construct phylogenic trees in the study of evolution.

5.4 Selected References

Churchill, G.A. (1989) Stochastic models for heterogeneous DNA sequences. *Bull Math Bio* **51**, 79–94.

Churchill, G.A. (1992) Hidden Markov chains and the analysis of genome structure. *Comp Chem* **16**, 107–115.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids.* Cambridge: Cambridge University Press.

Elliot, R.J., Aggoun, L., and Moore, J.B. (1995) *Hidden Markov Models.* Berlin: Springer-Verlag.

Juang, B.H. and Rabiner, L.R. (1991) Hidden Markov models for speech recognition. *Technometrics* **33**, 251–272.

MacDonald, I.L. and Zucchini, W. (1997) *Hidden Markov and other Models for Discrete-valued Time Series.* London: Chapman & Hall.

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* **77**, 257–286.