Bioinformatics and Bioengineering Summer Institute
Identification of genes and other features by Markov models

I. Markov models

Position-specific scoring matrices are great for what they're great for, but what about those situations where you want to identify features that don't come in columns? For example, if you want to find genes, distinguishing open reading frames of biological significance from those that happen merely to appear in the genome, what then? Outside of the start and stop codons, there are no obvious positions of the gene that have predictable nucleotides.

And yet, the gene as a whole is predictable. Open reading frames that are real genes may be, indeed ought to be distinguishable from random sequences of DNA that begin with a start codon and end with a stop codon. We may not be able to set forth exactly how to distinguish the two, but there ought to be some way of feeding a clever program a set of known genes and say, Analyze this! Then give the program unknown open reading frames, and ask whether they share the characteristics found in true genes.

If this sounds fanciful, consider that you are able to do just this in another area. Read the phrases below and ask based on your knowledge of known texts whether they were written by George Lucas:

> *But you to other galled eyes, and this author of thousand moments; while most on to tell uses offence 'twill stale or there!*
>
> *That wickled all though thin infines, my somes life itsell; And wift, anday to to beasy;*
>
> *To be as into night, what fried indeed willion, Or lord; that I am not shall a said was it rant.*
>
> *Hold it truly; it escoted? Will my her in you know a knave.*

Who wrote these words? Francis Bacon? Christopher Marlowe? Edward de Vere? Perhaps even ***The Bard?*** Actually, they were written by a Dell PC running *Hamlet* (after having digested all words uttered by Hamlet). At times the texts produced seem like monkeys at a typewriter, at other times Hamlet seems to shine through. The program knew nothing of English, let alone the human condition. All it did was analyze the tendencies of one letter to follow another and produce random text that followed those tendencies.[1]

Now try these:

> *Away to the little for your knees!*
> *Oh, gathe little to the new angels sight*
> *Round gather and gathey shepherds quakes.*
>
> *Hark! the feast who,*
> *lowinter Proclaid dream*
> *from her and every worth.*
> *A ther King Glory hear thy dearth.*

[1] The program is freely available and might be useful for those writing theses at 3 am.

Mary wondering love's because of old:
Peace. Sleep on the Saviour knees!
O night nightly night, his the the sing heaven.
Bright, O holy shephen on Mary world Jesus,
holy her and child in and nature see the star,
He ago in their watch are shepher King;
Let evenly peat ther King willness loods,
whild, angels sing to the hay.

Definitely not Shakespeare, but the writer is the same: Hamlet.pl, but using a different input text, composed of a variety of Christmas carols.

Hamlet.pl begins by performing a Markov analysis of the input text, creating a table of tendencies. Most often, we use that table to ***identify*** unknown input. For example, if you had access to the two tables underlying the first set of texts and the second, but you did not have access to either bona fide Hamlet or Christmas carols and had never heard either, you would no doubt be able to identify whether a text given to you was closer to Hamlet's speech or the conventions of Christmas carols.

IV. Markov models behind the scenes

A Markov model is a table of frequencies, answering the question: given that such-and-such states have occurred, what's the probability that another specified state will follow? The model is calculated from a training set, a set of sequences that are known to have a certain property. You can build a training set from proven open reading frames, ribosome binding sites,… just about any sequence feature.

Markov models are characterized by an ***order***. For example, a third-order Markov model uses three states to predict a fourth. To make a Markov model, it is necessary to have a set of sequences known to possess the desired characteristic. The process is illustrated for a third-order Markov model in Fig. 1.

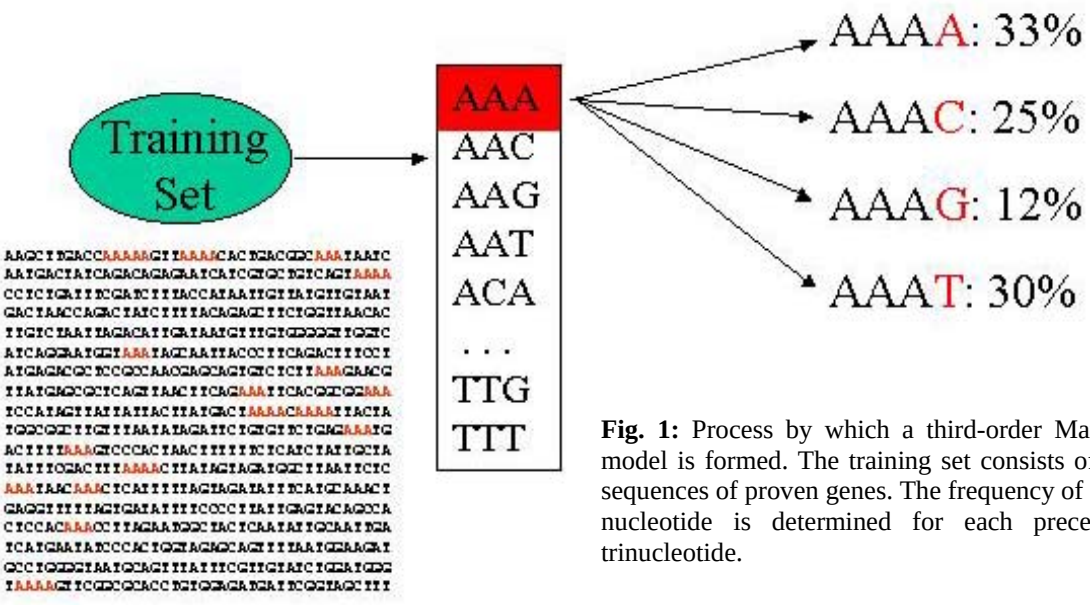

Fig. 1: Process by which a third-order Markov model is formed. The training set consists of the sequences of proven genes. The frequency of each nucleotide is determined for each preceding trinucleotide.

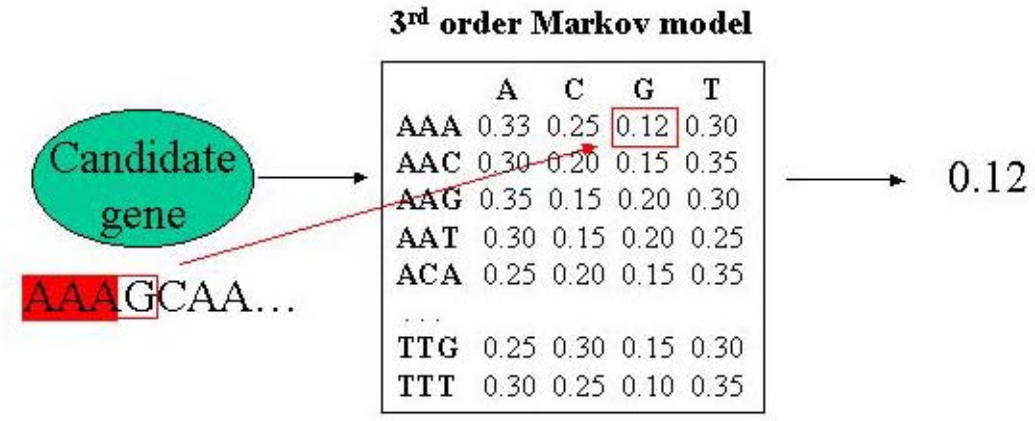

Fig. 4. Use of a Markov model to predict whether a sequence is or is not part of a gene. A four-nucleotide window scans a nucleotide sequence. For each position, the probability of the fourth nucleotide, given the prior three nucleotides, is found in the table constructed from proven genes (**Fig. 5**).

Figure 4 illustrates how a Markov model is used. We will see next time how these probabilities calculated for each position in a sequence are combined to form a single prediction as to whether a gene is closer in its structure to a native gene of an organism or one foreign to it. We'll do this by unpacking the program *Hamlet* figuring out how to modify it to work on genes.