VCU Bioinformatics and Bioengineering Summer Institute

Markov Models and Set Distinction - Introduction

Outline:

I. Orthologous genes
II. Markov models in action
III. Markov models behind the scenes

I. Orthologous genes

The Scenario describes an attempt to gain insight into the functions of unknown genes related to a complex phenomenon. The central trick was collecting all genes that are ***orthologs*** of each other and are found in all available genomes of photosynthetic bacteria. The notion of ortholog is critical to many bioinformatic applications, yet it is widely misunderstood. Imagine that you can see time as a physical dimension. If, by following over evolutionary time the paths of two genes, you find that they converge on a common ancestor without being part of a gene duplication, then those two genes are orthologs.

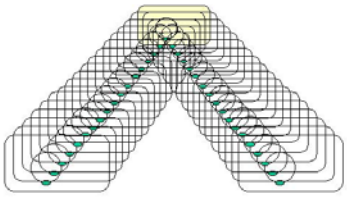

Fig. 1: Two orthologous genes arise by lineal descent from a common ancestor.

Notice that ***orthologous*** is a different property from ***similar*** (sometimes erroneously called ***homologous***). Two orthologous genes may be not very similar, if they have mutated at a relatively rapid rate. This would be the case in proteins whose functions do not make strict demands over their full lengths. Likewise, two very similar genes may not be orthologous, if a gene duplication occurred.

Whether two genes are orthologs, therefore, is not a question of classification, which might differ according to tastes, but rather a matter of historical fact. Unfortunately, since we ***can't*** see back in time, we can only make inferences about a gene's history.

Genes do duplicate, however, and that complicates matters, giving rise to sister genes called ***paralogs***. Gene duplication is the raw material for the evolution of new protein functions. For example, the cyanobacterium *Nostoc punctiforme* possesses over a hundred genes encoding regulatory proteins called histidine kinase. They surely arose from gene duplication.

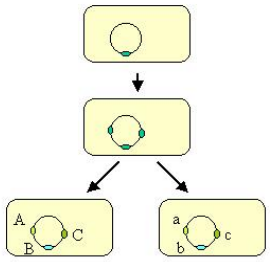

Fig. 2: A gene duplication event gives rise to paralogous genes in two related organisms, plus three pairs of orthologs.

Many working definitions of orthologous genes have been proposed, to overcome our inability to see directly the actual lineage of a gene. The most common definition is called ***two-way best hit*** or ***bi-directional best hit***. If gene ***A*** is the most similar to gene ***a*** in another organism, and gene ***a*** is the most similar gene to gene ***A***, then the two genes are declared to be orthologs. Note that this is not a definition but a prediction.

SQ1. Suppose that a gene encoding a liver enzyme in hamsters that detoxifies caffeine is similar to an enzyme in dogs that detoxifies a compound found in chocolate. The latter enzyme has no effect on caffeine. Describe the evolutionary history of these

two genes that would (a) convince you to call them orthologs or (b) convince you to call them paralogs.

II. Markov models in action

The scenario ends by pointing to the need for an analytical method to identify regions likely to be part of coding sequences. The method should be independent of the presence of start codons, since the presence or absence of conventional start codons in front of at least two orthologs of Ssr1600 is precisely the matter at issue.

One of the most general and powerful methods to identify members of a class is Markov analysis. Before describing it, let me show you Markov analysis in action.

Consider the following excerpts:

> *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.pl* (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.*
>
> *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.*

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

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.

SQ2. Load and run *Hamlet.pl*

III. 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. 3.

To gain insight into how a Markov model is built, download *Hamlet.pl* and *Display-hash.pl* and run each. *Hamlet.pl* outputs the model it creates to the file *model.dat*, and *Display-hash.pl* reads that model and displays it on the screen or (better) saves it in a file. *Display-hash.pl* shows the number of times each combination of letters/symbols appears in the utterances of Hamlet (or whatever text it's given and, for each three letter/symbol combination, the total number instances scored.

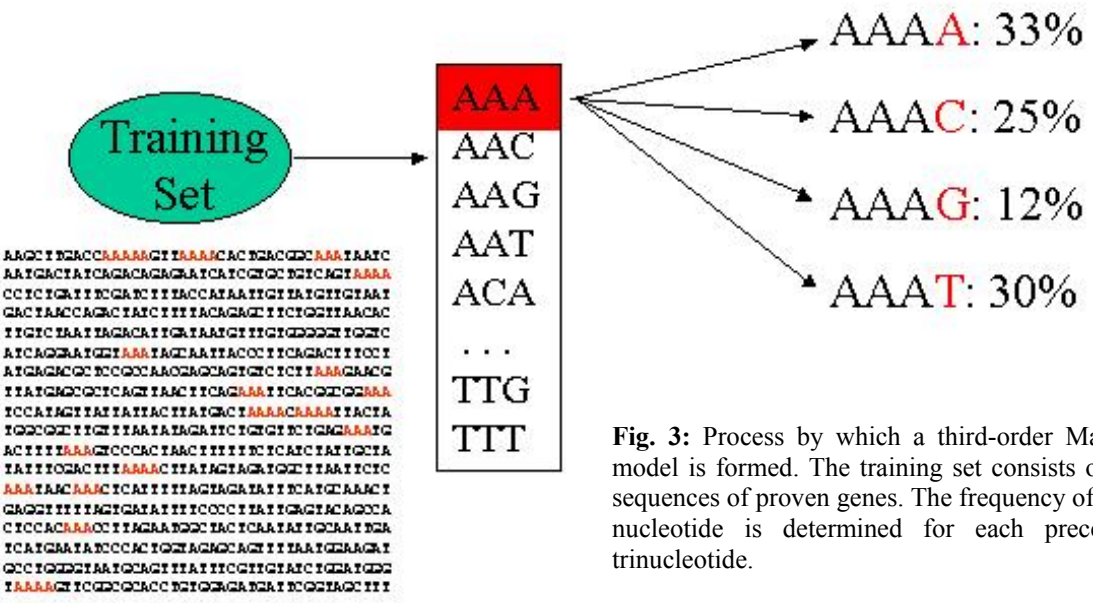

Fig. 3: Process by which a third-order Markov model is formed. The training set consists of the sequences of proven genes. The frequency of each nucleotide is determined for each preceding trinucleotide.

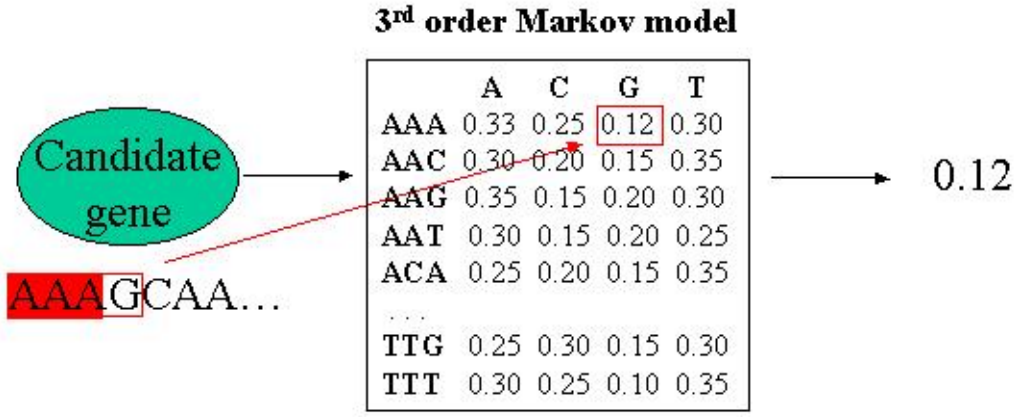

Fig. 4. Use of a Markov model to predict whether a sequence is or is not part of a gene. A four-nucleotide window scans a nucleotide sequence. For each position, the probability of the fourth nucleotide, given the prior three nucleotides, is found in the table constructed from proven genes (**Fig. 5**).

SQ3. Using *Display-hash.pl,* examine the model created by *Hamlet.pl.* Predict which letters will be the most frequent for a three letter/symbol combination and look at the table to see if your predictions pan out.

SQ4. Go into *Hamlet.pl* and change the order from 3rd-order to 2nd-order and notice what effect this change has on the output of the program. Do the same after changing the order to 4th-order.

SQ5. Try your own text as input to *Hamlet.pl.* Who knows? Maybe that's how income-tax instructions are written.

Figure 4 illustrates how a Markov model is used. We will see on Tuesday how these probabilities calculated for each position in a sequence are combined to form a single prediction as to whether a sequence is part of a gene or not. We'll also unpack *Hamlet.pl* and see how it works.