Genetics and Molecular Biology, 26, 4, 529-535 (2003)

www.sbg.org.br

Research Article

Hidden Markov models applied to a subsequence of the *Xylella fastidiosa* genome

Cibele Q. da-Silva

Universidade Federal de Minas Gerais, Departmento de Estatística, Belo Horizonte, MG, Brazil.

Abstract

Dependencies in DNA sequences are frequently modeled using Markov models. However, Markov chains cannot account for heterogeneity that may be present in different regions of the same DNA sequence. Hidden Markov models are more realistic than Markov models since they allow for the identification of heterogeneous regions of a DNA sequence. In this study we present an application of hidden Markov models to a subsequence of the *Xylella fastidiosa* DNA data. We found that a three-state model provides a good description for the data considered.

Key words: DNA, *Xylella fastidiosa*, hidden Markov models.

Received: April 16, 2002; Accepted: June 26, 2003.

Introduction

The rate of sequence data generation in recent years has provided abundant opportunities not only for the development of new approaches to problems in computational biology but also for the exploration of the already known techniques on data that have never been analysed before.

The starting point in most data analysis consists of the use of well established methodology. As the analysis progress, data particularities may require the development of specific tools that are more suitable to better describe and model the data. Creation of new methods requires deep understanding of the current ones, especially when these methods are incredibly powerful and are not as known as they should be due to their mathematical and computational complexity. We consider that hidden Markov Models (HMM) exemplify this notion very well since although these models are not new, we believe that molecular biologists are not aware of the possibilities that these models provide.

Our aim in this study is to discuss dependencies and heterogeneity in DNA data and how they can be appropriately accounted for by the use of HMM. We applied this sort of model to a subsequence of the *Xylella fastidiosa* (Xf) genome as a way to suggest possible analysis for the whole genome.

According to Lambais *et al.* (2000), *Xylella fastidiosa* is a bacteria associated with diseases that cause tremendous losses in many economically important plants, including citrus. *Xylella fastidiosa* is the causal agent of Citrus Variegated Chlorosis (CVC), a disease that affects all commercial sweet orange varieties and that represents a major concern to the Brazilian citrus industry. The plant pathogen attacks citrus fruits resulting in juiceless fruits of no commercial value. *Xylella fastidiosa* is the first plant pathogen to have its genome (the total genetic information stored in the chromosomes of an organism) completely sequenced. In addition, it is probably the least previously studied of any organism for which the complete genome sequence is available.

Data sets generated by sequencing the entire *Xylella fastidiosa* genome pose new challenges since now biologists need quantitative tools and statistical methods to help them to analyze sequences. Some recent publications about *Xylella fastidiosa* signal the need not only for the application of current statistical methods available to analyse its sequenced data but also for statistical research to attack its particularities. Chen *et al*. (2000) analysed sequenced data from 16 strains of *Xylella fastidiosa* originating from nine different hosts. They studied aspects such as sequence heterogeneity in the classification of *X. fastidiosa* at the subspecies level. The studies by Qin *et al.* (2000) and Mehta *et al*. (2001) are concerned with the evaluation of *Xylella fastidiosa* genetic diversity isolated from diseased citrus and coffee in Brazil.

Due to the huge size of the datasets, statistical analyses for the whole genome of many organisms demand the use of high power state of the art computers. That may represent a major problem since we do not have enough available for this purpose.

Send correspondence to Cibele Q. da-Silva. Universidade Federal de Minas Gerais, Departmento de Estatística, Belo Horizonte, MG, Brazil. E-mail: cibeleqs@est.ufmg.br.

In this study we fit hidden Markov models to a dataset of the bacteria *Xylella fastidiosa* genome. Model selection is performed using the Bayesian Information Criterion (BIC) and Akaike's Information Criteria (AIC). In section 2 we talk about dependencies in DNA data. In section 3 we discuss heterogeneity in DNA sequences. Hidden Markov Models are introduced in section 4. In section 5 we briefly introduce AIC and BIC for model selection. Phage lambda and *Xylella fastidiosa* datasets are analysed in section 6.

Dependencies in DNA Data

An obvious first summary of a DNA sequence is just the distribution of the four base types. Although it would be convenient for mathematical modeling if the four bases were equally frequent, almost all empirical studies show an unequal distribution. That means that a simple independence model for DNA sequences have their uses, but only go a little way.

We need to take into account in a model the fact that neighboring bases in DNA sequences are not independent. According to Tavaré and Giddings (1989), associations between adjacent bases will lead to associations between more distant bases and an estimate of how far the relations extend may be found from Markov chain theory.

According to Weir (1996), Markov chain analyses are of use at the genome level, rather than at the level of an individual gene, since the last may involve very short sequences that are not sufficient to demonstrate the presence of higher order chains. The same author observes that it is unlikely that the same Markov chain can describe the whole genome, and if a Markov chain has been fitted to a genome, no biological mechanism is implied, but useful questions can be answered. For example, the frequency of particular subsequences (words) can be predicted.

According to the website http://www.accessexcellence.org/AE/AEC/, in genetic engineering it is common to use the many enzymes that are able to modify or join existing DNA molecules, or to aid in the synthesis of new DNA molecules. For example, the enzyme DNA polymerase makes possible the attachment of two or more DNA molecules to one another. The enzyme DNA ligase breaks DNA molecules into fragments, while the so called restriction endonuclease enzyme (REE) functions by "scanning" the length of a DNA molecule. Once the REE encounters its particular specific recognition sequence (word), it will bond to the DNA molecule and cut it in a predictable and reproducible way. It is important to use Markov chains to help a biologist to estimate the expected number of fragments produced when a specific restriction enzyme is applied to the genome.

Markov chains might describe DNA sequences in terms of their nucleotide composition, *i.e.*, as a string of letters from a four-letter alphabet, $\{A, C, G, T\}$. Let us denote each one of the four base types as *states.* We are going to introduce some terminology and notation useful for Markov chains.

Generally speaking, for a given subject, let X_t denote the response on a categorical variable at time t, t = 0, 1, ..., T. The sequence $(X_0, X_1, X_2, ...)$ is an example of a stochastic process, an indexed family of random variables. In this paper X_t indicates the nucleotide at position *t* in the sequence.

Without invoking any biological mechanism, a Markov chain of order *r* implies that the base present at certain position in a sequence depends only on the bases present at the previous *r* positions. In more formal grounds, a stochastic process is a *rth*-order Markov chain if, for all *t,* the conditional distribution of X_{t+1}, given X_0, ..., X_t, is identical to the conditional distribution of X_{t+1}, given X_t, ..., X_{t-r+1}. Given the states at the previous *r* times, the future behavior of the chain is independent of the past behavior before those *r* times. For a first-order Markov chain with *I* possible states, the conditional probabilities

$$\eta_{ij}(t) = \Pr(X_t = j \mid X_{t-1} = i) \quad (1)$$

with i, j = 1, ..., I are called *transition probabilities*. The extension for higher orders is immediate. If $\eta_{ij}(t)$ does not depend on *t*, the Markov chain is called homogeneous.

Statistical inference for Markov chain uses standard methods of categorical data analysis, such as log-linear models. Some useful references are Anderson and Goodman (1957), Birch (1963), Bishop *et al.* (1975), McCullagh and Nelder (1989), Agresti (1990), and Avery *et al.* (1999).

Heterogeneity in DNA Sequences

Markov chains and log-linear models are important tools to help us describe local properties of DNA sequences. However, Markov chains cannot account for the heterogeneity that may be present in different regions of the same DNA sequence. The basic assumption of this kind of model is that the chain is homogeneous, meaning that the same transition probability matrix is assumed true for the whole sequence being analysed. However, biologists know that coding and non-coding regions of DNA present different nucleotide frequencies. Thus a Markov model would predict some behavior that is not observed in the data. Therefore, this kind of model may be of little practical use in a variety of problems.

An example of heterogeneous DNA is presented by Bernardi and Bernardi (1986). Working with biochemical aspects of DNA, they explain that the nuclear genome of warm-blooded vertebrates exhibits a compositional compartmentalization, in that it consists mainly of a mosaic of very long DNA segments, the isochores. According to the authors, isochores are characterized by fairly homogeneous regions in $C + G$ content, and distinct isochores present distinct proportions of $C + G$. The authors also state that genome does not present very many isochores and that het-

erogeneity within an isochore is very low but is high between isochores. Heterogeneity may be due to differences in patterns of base composition and dependence between neighboring bases, and it might reflect functional and structural differences between regions.

It is possible to describe those heterogeneous unobserved regions of the genome of a given organism by using statistical tools instead of biochemical ones that would then be used more parsimoniously. The referred tools are statistical models that can account for heterogeneity that is present in the sequences. This is the subject of our next discussion.

A Hidden Markov Model for DNA Sequences

In this section we are going to present some hidden Markov models developed by Churchill (1989). These models are still very popular (see Boys *et al.*, 2000). We will make a brief description restating some aspects of section 4 in Churchill (1989). For major details about this issue the referred paper should be consulted.

While the bases *A, C, G, T* represent *observed outcomes* and for short will be denoted *outcomes*, the homogenous unobserved regions we are looking for will be called *hidden states* and for brevity will be denoted *states*. Our job is to estimate how many hidden states there are and to present a map describing where they are located. The number of states is considered to be finite and fixed and corresponds to the different regions of the DNA. We introduce now some notation and definitions needed for describing hidden Markov models for DNA sequences.

Consider a sequence of random variable $\{Y_i: i = 1, ..., n\}$ with distribution determined by a corresponding sequence of unobserved states $\{s_i\}$. Denote the sequence of observed outcomes and states up to time *t* by, respectively, $y^t = \{y_1, ..., y_t\}$ and $s^t = \{s_1, ..., s_t\}$.

Admitting a fixed number of states and multinomial outcomes, let $y_t = (y_{t,0}, ..., y_{t,m-1})$ be a

vector whose components are all zero except for one equal to unity, indicating which of the m possible outcomes is observed. Each observation is associated with one of r states indicated by the vector $s_t = (s_{t,0}, ..., s_{t,r-1})$. There is a vector π_0 of initial probabilities associated to s_1, such that $\Sigma_i \pi_{0i} = 1$. Thus, for the π_{0i}'s there are r - 1 parameters to estimate.

The distribution of y_t given that the state at time t is k is multinomial, that is, $y_t \mid s_{t,k}$ Multinomial$(1, p_{0,k}, ..., p_{m-1,k})$. The parameter $p_{i,k}$ is the probability of observing outcome i when the current state is k, subject to the constraint

$$\sum_{i=0}^{m-1} p_{i,k} = 1 \qquad (2)$$

Therefore, for the p_{ij}'s there are r x (m-1) parameters to estimate.

The *observation equations,* considering independence between the outcomes are

$$P(y_t \mid s_{t,k}) = \prod_{i=0}^{m-1} p_{i,k}^{y_{t,i}} \qquad (3)$$

It is also possible to allow for Markov dependence between the observed outcomes. In the case of first-order dependence, the probability of observing outcome *j* given that the previous outcome was *i* and the current state is *k* is $p_{ij,k}$, where

$$\sum_{j=0}^{m-1} p_{ij,k} = 1 \qquad (4)$$

Therefore for the $p_{ij,k}$'s there are r x m x (m-1) parameters to estimate.

The *observation equations* allowing for first-order dependence are

$$P(y_t \mid y_{t-1}, s_{t,k}) = \prod_{i=0}^{m-1} \prod_{j=0}^{m-1} p_{ij,k}^{y_{t-1,i} \cdot y_{t,j}} \qquad (5)$$

The state process is a Markov chain on the r states. Denote the r x r matrix of state transition probabilities by $\Lambda = (\lambda_{ij})$. Thus, for the λ_{ij}'s there are r x (r - 1) parameters to estimate.

The *system equations* can be written as:

$$P(s_t \mid s_{t-1}) = \prod_{i=0}^{r-1} \prod_{j=0}^{r-1} (\lambda_{i,j})^{s_{t,i} s_{t-1,j}} \qquad (6)$$

The observation and system equations are assumed to be completely specified. The marginal posterior distribution of the state at time *t*, $Pr(s_t \mid y^n)$ is called the smoothed estimate of s_t. Graphic displays of the underlying state process are produced by plotting the smoothed estimates against the sequence index *t*. That represents the mentioned map that describes where the homogeneous regions of the genome are located.

A recursive updating algorithm can be applied as follows to reconstruct the underlying state process. The general filtering and smoothing equations needed in the algorithm are

Filter: to begin, suppose that $Pr(s_{t-1} \mid y^{t-1})$ is known. A prediction of the state at time *t* (predictive equations) can be computed using the system equation

$$P(s_{t,j} \mid y^{t-1}) = \sum_{i=0}^{r-1} \lambda_{i,j} P(s_{t-1,i} \mid y^{t-1}) \qquad (7)$$

and the filtered densities are:

$$P(s_{t,j} \mid y^{t-1}) = \frac{P(y_t \mid s_{t,j}) P(s_{t,j} \mid y^{t-1})}{\sum_{i=0}^{r-1} \lambda_{i,j} P(y_t \mid s_{t,i}) P(s_{t,i} \mid y^{t-1})} \qquad (8)$$

Smoother: the joint distribution of adjacent states is:

$$P(s_{t,i},\ s_{t+1,j}\,|\,y^n) = \frac{P(s_{t+1,j}\,|\,y^n)\,\lambda_{ij}\,P(s_{t,i}\,|\,y^t)}{P(s_{t+1,j}\,|\,y^t)} \qquad (9)$$

and the smoothed estimates are:

$$P(s_{t,i}\,|\,y^n) = P(s_{t,i}\,|\,y^t)\sum_{i=0}^{r-1}\frac{P(s_{t+1,j}\,|\,y^n)\lambda_{i,j}}{P(s_{t+1,j}\,|\,y^t)} \qquad (10)$$

The recursive updating algorithm requires that the parameters of the observation and system equations be specified. The parameter vector $\Theta = \{\pi_0, p, \Lambda\}$ has to be estimated from the data using the EM-algorithm (Dempster, Laird and Rubin, 1977) where the *missing information* is the state at each time $\{s_t, t = 1, ..., n\}$.

The likelihood of the *incomplete data* is

$$P(y^n) = \prod_{t=1}^{n}\sum_{j=1}^{r} P(y_t\,|\,s_{t,j}, y^{t+1})P(s_{t,j}\,|\,y^{t-1}) \qquad (11)$$

and the likelihood of the *complete data* is

$$P(y^n, s^n) = \prod_{t=1}^{n} P(y_t\,|\,s_t, y^{t-1})(s_t\,|\,s^{t-1}) \qquad (12)$$

The closed-form solutions for the likelihood of the complete data are:

$$\hat{p}_{i,j} = \frac{\sum_{t=1}^{n} y_{t,i}s_{t,j}}{\sum_{t=1}^{n} s_{t,j}} \qquad \hat{\lambda}_{i,j} = \frac{\sum_{t=1}^{n} s_{t-1,i}s_{t,j}}{\sum_{t=1}^{n} s_{t-1,i}} \qquad (13)$$

When first order Markov dependence between outcomes is present

$$\hat{p}_{ij,k} = \frac{\sum_{t=1}^{n} y_{t-1,i}y_{t,j}s_{t,k}}{\sum_{t=1}^{n} y_{t-1,i}s_{t,k}} \qquad (14)$$

The initial probability vector is estimated by

$$\hat{\pi}_0 = E(s_1\,|\,y^b) \qquad (15)$$

The EM-algorithm is implemented as follows:

1. Start with an initial guess $\Theta^{(0)}$ of the parameter vector;

2. **E-step**. run the recursive updating algorithm with current parameter estimate $\Theta^{(p)}$. Estimate the states by their conditional expectations

$$E(s_t\,|\,y^n, \Theta^{(p)}) = \Big(P(s_{t,0}\,|\,y^n, \Theta^{(p)}), \ldots, P(s_{t,r-1}\,|\,y^n, \Theta^{(p)})\Big) \qquad (16)$$

M-step. treat the estimated states as data, solve the complete-data likelihood equations to obtain an updated estimate $\Theta^{(p+1)}$.

The recursive updating algorithm is then updated in step (2) above until convergence.

BIC and AIC for Hypothesis Testing

Due to the large size of the *Xylella* subsequence we are working with, traditional Chi-square tests for comparing competing models systematically reject simpler models (the ones with fewer parameters to estimate) in favor of larger ones. That means we need a different methodology to perform our tests. Such methodology was developed by Schwarz (1978) and Sakamato *et al.* (1986), and applied in our tests in section 6.

Sakamoto *et al.* developed Akaike's Information Criterion (AIC) that has the form of a penalized maximum likelihood function where the size of the penalty depends on the number of units required to encode the parameters. Schwarz (1978) developed the Bayesian Information Criterion (BIC) (also known as Schwarz's Bayesian Criterion (SBC)) which is based on Bayesian theory. Raftery (1986), presents a very good description of the BIC.

Let θ be a vector of parameters, λ be the log-likelihood function in study, K as the number of parameters in the model (degrees of freedom), and n be the sample size. Then

$$AIC = -2\ell(\hat{\theta}) + 2K \qquad (17)$$

and

$$BIC = -2\ell(\hat{\theta}) + K\log(n) \qquad (18)$$

When comparing fitted objects, the smaller the AIC (and BIC) the better the fit is. AIC and BIC values have no intrinsic meaning, except in relation to other models based upon the same dataset.

Examples: Bacteriophage Lambda and *Xylella*

In this section we illustrate the application of the hidden Markov models discussed above using a subsequence of the *Xylella* genome and the data for the entire genome of the virus Bacteriophage lambda which has been studied by Churchill (1989). Our codes have been written in FORTRAN and we used the Bacteriophage lambda data to illustrate the methodology and also to check whether our results match Churchill's.

Bacteriophage lambda. The DNA sequence for the Bacteriophage lambda has been acquired from the Genbank website http://www.ncbi.nlm.nih.gov/Genbank/. According to the website http://latin.arizona.edu/~plpweb/lecture/lect39, bacteriophages are basically viruses that specialise in the infection of bacteria. Bacteriophage uses the bacteria Escherichia coli as a host. The bacteriophage lambda ge-

nome size is 48,514 bases and is by far the most completely studied bacteriophage known. A total of 46 genes have been identified on the circular lambda map. The bacteriophage's $G + C$ content has been studied by Skalka *et al*. (1967). Using chemical analyses, they concluded that the genome is composed of six segments with different $G + C$ content. Churchill (1989) found a three-state first-order dependent model as the one that best fits the data. Following Churchill (1989), in Figure 1 we show the smoothed estimates of local $A + T$ composition based on a four-state independent outcome model. Our map is largely in agreement with the one produced by Churchill (1989).

(b) *Xylella* subsequence. According to information obtained from the website http://aeg.lbi.ic.unicamp.br/xf/, the main chromosome of the *Xylella fastidiosa* (cataloged by the National Center for Biotechnology Information (NCBI) by code number AE003849) is composed of 2,838 genes. This chromosome has a total number of 2,679,305 *nucleotide* bases (*A, C, G, T).* Among them, 1,411,300 (52.67%) are *C* or *G.* The website http://aeg.lbi.ic.unicamp.br/xf/ makes available the main chromossome gene map which lists adjacent genes. This gene map has the advantage of being presented in the form of colored horizontal bars, where each color represents the predominant gene function. As we mentioned in the introductory section, we worked with only a subsequence of the genome of *Xylella fastidiosa*. We chose *Xylella* main chromossome subsequence composed of genes XF1141 and XF1196, a total of 38,730 bases. This subsequence is mainly formed by genes that describe three well-defined and contiguous regions. Region 1 starts at base 1 and goes as far as approximately base 7,401. It contains genes related to *energy metabolism*. Region 2 is located between bases 7,402 and 21,723 and it is basically composed of *RNA processing* genes. After a very heterogeneous region including several small genes, each with a particular major function, we find region 3 located between bases 26,767 and 32,514. Region 3 is basically composed of *macromolecule metabolism* genes. Therefore we are aware of the existence of three functionally speaking different regions (energy metabolism, RNA processing and macromolecule metabolism). That may have an impact on the dependence structure between nearby bases. If that is so, hidden Markov models can help us to locate them.

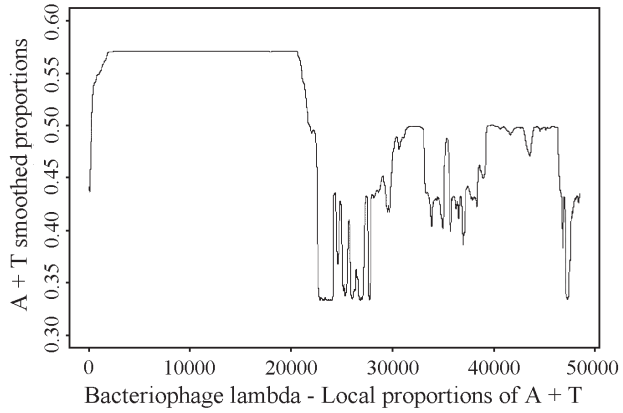

Figure 1 - Plot of the local proportions of A+T based on a four-state model fit by maximum likelihood.

We fitted independent or first-order dependent outcomes and up to four latent states. Table 1 summarizes the results. BIC and AIC (see Schwarz, 1978 and Sakamoto *et al.*, 1986) pointed out a three-state independent outcome model as the one providing the best fit.

Figure 2 shows the smoothed estimates of local $C + G$ composition based on a three-state independent outcome model . The local composition was estimated as:

$$\hat{\pi}_t = \sum_{i=1}^{3} P(y_{t,1} \mid s_{t,1}) P(s_{t,1} \mid y^n) \qquad (19)$$

In (19) $y_{t,1}$ implies we are dealing with local proportion of $C + G$ - there are only two outcomes: 1 (for C or G) or 2 (for A or T).

In Figure 2 we notice that there are four regions. The first and the last one present comparable $C + G$ smoothed proportions.

The second region is the largest and presents the largest $C + G$ concentration. Region three is narrow and spiked. Approximately first region ranges from base 1 to 7,000, the second region goes up to base 21,000, and the third goes from base 22,000 to 24,000. Regions one and two are in very good agreement with our data description, matching *energy metabolism* and *RNA processing* genes. Region three incorporates, functionally speaking, very heterogeneous data, so it is not possible to characterize it in those

Table 1 - Summary statistics used for selecting the best HMM model for the *Xylella* dataset.

Hypothesized number of hidden states	Order of dependence of the outcomes	Degrees of freedom	Likelihood	BIC	AIC
2	0	1 + 2 + 2 = 5	-26795.9	53644	53601
3	0	2 + 6 + 3 = 11	-26749.9	53616	53521
4	0	3 + 12 + 4 = 19	-26745.8	53792	53529
2	1	1 + 2 + 4 = 7	-26795.3	53664	53604
3	1	2 + 6 + 6 = 14	-26795.2	53738	53618
4	1	3 + 12 + 8 = 23	-26739.6	53833	53636

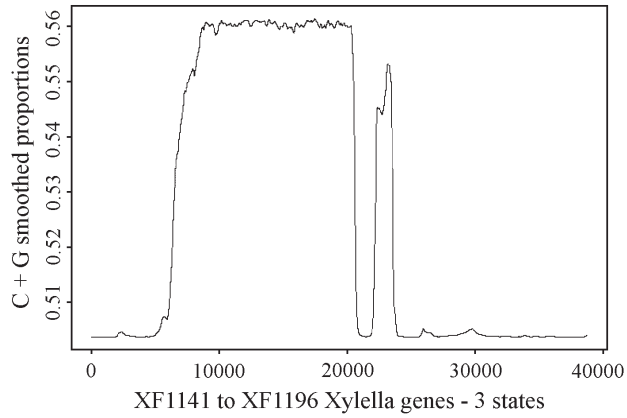

Figure 2 - Plot of the local proportions of C+G based on a three-state model fit by maximum likelihood.

terms. Region four is well defined with a regular low expression of $C + G$. It does not contradict the data description in the sense that approximately from bases 26,000 to 32,000 the smoothed proportions show homogeneity which matches the location of *macromolecule metabolism* genes. These results confirm that hidden Markov models are useful tools to reveal homogeneous regions of DNA data.

We notice that there are non negligible differences in scale of the maps obtained for the studied organisms. We can observe that the homogeneous regions are separated more clearly in the case of the phage. This might be due to the weaker kind of dependencies in the *Xylella* data we worked with compared to the phage data. Despite the fact that the best model was found to be a three state one, model selection procedures did not support first order dependencies among *Xylella* outcomes in contrast with the phage outcomes. Therefore, it is reasonable that the phage map presents a better discrimination of homogeneous regions than the *Xylella* map.

Finally, even though we could tell the major function of the genes included in the *Xylella* dataset we used by inspecting the chromossome map (website http://aeg.lbi.ic.unicamp.br/xf/), this is no longer feasible when dealing with the whole genome of the *Xylella* genome. Thus, computational methods are needed to extract and summarize major underlying features that help the analyst to understand DNA structure and function. Hidden Markov models seem to be a good option.

Discussion

Hidden Markov models are useful for describing and revealing some special features of temporal biological series. The main advantage over the regular Markov models is the possibility that HMMs have of accounting for heterogeneity that may be present in the data. As a result, more sensible models and better data descriptions might be available to the analyst.

In this work we applied HMM methods to a subsequence of the *Xylella fastidiosa* genome. In order to describe possible variations in the expression of $C + G$, we worked with the $C + G$ sequence data instead of with the original DNA sequence. The HMM that better describes the data was able to correctly discriminate regions in the data corresponding to distinct biological functions. In the future, this kind of model may be used in the study of larger subsequences or even the whole *Xylella* genome.

There are other already known uses of HMM in computational biology. Those may be tried on the *Xylella* data. For example, HMMs are used for obtaining multiple aligned sequences and also for gene detection.

According to Hugley and Krogh (1996), HMM are a highly effective means of modeling a family of unaligned sequences or a common motif within a set of unaligned sequences. HMMs are particularly useful in the study of protein molecules since they make possible the automatic discrimination of evolutionarily close proteins. Proteins are built from an alphabet of twenty smaller molecules known as amino acids. According to http://www.cse.ucsc.edu/research/compbio/ismb99.handouts/KK185FP.html, when a cell reproduces, a protein inside the cell is most of the time exactly duplicated in the daughter cell. However, over long periods of time, errors occur in the copy process. When this happens, a protein in the daughter cell is slightly different from its counterpart in the parent. The three most common errors are “substitution” of an amino acid in a given position, “insertion” of one or more new amino acids, and “deletion” of one or more amino acids. As a result of these errors, proteins which share a common ancestor are not exactly alike. However, they inherit many similarities in primary structure from their ancestor. This is known as “conservation” of primary structure in a protein family. In an HMM for multiple alignment the states are S (substitution), I (insertion), D (deletion) and M (matching of amino acids). According to Krogh *et al*. (1994), an HMM used for multiple alignment identifies a set of positions that describes the conserved primary structure in the sequences from a given family of proteins, *i.e.*, the model identifies the core elements of homologous proteins. In the case of the *Xylella* it may be important to know how evolutionarily close is this organism to another, since similar methods used in the combat of the latter could be tried.

HMMs are also a very useful tool in gene prediction. According to Stormo (2000), the states correspond to exons, introns, and any other class of sequences desired (such as 5’ and 3’ UTRs, promoters regions, intergenic regions, repetitive DNA, etc). These genetic entities signal the points in the DNA where a new gene starts or ends. The probability of changing from an intron to an exon depends on the local sequence such that it is high only at plausible splice junctions. Stormo (2000) explains that the “hidden” in these HMMs denotes that fact that we see only the DNA sequence directly, and the state that generated the sequence

(exon, intron, etc) is not visible. These methods are important because genes can be located using computational biology tools.

Acknowledgments

We would like to thank the anonymous referees for valuable comments and suggestions on this study.

References

Agresti A (1990) Categorical Data Analysis. 1st ed. John Wiley & Sons, New York.

Anderson TW and Goodman LA (1957) Statistical Inference About Markov Chains. Annals Math Statist 28:89-110.

Arnold J, Cuticchia AJ, Newsome DA, Jennings III, WW and Ivarie R (1988) Mono-through hexanucleotide composition of the sense strand of yeast DNA: A Markov chain analysis. Nucleic Acids Res 16:7145-7158.

Avery PJ (1987) The analysis of Intron data and their use in the detection of short signals. Journal of Molecular Evolution 26:335-340.

Avery PJ and Henderson DA (1999) Fitting Markov chain models to discrete state series such as DNA sequences. Applied Statistics 48, Part 1:53-61.

Bernardi G and Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1-11.

Bishop YMM, Fienberg SE, Holland PW, Mosteller F and Light R (1975) Discrete Multivariate Analysis: Theory and Practice. 1st ed. The MIT Press, Massachusetts.

Boys RJ, Henderson DA and Wilkinson DJ (2000) Detecting homogeneous segments in DNA sequences by using hidden Markov models. Applied Statistics 49 Part 2:269:285.

Chen J, Jarret RL, Qin X, Hartung JS, Banks D, Chang CJ and Hopkins DL (2000) 16S rDNA sequence analysis of *Xylella fastidiosa* strains. Systematic and Applied Microbiology 23(3):349-354.

Churchill GA (1989) Stochastic models for heterogenous DNA sequences. Bulletin of Mathematical Biology 51(1):79-94.

Guilhabert MR, Hoffman LM, Mills DA and Kirkpatrick BC (2001) Transposon mutagenesis of *Xylella fastidiosa* by electroporation of Tn5 synaptic complexes. Molecular Plant-Microbe Interactions 14(6):701-706.

Hughey R and Krogh A (1996) Hidden Markov models for sequence analysis: extensions and analysis of the basic method. CABIOS 12(2):95-107.

Krogh A, Brown M, Mian S, Sjolander K and Haussler D (1994) Hidden Markov models in computational biology. Recent Methods for RNA Modeling Using Stochastic Context-Free Grammars. CPM: 289-306

Lambais MR, Goldman MHS, Camargo LEA and Goldman GH (2000) A genomic approach to the understanding of *Xylella fastidiosa* pathogenicity. Current Opinion in Microbiology 3(5):459-462.

McCullagh P and Nelder JA (1989) Generalized Linear Models. 2nd ed. Chapman & Hall, London.

Mehta A, Leite RP and Rosato YB (2001) Assessment of the genetic diversity of *Xylella fastidiosa* isolated from citrus in Brazil by PCR-RFLP of the 16S rDNA and 16S-23S intergenic spacer and rep-PCR fingerprinting. Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology 79(1):53-59.

Qin XT, Miranda VS, Machado MA, Lemos EGM and Hartung JS (2000) An evaluation of the genetic diversity of *Xylella fastidiosa* isolated from diseased citrus and coffee in Sao Paulo, Brazil. Phytopathology 91(6):599-605.

Raftery A (1986) Choosing Models for Cross-classifications, Amer Sociol Rev 51:145-146.

Raftery A (1995) Bayesian Model Selection in Social Research (with Discussion). University of Washington Demography Center Working Paper . 94-12. A revised version appeared in Sociological Methodology 1995, pp 111-196.

Raftery A and Tavare S (1994) Estimation and Modelling Repeated Patterns in High Order Markov Chains with the Mixture Transition Distribution Model, Applied Statistics 43(1):179-199.

Sakamoto Y, Ishiguro M and Kitagawa G (1986) Akaike Information Criterion Statistics, D. Reidel Publishing Company.

Schwarz G (1978) Estimating the dimension of a model. Annals of Statistics 6:461-464.

Skalka A, Burgi E and Hershey AD (1968) Segmental distribution of nucleotide sequence of Bacteriophage lambda DNA. J Mol Biol 162:729-773.

Stormo GD (2000) Gene-finding approaches for eukaryots. Genome Research 10:394-397.

Tavaré S and Giddings BW (1989) Some statistical aspects of the primary structure of nucleotide sequences. In: Waterman MS (ed) Mathematical Methods for DNA Sequences. CRC Press, Boca Raton, FL, pp 117-132.

Weir BS (1996) Genetic Data Analysis II, 1st. ed., Sinauer Associates, Inc., Sunderland.

Editor: Darcy Fontoura de Almeida