Applications of hidden Markov models in Bioinformatics

Richard J Boys
Newcastle University, Department of Statistics
Newcastle upon Tyne
U.K.
Richard.Boys@ncl.ac.uk

1. Introduction

Biological science is now one of the major areas of focus of scientific research worldwide. The increasing amounts of data being produced often require sophisticated techniques to determine biological functions and mechanisms. Many statisticians and computer scientists are now working with biologists at the interface between these subjects, which has become known as "Bioinformatics". This paper describes some methods that have been developed recently for the analysis of biological sequence data which are based on hidden Markov models (HMMs).

Hidden Markov models provide a natural and flexible framework for describing heterogeneous time series, and are therefore ideal for modelling biological sequence data; see MacDonald and Zucchini (1997). These models are defined by two connected processes, one describing the evolution (in discrete time) of the observed process $\{Y_t : \ t \in \mathbb{N}\}$ and the other describing that of the hidden process $\{S_t : \ t \in \mathbb{N}\}$. The hidden process is an integer-valued Markov chain with state space $\{1, 2, \ldots, r\}$ and the observed process evolves conditionally on this hidden process. In many engineering applications such as speech recognition (Rabiner, 1989), the random variables Y_t are conditionally independent given their hidden counterpart S_t. However, some bioinformatics applications benefit from more complex dependence structures in order to account for the observed correlations in biological sequences.

Molecular biologists now commonly try to identify genes by using HMM-based gene-finding algorithms; see, for example, the `Veil` algorithm by Henderson *et al.* (1997). Such algorithms typically adopt a pattern-based approach, that is, model genes as a large number of homogeneous components (patterns). These DNA patterns are described by models with known parameters obtained from large training samples of known genes. The location of genes is modelled by the hidden Markov chain and the predicted gene structure is determined by the Viterbi algorithm (Forney, 1973), a dynamic programming method to find the optimal state sequence. Extensions that use semi-Markov models, which explicitly model (hidden) pattern lengths, such as the `Genscan` algorithm (Burge and Karlin, 1997) and the `Fgenesh` algorithm (Salamov and Solovyev, 2000), have also proved very successful in identifying candidate genes.

Another activity which has employed HMMs is the search for regulatory regions in DNA. These regions are important as they are areas of DNA to which specific proteins bind in order

to control the regulation of proteins produced by a cell. However, locating these regions is a challenging task due to the shortness of the protein binding elements (motifs), typically 5–20 base pairs long. Strategies have been developed using HMMs to find known elements (Crowley *et al.*, 1997) and to locate new (unknown) motifs of known length (Liu, Neuwald and Lawrence, 1995). For example, Crowley *et al.* (1997) identify the positions of protein binding elements using a reference catalogue of "words". The DNA sequence is divided into short intervals (equal to the length of the longest element) and their method employs data in the form of binary variables $X_{ij} = 1$ if word j is found in interval i, and 0 otherwise. The probabilities $Pr(X_{ij} = 1)$ are then modelled as functions of the hidden layer, this layer indicating whether or not the region is considered to be a regulatory region. The HMM attempts to locate correct sightings of the motif from the many spurious matches. Although Liu, Neuwald and Lawrence (1995) use a similar hidden Markov structure, they model motifs by a set of independent base probabilities which depend on position (in the motif) and predict the location of motifs by contrasting these probabilities with those in the "background" sequence, which itself may follow an HMM pattern. Both these methods adopt a Bayesian approach facilitated by modern computational Markov chain Monte Carlo (MCMC) techniques.

A common theme in the methods described thus far is the need for models which capture the inherent heterogeneity in base composition; for instance, when modelling the "background" in the search for motifs. The next section describes how HMMs can be used in this context and, in particular, outlines a Bayesian approach which permits inferences for the number (r) of homogeneous segments and the order (q) of Markov dependence within these segments.

2. Locating regions of unknown structure

Suppose we observe a DNA sequence $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ of unknown origin and composition. HMM-based procedures can be used to discover whether the sequence contains any homogeneous patterns and where they are located; see, for example, Churchill (1989) and Muri (1998). We assume that base updates within each pattern type evolve as a qth order Markov chain, with for $t = q+1, \ldots, n$

$$\Pr(Y_t|\boldsymbol{S}, \boldsymbol{Y} \setminus \{Y_t\}) = \Pr(Y_t = j|S_t = k, Y_{t-q} = i_1, Y_{t-q+1} = i_2, \ldots, Y_{t-1} = i_q)$$

where $i_1, i_2, \ldots, i_q, j \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and $k \in \{1, 2, \ldots, r\}$, and denote the collection of these probabilities (conditional on r and q) by $\mathcal{P}$. We also assume a first order (hidden) Markov chain for the segment label process, with for $t = 2, 3, \ldots, n$

$$\Pr(S_t|\boldsymbol{S} \setminus \{S_t\}) = \Pr(S_t = j|S_{t-1} = i)$$

where $i, j \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, and denote the collection of these probabilities (conditional on r) by Λ.

We shall adopt a Bayesian approach which allows us to incorporate prior information about likely segment lengths and possibly different base transition patterns. For given r and q,

it is convenient to express the prior information for probability transitions as a collection of independent Dirichlet distributions. This choice permits a tractable analysis which caters for the unknown segmentation by using standard MCMC methods with data augmentation to produce realisations from the posterior distribution $\pi(\mathcal{P}, \Lambda, \boldsymbol{S}|\boldsymbol{Y}, r, q)$; see Boys *et al.* (2000) for more details. Inferences about the dimensionality parameters r and q can be made by augmenting this MCMC procedure using techniques such as reversible jump algorithms which allow the underlying Markov chain to jump between subspaces of different dimension; see Richardson and Green (1997). These techniques have recently been extended to the more complicated HMM context by Robert *et al.* (2000), who illustrate their general procedure using an HMM with zero mean Gaussian observables. They describe *split/combine* and *birth/death* moves for updating a "number of components" parameter which are suitable for the parameter r. Updates are made by matching moments between the two subspaces. We have adapted these moves to the model described above but, unfortunately, found that they exhibited very low acceptance rates which, in turn, resulted in poor mixing of the Markov chain. An alternative approach is suggested by the work of Viallefont *et al.* (2001) on Poisson mixture models. We have found that moves similar to these can overcome the problems of poor mixing associated with the more complicated split/combine strategies above. Determining moves with good mixing properties for updating q are not so straightforward. We have found little success when using moves based on matching joint and marginal probabilities. The search for more successful strategies is the subject of on-going research. Further details on the latest results can be found in Boys and Henderson (2001).

REFERENCES

Boys, R.J. and Henderson, D.A. (2001). A Bayesian approach to DNA sequence segmentation using reversible jump MCMC methods. Statistics Preprint STA01.3, Newcastle University, U.K.

Boys, R.J., Henderson, D.A. and Wilkinson, D.J. (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Appl. Statist.*, **49**, 269-285.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78-94.

Crowley, E.M., Roeder, K. and Bina, M. (1997). A statistical model for locating regulatory regions in genomic DNA. *J. Mol. Biol.* **268**, 8-14.

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

Forney, G.D. (1973). The Viterbi algorithm. *Proc. IEEE*, **61**, 268-278.

Henderson, J., Salzberg, S. and Fasman, K. (1997). Finding genes in DNA with a hidden Markov model. *J. Comp. Biol.*, **4**, 127-141.

Liu, J.S., Neuwald, A.F. and Lawrence, C.E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist. Assoc.*, **90**, 1156-1170.

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

Muri, F. (1998). Modelling bacterial genomes using hidden Markov models. In *COMPSTAT '98* (Eds. R.W. Payne and P.J. Green), pp. 89-100. Physica.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257-285.

Richardson, S. and Green, P.J. (1997). On Bayesian mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. B*, **59**, 731-792.

Robert, C.P., Rydén, T. and Titterington, D.M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. Roy. Statist. Soc. B*, **62**, 57-75.

Salamov, A. and Solovyev, V. (2000). Ab initio gene finding in Drosophila genomic DNA. *Genom. Res.*, **10**, 516-522.

Viallefont, V., Richardson, S. and Green, P.J. (2001). Bayesian analysis of Poisson mixtures. *J. Nonpara. Statist.*, to appear.

RESUME

Les HMM (Hidden Markov models, ou modèles de Markov cachés) sont maintenant largement utilisés pour traiter des problèmes complexes en bio-informatique. Dans ce papier, nous présentons brièvement plusieurs applications des HMM dont le but est de localiser et de décrire les structures dans une séquence de données ADN. Nous discutons également de techniques pour analyser l'hétérogénéité de composition, puis nous concentrons sur les méthodes Bayesienne pour déterminer le nombre de segments homogènes ainsi que l'ordre de dépendance de Markov au sein de ces segments.