



## Bayesian hidden Markov model for DNA sequence segmentation: A prior sensitivity analysis

Darfiana Nur<sup>a,\*</sup>, David Allingham<sup>b</sup>, Judith Rousseau<sup>c</sup>, Kerrie L. Mengersen<sup>b</sup>, Ross McVinish<sup>d</sup>

<sup>a</sup> School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia

<sup>b</sup> ARC Centre of Excellence for Complex Dynamic Systems and Control, University of Newcastle, Callaghan, NSW 2308, Australia

<sup>c</sup> CEREMADE, Université Paris-Dauphine, Place du maréchal de Lattre de Tassigny, Paris 75016, France

<sup>d</sup> School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD 4001, Australia

### ARTICLE INFO

#### Article history:

Available online 8 July 2008

### ABSTRACT

The sensitivity to the specification of the prior in a hidden Markov model describing homogeneous segments of DNA sequences is considered. An intron from the chimpanzee  $\alpha$ -fetoprotein gene, which plays an important role in embryonic development in mammals, is analysed. Three main aims are considered: (i) to assess the sensitivity to prior specification in Bayesian hidden Markov models for DNA sequence segmentation; (ii) to examine the impact of replacing the standard Dirichlet prior with a mixture Dirichlet prior; and (iii) to propose and illustrate a more comprehensive approach to sensitivity analysis, using importance sampling. It is obtained that (i) the posterior estimates obtained under a Bayesian hidden Markov model are indeed sensitive to the specification of the prior distributions; (ii) compared with the standard Dirichlet prior, the mixture Dirichlet prior is more flexible, less sensitive to the choice of hyperparameters and less constraining in the analysis, thus improving posterior estimates; and (iii) importance sampling was computationally feasible, fast and effective in allowing a richer sensitivity analysis.

© 2008 Elsevier B.V. All rights reserved.

### 1. Introduction

Many genome sequences display heterogeneity in base composition in the form of segments of similar structure. A number of statistical techniques have been developed to identify these homogeneous DNA segments, as reviewed in Braun and Müller (1998). One technique, proposed in Churchill (1989), describes DNA sequence structure using a hidden Markov model (HMM) which is, in essence, a mixture model with Markov-dependent component indicators (MacDonald and Zucchini, 1997). Sequence analysis using HMMs is now a standard approach (Durbin et al., 1998) in the comparatively young science of bioinformatics and is a fundamental component of many gene-finding algorithms which identify and delineate genes in the human and other genomes (De Fonzo et al., 2007).

Bayesian inference procedures and algorithms have revolutionized the field of computational biology (Liu and Logvinenko, 2003) due to the development of computationally-intensive simulation-based methods such as Markov chain Monte Carlo (MCMC), which are available in software such as WinBUGS (Lunn et al., 2000), and has led to the adoption of increasingly complex models in many situations.

A sometimes controversial aspect of the Bayesian approach is the need to specify prior distributions for the unknown parameters. In certain situations these priors may be very well defined. However, for complex models with many

\* Corresponding author. Tel.: +61 2 49215547; fax: +61 2 49216898.

E-mail addresses: [Darfiana.Nur@newcastle.edu.au](mailto:Darfiana.Nur@newcastle.edu.au) (D. Nur), [David.Allingham@newcastle.edu.au](mailto:David.Allingham@newcastle.edu.au) (D. Allingham), [rousseau@ceremade.dauphine.fr](mailto:rousseau@ceremade.dauphine.fr) (J. Rousseau), [k.mengersen@qut.edu.au](mailto:k.mengersen@qut.edu.au) (K.L. Mengersen), [r.mcvinish@qut.edu.au](mailto:r.mcvinish@qut.edu.au) (R. McVinish).

parameters, the choice of priors and conclusions of the subsequent Bayesian analysis are usually validated through a prior sensitivity analysis, as presented here.

For DNA sequence segmentation, a DNA sequence can be thought of as the observed process which evolves independently or dependently given an unobserved Markov chain which locates the position of the segment types. The parameters in this model are the base (nucleotide) transition probabilities for the segment types and the transition matrix of segment types. [Boys et al. \(2000\)](#) presented a Bayesian solution to the segmentation problem using HMMs when the number of segments is known. These results were generalised in [Boys and Henderson \(2004\)](#) to the case in which the number of segments is unknown. In [Boys et al. \(2000\)](#) and [Boys and Henderson \(2004\)](#), the prior knowledge for base transition probabilities in each segment was weak but the prior beliefs about the transition matrix for the segment types were strong. The authors discussed briefly the sensitivity of their conclusion to the choice of prior, especially for the transition matrix for the segment types, but no details were given. Their articles raise fundamental questions about limitations in model specification and bring to the forefront the issue of how far one can refrain from making prior assumptions about a model while keeping it feasible in practice. This prompts the important question of the impact of these priors on resultant inferences.

This paper has three main aims. The primary aim is to undertake a sensitivity analysis of the priors of a Bayesian hidden Markov model for DNA sequence segmentation. We employ Markov chain Monte Carlo via a short and easy-to-use program in BRUGS (“Bayesian analysis using Gibbs Sampler in R”). The sensitivity analysis includes a traditional approach, varying the prior distributions for base transition probabilities for each segment type and for the transition matrix of segment types. A sequence of Dirichlet priors is considered for the former and Dirichlet and mixture Dirichlet priors for the latter. The second aim of this paper is to introduce an alternative approach to sensitivity analysis that employs importance sampling of an MCMC chain obtained from the traditional approach. Our focus is on the feasibility and computational efficiency of this approach for comparing a large number of priors simultaneously in a more comprehensive sensitivity analysis. The results are applied to the segmentation of a benchmarking DNA sequence, intron 7 of the chimpanzee  $\alpha$ -fetoprotein gene.

## 2. Methods

### 2.1. The hidden Markov model

A DNA sequence  $\mathbf{y} = y_1, y_2, \dots, y_n$  can be considered as a realisation of a random process  $Y_1, Y_2, \dots, Y_n$  where  $Y_t \in \{a, c, g, t\}$ ,  $t = 1, 2, \dots, n$ , represent the four nucleotides adenine, cytosine, guanine and thymine, respectively, and  $n$  represents the length of the sequence. For convenience, the data can be encoded as 1, 2, 3, 4 for  $a, c, g$  and  $t$ , respectively. Suppose that there are at most  $r$  types of homogeneous segment within the DNA sequence. The (hidden) segment type at location  $t$  will be denoted by  $S_t \in \{1, 2, \dots, r\}$  for  $t = 1, 2, \dots, n$ .

Assume that transitions between bases,  $Y_{t-1} \rightarrow Y_t$ , follow a first-order Markov chain, where the choice of transition matrix is determined by the hidden state  $S_t$ . Following [Boys et al. \(2000\)](#), we denote the  $4 \times 4$  transition matrices for each segment type by  $\mathcal{P} = \{P^{(1)}, P^{(2)}, \dots, P^{(r)}\}$ , where  $P^{(k)} = (P_{ij}^{(k)})$ . The update equations for the base transitions are

$$\begin{aligned} P(Y_t = y_t | S_t = s_t, y_1, \dots, y_{t-1}, \mathcal{P}) &= P(Y_t = y_t | S_t = s_t, y_{t-1}, \mathcal{P}) \\ &= P_{y_{t-1}y_t}^{(s_t)}. \end{aligned} \quad (1)$$

The hidden state process of segment types is assumed to be a homogeneous first-order Markov chain with  $r \times r$  transition matrix  $\Lambda = (\lambda_{ij})$  such that

$$P(S_t = s_t | s_1, \dots, s_{t-1}, \Lambda) = P(S_t = s_t | s_{t-1}, \Lambda) = \lambda_{s_{t-1}s_t}. \quad (2)$$

Assuming that  $Y_1$  and  $S_1$  follow independent discrete uniform distributions and by using (1) and (2), the likelihood for the model parameters  $\mathcal{P}$  and  $\Lambda$ , given the observed DNA sequence  $\mathbf{y}$  and the hidden segment types  $\mathbf{s}$ , is

$$L(\mathcal{P}, \Lambda | \mathbf{y}, \mathbf{s}) = \frac{1}{4^r} \prod_{t=2}^n P_{y_{t-1}y_t}^{(s_t)} \lambda_{s_{t-1}s_t}.$$

The posterior distribution for the model parameters  $\mathcal{P}$  and  $\Lambda$  and unobserved segment types  $\mathbf{s}$  can be obtained by using Gibbs sampling with data augmentation. Let  $\pi(\mathcal{P}, \Lambda | \mathbf{y}, \mathbf{s})$  be the posterior distribution of the parameters. Given the multinomial form of the likelihood function, we adopt a conjugate Dirichlet prior distribution, described in more detail below. Combining the likelihood with these priors using Bayes' theorem produces independent Dirichlet distributions for the rows of the transition matrices.

### 2.2. The priors

A choice of priors is available for the base transition probabilities for the segment types and the transition matrix of segment types.

A typical row of a base transition matrix and a row of the segment type transition matrix are denoted by  $\mathbf{p}_i = (p_{ij})$  and  $\lambda_k = (\lambda_{kj})$ , respectively. Given that the likelihood is multinomial, the conjugate prior distributions for  $\mathbf{p}_i$  and  $\lambda_k$

متن کامل مقاله

دریافت فوری ←

**ISI**Articles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات