Computational and Structural Biotechnology Journal 10 (2014) 107–115

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Investigating genomic structure using *changept*: A Bayesian segmentation model

Manjula Algama, Jonathan M. Keith *

School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia

ARTICLE INFO

Available online 27 August 2014

Keywords:
Sequence segmentation
Bayesian modelling
Generalised Gibbs sampler
Conservation levels
GC content
Non-coding RNA

ABSTRACT

Genomes are composed of a wide variety of elements with distinct roles and characteristics. Some of these elements are well-characterised functional components such as protein-coding exons. Other elements play regulatory or structural roles, encode functional non-protein-coding RNAs, or perform some other function yet to be characterised. Still others may have no functional importance, though they may nevertheless be of interest to biologists. One technique for investigating the composition of genomes is to segment sequences into compositionally homogenous blocks. This technique, known as 'sequence segmentation' or 'change-point analysis', is used to identify patterns of variation across genomes such as GC-rich and GC-poor regions, coding and non-coding regions, slowly evolving and rapidly evolving regions and many other types of variation. In this mini-review we outline many of the genome segmentation methods currently available and then focus on a Bayesian DNA segmentation algorithm, with examples of its various applications.

Contents

* Corresponding author. Tel.: +61 3 990 20890; fax: +61 3 9905 4403.
E-mail address: jonathan.keith@monash.edu (J.M. Keith).

http://dx.doi.org/10.1016/j.csbj.2014.08.003
2001-0370/

1. Role of genome segmentation

Identifying the distinct components of the human and other genomes is a core task in current bioinformatics, and a necessary pre-requisite to a full understanding of the connections between genomes and phenotypes. Yet the annotation of complex eukaryotic genomes is still far from complete. Even the proportion of the genome that performs biological functions is still hotly debated, with estimates varying from 5% [1] to 80% [2]. Whatever the true figure may be, it is clear that a vast amount of the biology underlying the structure of genomes remains to be discovered. Bioinformatics has an important role to play in this endeavour, and one of its tasks is to identify segments of the genome representing elements that require annotation.

2. Segmentation methods

Several techniques have been developed to analyse variation in properties of interest across a genome and to provide clues to the nature of its components. In this article we review some of the most widely used segmentation methods and discuss the main ideas behind each technique.

2.1. Sliding window analysis

Although not technically a segmentation method, 'sliding window analysis' is the most commonly used way to profile variation in a property of interest across a genome. This technique involves averaging the property of interest over a sliding window of a predetermined length along the sequence. For example if the window size is 10, the first point is obtained by averaging the property of interest over nucleotides 1–10, the second point is the average over nucleotides 2–11, and so on. Determining the window size can be crucial: a smaller window allows for a more precise localisation of changes, however this can increase the noise. Tajima in 1991 has proposed an algorithm to determine window size [3]. The main drawback of the sliding window analysis is that it does not identify boundaries where statistically significant changes to the property in question occur. To avoid some of the disadvantages of the sliding window approach, a windowless technique based on the Z curve was introduced to analyse GC content of genomic sequences [4]. This method enables calculation of GC content at any resolution, even at a base position. Some applications of the sliding window analysis can be found in papers [5–16].

2.2. Hidden Markov models

More precise segmentation methods have been developed to identify homogenous segments as well as the locations (change-points) at which sharp changes in a particular property of interest occurs. Hidden Markov models (HMMs) are one approach capable of inferring segment boundaries. The HMM methodology is well-established, dating from the 1950s [17]. In these models, the observed sequence is considered to be composed of segments, with the sequence of each segment generated by a Markov process. The transition probabilities for each segment are determined by a hidden state, and transitions between hidden states occur at segment boundaries. The sequence of hidden states is also modelled as a Markov process. A key parameter of an HMM is the *order* of the Markov chain, that is, the number of preceding sequence positions required to condition the transition probabilities of the observed sequence. This is unknown a priori, and usually needs to be specified, although some approaches are able to infer the order, or determine it adaptively.

HMMs were first used in biological sequence analysis by Churchill [18,19]. The parameters of the model, including segment boundaries, were estimated by using the maximum likelihood method based on the expectation–maximisation (EM) algorithm [20]. HMMs have since been widely used for sequence analysis problems in bioinformatics, and an extensive literature now exists. Two important developments were the 1998 GeneMark.hmm algorithm which used an HMM to find exact gene boundaries [21] and an HMM developed by Peshkin and Gelfand in 1999 to segment yeast DNA sequences [22]. Some other important examples are included in [23–29]. The Sarment package of Python modules built by Gueguen for easy building and manipulation of sequence segmentations uses both sliding window and HMM methods [30].

HMM models have also been implemented from a Bayesian perspective. One advantage of adopting a Bayesian approach is that it provides quantification of the uncertainties in parameter estimates in the form of probability distributions. In fact, one can dispense with point estimates of parameters altogether, instead reporting marginal distributions for key parameters, such as the locations of change-points. Boys et al. in 2000 presented a Bayesian method of segmentation using HMMs when the number of segments is known [31] and later generalised this method for an unknown number of segments [32]. In 2006, the segmentation method developed by Kedzierska and Husmeier was a combination of the sliding window analysis and the Bayesian HMM [33]. Nur and co-workers in 2009 performed sensitivity analysis on priors used in the Bayesian HMM to show the impact of prior choice on posterior inference [34]. One challenge for Bayesian HMM approaches is that they are computationally intensive and are typically infeasible for segmenting large-scale sequences, without simplifying heuristics.

2.3. Multiple change-point analysis

This approach arose independently of HMMs, and has an extensive literature dating back to the 1970s [35,36]. Change-point analysis differs from HMMs in that it typically assumes no Markov dependence in either the observed sequence or the underlying sequence of hidden states. In this sense change-point models are simpler than HMMs, and have fewer parameters. However, the two types of analysis are clearly related, and it may be useful to think of change-point models as zeroth order HMMs. A key advantage of change-point models, due to their simplicity, is their reduced computational burden, a point which is of particular relevance when implementing them within a Bayesian framework.

The use of multiple change-point models in bioinformatics was pioneered by Liu and Lawrence in 1999, using a Bayesian framework [37]. In 2000, Ramensky et al. developed a similar method which uses a Bayesian estimator to measure the degree of homogeneity in segmentation [38]. In this method, optimal segmentation is obtained by maximising the likelihood function using the dynamic programming technique presented in [39]. After completion, the partition function approach is used to obtain segmentation with longer segments by filtering the boundaries. In contrast to the approach of Liu and Lawrence, this method does not use probability distributions for segment boundaries and does not use sampling. A related method is presented in [40], which uses reversible jump Markov chain Monte Carlo (RJMCMC) sampling method to estimate posterior probabilities [41]. In contrast to Liu and Lawrence, they have used Poisson intensity models as the underlying model (as opposed to multinomial likelihood). The method has been tested by applying to modelling the occurrence of ORFs along the human genome. Another Bayesian model can be found in [42].

The method on which we focus in the main part of this article [43,44] is also of this type. The method can be described as a segmentation–classification model as it not only detects change-points but also groups segments based on their sequence characteristics. The group to which a segment belongs is essentially a hidden state, in the terminology of HMMs, and the classification is unsupervised, in the terminology of machine learning. There are two main innovations in this method. The first is that the character frequencies (emission probabilities) for a given segment are not constant for all segments in a group. Instead, the character frequencies are drawn from a Dirichlet distribution specific to the group to which that segment belongs, and it is the