BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 5 2011, pages 604–610
doi:10.1093/bioinformatics/btq716

Genome analysis

Advance Access publication January 5, 2011

Model selection in Bayesian segmentation of multiple DNA alignments

Christopher Oldmeadow[1,*] and Jonathan M. Keith[2,*]

[1]Centre for Clinical Epidemiology and Biostatistics, University of Newcastle, NSW and [2]School of Mathematical Sciences, Monash University, Victoria, Australia

Associate Editor: David Posada

ABSTRACT

Motivation: The analysis of multiple sequence alignments is allowing researchers to glean valuable insights into evolution, as well as identify genomic regions that may be functional, or discover novel classes of functional elements. Understanding the distribution of conservation levels that constitutes the evolutionary landscape is crucial to distinguishing functional regions from non-functional. Recent evidence suggests that a binary classification of evolutionary rates is inappropriate for this purpose and finds only highly conserved functional elements. Given that the distribution of evolutionary rates is multi-modal, determining the number of modes is of paramount concern. Through simulation, we evaluate the performance of a number of information criterion approaches derived from MCMC simulations in determining the dimension of a model.

Results: We utilize a deviance information criterion (DIC) approximation that is more robust than the approximations from other information criteria, and show our information criteria approximations do not produce superfluous modes when estimating conservation distributions under a variety of circumstances. We analyse the distribution of conservation for a multiple alignment comprising four primate species and mouse, and repeat this on two additional multiple alignments of similar species. We find evidence of six distinct classes of evolutionary rates that appear to be robust to the species used.

Availability: Source code and data are available at http://dl.dropbox.com/u/477240/changept.zip

Contact: jonathan.keith@monash.edu; christopher.oldmeadow@newcastle.edu.au.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 2, 2010; revised on November 24, 2010; accepted on December 21, 2010

1 INTRODUCTION

Much attention has been given to delineating protein-coding regions, believed to account for ∼1.2% of the human genome (Collins *et al.*, 2004). However, it is the extensive non-protein-coding regions that remain a largely untapped resource for the identification of novel classes of functional elements. Recent studies have indicated that there exists a vast amount of functionality in the non-coding regions (Mattick, 2005; Oldmeadow *et al.*, 2010; Pheasant and Mattick, 2007). The task of delineating and identifying the functional elements is hindered by the lack of understanding of the constraints shaping such elements.

Studying rates of evolution in genomes is also key to understanding genomic function. Human–mouse and human–dog sequence comparisons suggest that ∼5% of the mammalian genome is subject to purifying evolutionary selection (Lindblad-Toh *et al.*, 2005; Waterston *et al.*, 2002). This figure is supported by the recent finding that 5% of bases are confidently predicted as being under evolutionary constraint in mammals by two out of three algorithms employed in the ENCODE project analysis (Birney *et al.*, 2007) and in analyses of large alignments of mammals (Pollard *et al.*, 2010). What remains unclear is the relationship between the rate at which a functional element evolves and its functional type. It is understood that conserved sequences are likely to harbour functional elements: the length of evolutionary time the segments have remained unchanged is evidence of biological or functional significance. It should be stressed, however, that the amount of conserved sequence is a lower bound on the amount of functional sequence, since lack of evolutionary constraint on a sequence does not imply lack of function (Pang *et al.*, 2006). Indeed, the ENCODE pilot study found many functional elements that are seemingly unconstrained across mammalian evolution.

There is a strong demand for the development of more sophisticated bioinformatic approaches to studying distributions of evolutionary rates. Current approaches such as the simple per cent identity within a sliding window (Lindblad-Toh *et al.*, 2005; Waterston *et al.*, 2002) and the two state phylo-Hidden Markov Model (Pollard *et al.*, 2010; Siepel and Haussler, 2004) hide a multitude of sins: most importantly, the implicit assumption of a dichotomy of evolutionary rates: conserved and non-conserved. Conservation estimates are based on a continuous model of a known distribution of evolutionary rates for ancient repeat (AR) sequences, assumed to be neutrally evolving, and an unknown distribution for functional sequences. This unknown distribution is estimated through the computationally efficient sliding window technique; however, results are affected by window size. It is assumed that the conserved fraction is the difference, estimated by mixture decomposition, between the level of similarity across all windows in the alignable fraction of the genome versus that in AR sequences or other supposedly neutrally evolving sequences.

This fundamental assumption of a dichotomy of conservation levels does not adequately capture the complexity of the evolutionary rate distribution and may lead to invalid estimates of

*To whom correspondence should be addressed.

the amount of functional sequence. The knowledge of variation in mutation rates along a sequence has been acknowledged for some time [see Oldmeadow *et al.* (2010) and references therein for an overview]. The discrete-gamma model (Yang, 1994, 1996) was developed to account for rate variation when inferring phylogenetic relationships from molecular sequences. This model allows a variable number of substitution rate classes to approximate a gamma distribution of rates for computational efficiency; however, the underlying distribution of rates is still considered to be a single univariate gamma distribution. Recently, by using a more sensitive statistical approach, the authors showed that a multi-modal distribution of evolutionary rates is more appropriate, and one can no longer confidently separate conserved from non-conserved sequences using existing approaches (Oldmeadow *et al.*, 2010). This discovery was based on a method of estimating distributions of evolutionary rates using a Bayesian multiple changepoint model (Keith, 2006; Keith *et al.*, 2004): a genomic segmentation technique that simultaneously segments a genome and classifies the segments, with computational speed that enables chromosome or even genome-wide analysis. The model, denoted here as the Bayesian Segmentation and Classification Model (BSCM), has been used with success at analysing pairwise (Keith *et al.*, 2008) and multiple (Oldmeadow *et al.*, 2010) alignments. In summary, the algorithm takes as input a sequence of characters (which may represent pair-wise or multiple alignments) and output positions that delineate homogeneous segments, the number of which is unknown *a priori*. Each segment is also classified based on the probability of observing each of the constituting input characters. Most importantly, the algorithm allows for an unknown number of segment classes; a facet of the model recently exploited for finding multiple classes of evolutionary rates in animal genomes (Oldmeadow *et al.*, 2010). The optimal number of classes is determined by fitting a number of models differing only in the number of classes, and calculating an information criterion for each model, based on a tradeoff between model fit and model complexity. The model with the smallest information criterion is generally considered the optimal; however, in practice, the model for which the majority of reduction in information criterion has occurred is chosen for further investigation, as there may not be a true optimum.

The ability to take advantage of already sampled MCMC values is key to efficient analysis of these highly complex hierarchical models. Approximations to information criteria such as the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) have been used previously (Keith, 2006; Keith *et al.*, 2008; Oldmeadow *et al.*, 2010); however, the behaviour of these approximations under different circumstances has not been investigated. The problem of generating an effective tool for model comparison is a challenging statistical problem, especially for computationally demanding hierarchical models. We consider the problem of selecting the optimal number of classes for the BSCM through information criteria and an approximation to the Deviance Information Criterion (DIC) for the BSCM is also presented. The behaviour and capabilities of three information criteria are explored through simulated data, examining the potential of the chosen model to detect superfluous changepoints and mixture modes and also the effect of sequence size on the resulting optimal model. Recent advances in DNA sequencing have lead to the sequencing of a number of species that are closely related to humans. We apply the BSCM to a sequence derived from the region corresponding to human chromosome 21 from a multiple alignment of four primates and mouse using the most recent assemblies. We explore the problem of identifying the optimal number of evolutionary rate classes for this input sequence.

2 APPROACH

2.1 The BSCM

A full description of this model can be found in Keith *et al.* (2008) or the Supplementary Material in Oldmeadow *et al.* (2010). A brief summary of the model is as follows: each alignment position is reduced to a single character, summarizing some feature of interest, with examples including GC content, match/mismatch for pairwise alignments and parsimony score (Fitch, 1971) for multiple alignments. A sequence of such characters (of dimension D) defines the input sequence (S), assumed to be possible to partitioned into a series of homogeneous segments of length L, defined by a vector of changepoint positions, denoted by A. A uniform prior distribution is assumed for the locations of the changepoints, some of which are fixed, based on the blocks in the input alignment and a variable number, k, are estimated. The characters within a segment are assumed independent and drawn from a multinomial distribution with parameters, each drawn from one of T Dirichlet (α) distributions with uniformly sampled mixture proportions , π. Uniform prior distributions are placed on the mean and SD of the Dirichlet parameters.

The posterior density function for the parameters of a BSCM is given as:

$$p(k,A,\pi,\alpha|S) \propto \Gamma(k+1)\Gamma(L-k) \times \prod_{i=1}^{k+1} f(m_i|\pi,\alpha), \qquad (1)$$

where m_{ij} is the number of occurrences of character j in segment i and,

$$f(m_i|\pi,\alpha) = \sum_t \left[\pi_t \frac{\Gamma(\sum_{j=1}^{D} \alpha_j^{(t)})}{\prod_{j=1}^{D} \Gamma(\alpha_j^{(t)})} \times \frac{\prod_{j=1}^{D} \Gamma(m_{ij}+\alpha_j^{(t)})}{\Gamma(\sum_{j=1}^{D}(m_{ij}+\alpha_j^{(t)}))} \right]. \qquad (2)$$

The generalized Gibbs sampler (GGS) is used to sample from this varying dimensional space: it allows the number of changepoints to vary. However, there is another dimension in this model that can vary—the number of classes T. It is possible to utilize the GGS again to allow the model dimension to vary in T, but it is unclear whether this approach would be advantageous. The availability of cluster computers makes information criterion methods of determining the best model an attractive choice. Typically, in this approach the number of classes is assumed known and the model parameters are then estimated for a range of numbers of classes. Each fitted model is then assessed in terms of model fit and model complexity, and the most parsimonious model that explains the data is chosen. There are a number of different choices for measuring model fit and complexity; in the following section, we give a brief overview of this problem in a general setting and provide a summary of how these measures are used in the BSCM.

2.2 AIC

The AIC (Akaike, 1974) is an estimate of the Kullback–Leibler distance between the hypothetical 'true' and estimated distributions.

It is defined as:

$$\mathrm{AIC} = -2\ln f(y|\hat{\theta}) + 2d \quad (3)$$

where y is the observed data, d is the number of free parameters and $\hat{\theta}$ the maximum likelihood estimate (MLE) for the model parameters θ. Strictly speaking, the AIC is not appropriate for mixture models, since its derivation relies on asymptotic theory of the MLE and the breakdown of regularity conditions namely, parameter values under the null hypothesis are on the boundary of the parameter space (Aitkin and Rubin, 1985; Titterington *et al.*, 1985). Nevertheless, it still is used with success in practice (Biernacki *et al.*, 1998). The AIC tends to penalize models with larger numbers of parameters less than other information criteria.

An approximation to Equation (3) described in Keith *et al.* (2008) is:

$$\widehat{\mathrm{AIC}} = -2\overline{\ln f(y|\theta)} + 2d.$$

This approximation has the advantage that it makes use of MCMC sampled values, with $-2\overline{\ln f(y|\theta)}$ being the posterior mean deviance, rather than the optimized deviance. The use of the posterior mean deviance as a measure of model fit has been adopted by many authors (Dempster, 1974; Gilks *et al.*, 1993; Richardson and Green, 1997). For the BSCM, the advantage is that the criterion can be estimated using the already sampled segmentations, and is defined as:

$$\widehat{\mathrm{AIC}} = -2\overline{\ln f(y|\theta)} + 2[\bar{k} + T \times (D+1)]$$

where $\bar{k}$ is the average number of changepoints over the set of segmentations sampled by MCMC.

2.3 BIC

The BIC is an information criterion derived from a Bayesian approach to model selection (Schwarz, 1978). It has been shown to be consistent in estimating the number of components for a finite mixture model (Keribin, 2000). The BIC is defined as:

$$\mathrm{BIC} = -2\ln f(y|\hat{\theta}) + d\ln n,$$

where $\hat{\theta}$ is the maximum likelihood solution. The BIC tends to favour models with fewer parameters than the AIC since the BIC penalizes parameters more heavily. An approximation to the BIC that makes use of posterior samples is defined for the BSCM in Oldmeadow *et al.* (2010) as:

$$\widehat{\mathrm{BIC}} = -2 \times \overline{\ln f(y|\theta)} + [\bar{k} + T(D+1)]\ln n.$$

2.4 DIC

The DIC (Spiegelhalter *et al.*, 2002) can be seen as a Bayesian version of the AIC and was originally developed as a model selection tool in generalized linear models. It takes on a similar form to the AIC; however, the complexity penalty is based on the number of effective model parameters, and not the actual number of free parameters as in AIC.

The complexity measure proposed by Spiegelhalter *et al.* (2002), p_D, is known as the effective number of parameters (or effective dimension) and is defined as the difference of the posterior mean deviance and the deviance at the posterior estimates of the parameters. Providing the posterior distribution can be reasonably approximated by a multivariate normal distribution, Spiegelhalter *et al.* (2002) noted that the model Deviance, $D(\theta) = -2\ln f(y|\theta)$ can be approximated by:

$$D(\theta) \approx D(\hat{\theta}) + \chi^2_p, \quad (4)$$

where χ^2_p is a chi-squared distribution with p degrees of freedom, p is in this case the number of model parameters. By taking expectations of both sides of 4, the formula for the effective dimension is given as:

$$p_D = \overline{D(\theta)} - D(\hat{\theta})$$

$\overline{D(\theta)}$ is the mean posterior deviance, ($E_\theta[D(\theta)]$) and $\hat{\theta}$ is an estimate of θ, often the posterior mean, median or mode.

The DIC is thus defined as:

$$\mathrm{DIC} = \overline{D(\theta)} + p_D \quad (5)$$

$$= D(\hat{\theta}) + 2p_D \quad (6)$$

The DIC in Equation (6) can be seen as analogous to the AIC in Equation (3), with two essential differences. One, the plug-in deviance in the AIC is calculated at maximum-likelihood estimates of parameters, whereas the plug-in deviance in the DIC is calculated at posterior means (or, alternatively, posterior medians or posterior modes). Two, the penalty function for model complexity in the AIC is determined by the nominal number of parameters in the model, whereas the penalty function for model complexity in the DIC is determined by an estimate of the effective number of parameters in the model. The nominal number of parameters assumes that parameters are independent, whereas the effective number of parameters allows for non-zero covariance among parameters. Furthermore, the nominal number of parameters can be difficult to ascertain, especially in hierarchical models, whereas the effective number of parameters is estimated from the data.

The problem for missing data models such as the mixture model is that p_D cannot be intrinsically defined. The use of DIC in missing data problems has been considered (Celeux *et al.*, 2006). In these models, the parameters θ may not always be identifiable [a problem known as label switching (Stephens, 2000)] and using the posterior mean for $\hat{\theta} = E[\theta|y]$ can be a poor estimate. A more relevant choice is the posterior median or mode.

For the Bayesian segmentation model, the set over which we are sampling is large. Therefore, finding the posterior mode is problematic and techniques such as simulated annealing (Kirkpatrick *et al.*, 1983) can be used.

Another problem with p_D is that it can be negative if the posterior distribution is very different from the normal distribution, hence $f(\hat{\theta})$ does not provide a very good estimate of θ. Also it is possible to get negative p_D , if the sampling distribution is non-log concave and there is strong prior-data conflict (Gelman *et al.*, 2004).

An alternative to p_D has been tentatively proposed by various sources [(Gelman *et al.*, 2004; Sturtz *et al.*, 2005) and discussed in Raftery *et al.* (2007)]. If instead of taking expectations of Equation (4), the variance is taken resulting in a different estimator of the effective dimension size. This estimate has been termed p_V and is simply:

$$p_v = \mathrm{Var}(D(\theta))/2$$

and,

$$\mathrm{DICV} = p_v + \overline{D(\theta)}$$

Although it has been suggested that this version tends to substantially overestimate the effective number of parameters for normal random effect models [discussion in Raftery *et al.* (2007)].

Throughout this article, we propose the term DICV to signify the use of p_V in the DIC.

3 METHODS

We simulated three sequences using a four-character alphabet, each sequence consisted of approximately 1×10^6 characters with only fixed changepoints. The positions of the fixed changepoints were determined using 1000 randomly sampled block lengths taken from the analysis of a parsimony sequence derived from an alignment of four mammals in a previous study by the authors (Oldmeadow *et al.*, 2010). The multinomial probabilities for each segment were drawn from a Dirichlet distribution with parameters set according to three scenarios:

(1) A single class distribution, with parameters set to those of the most abundant class identified in the mammalian analysis (Oldmeadow *et al.*, 2010). For each block, we sampled multinomial probabilities from a Dirichlet (854.29, 553.71, 25.87, 0.09) distribution.

(2) A three-class distribution, with parameters chosen as the posterior means from the three-class distribution identified in the analysis of Ancient Repeat sequence from the mammalian alignment (Oldmeadow *et al.*, 2010). Specifically: $\pi = \{0.07, 0.20, 0.73\}$, $\alpha_1 = \{41.99, 9.19, 0.33, 0.01\}$, $\alpha_2 = \{29.13, 29.24, 2.48, 0.01\}$, $\alpha_3 = \{1180.95, 693.89, 24.95, 0.13\}$

(3) A seven-class distribution, with parameters set to as the posterior means from the seven-class distribution identified in the analysis of the mammalian alignment (Oldmeadow *et al.*, 2010) $\pi = \{0.113, 0.012, 0.211, 0.073, 0.553, 0.006, 0.030\}$, $\alpha_1 = \{207.74, 179.43, 11.20, 0.06\}$, $\alpha_2 = \{15.22, 13.82, 3.68, 0.08\}$, $\alpha_3 = \{665.88, 310.82, 11.36, 0.02\}$, $\alpha_4 = \{346.22, 94.75, 3.27, 0.02\}$, $\alpha_5 = \{851.11, 553.63, 25.91, 0.09\}$, $\alpha_6 = \{121.64, 8.19, 0.17, 0.01\}$, $\alpha_7 = \{863.17, 127.51, 4.30, 0.01\}$.

To determine sensitivity to input sequence size, we simulated a sequence of approximately 1×10^7 characters using the same scheme as described above.

All models were run for a period of 20 000 iterations (a single iteration defined as a full sweep over all parameters), at which point convergence was determined by inspecting the trace plots of the model parameters, and each of the the cumulative information criteria were inspected to determine if the chains were stationary. The expected proportion of no mutations against the mixture proportions were plotted to ensure no systematic drift. The MCMC chains were then set to run for another 5000 samples, post burn-in.

We also ran the changepoint algorithm on a sequence of parsimony scores derived from a multiple DNA sequence alignment of four primates: Human (hg19), Chimp (panTro2), Gorilla (gorGor1) and Orangutang (ponAbe2) and Mouse(mm9). These data were obtained by extracting alignment blocks containing all five species from the 45-multi-species alignment (Kuhn *et al.*, 2009). Only blocks containing DNA sequence from standard chromosomes for all five species were retained. Further filtering included the removal of alignment columns that contained no-call or insertion–deletion characters. Parsimony scores were calculated for the filtered alignment columns using the method of Fitch (1971). We ran the changepoint algorithm over a range of class numbers (1..10) .

Finally, we repeated the above analysis on an additional two multiple alignments:

(1) Human (hg19) ,Gorilla (gorGor1), Rhesus (rheMac2), Tarsier (tarSyr1), Mouse (mm9) and

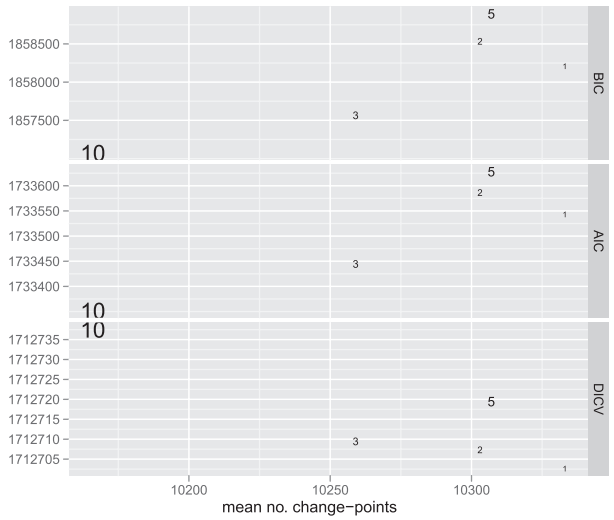

Fig. 1. Information criterion versus mean number of changepoints for 1×10^6 characters of simulated sequence (alphabet size = 4), drawn from a single Dirichlet distribution. The size of the numerals is related to the deviance (-2ln L), where larger numbers indicate a larger deviance and hence a poorer fit.

(2) Human (hg19), Tarsier (tarSyr1), Bushbaby (otoGar1), Tree shrew (tupBel1), Mouse (mm9).

4 DISCUSSION

4.1 Simulation

Upon first inspection of results from the one-class simulation (Fig. 1), we would choose the 10-class model as being superior using AIC or BIC. However, it is clear in Figure 1 that the 10-class model was the worst fit (the size of the numbers is proportional to the deviance), and that model resulted in a smaller number of changepoints than for other models. This would suggest that both AIC and BIC have been conservative and placed a heavy weighting for the complexity penalty. The DICV, however, appears to put more weighting for model fit (or less penalty on model complexity)—the one-class model has the minimum DICV. It should be noted that a degree of discretion should be used when choosing the best model, as inspection of the estimated mixture proportions tells another story. The mixture proportions for each fitted model all exhibit of a single dominant class, with the remainder of classes having much smaller weights ($<0.0003\%$) (with the exception of the five-class model, in which case there are two majority classes with similar Dirichlet parameters). This would suggest that one should choose the one-class model as the superior model, as indicated by DICV and as known to be correct. The median posterior number of changepoints for the one-class model was 10 373; this is a slightly larger increase in the true number of changepoints (9892). The posterior median Dirichlet parameters for the best model were $\hat{\alpha} = (762.8, 494.1, 23.1, 0.07)$. Parameter estimates for the 1×10^7 character sequence are even closer to the true parameter values and given in Supplementary Figure S1.

For the higher order simulated data (Scenarios 2 and 3), we found few differences among the information criteria. The model with the minimum information criterion was the true model for the three-class

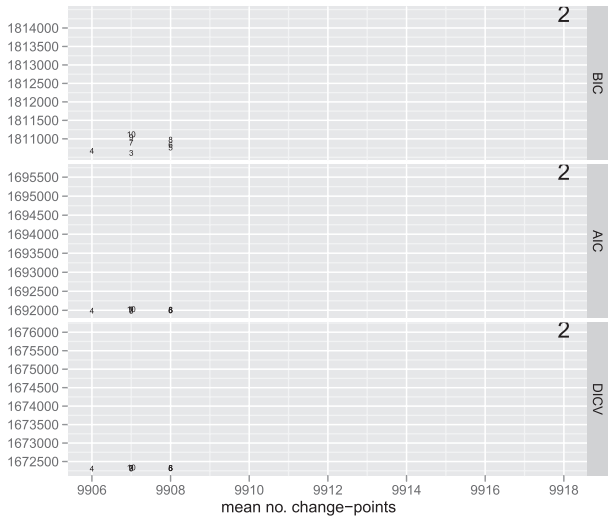

Fig. 2. Information criterion versus mean number of changepoints for 1×10^6 characters of simulated sequence (alphabet size = 4), drawn from a mixture of three Dirichlet distributions. The size of the numerals is related to the deviance (-2ln L), where larger numbers indicate a larger deviance and hence a poorer fit.

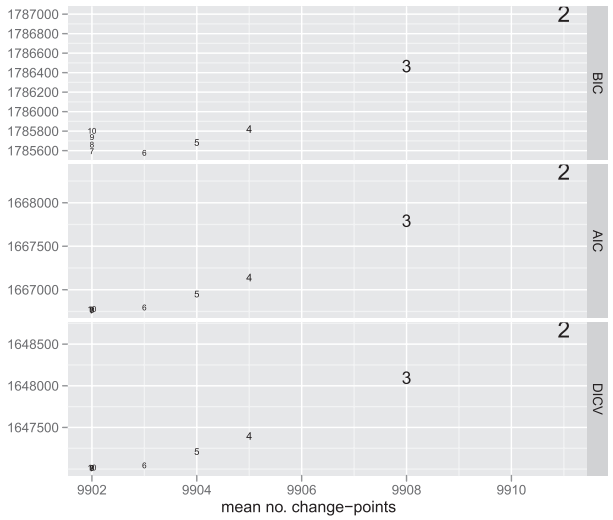

Fig. 3. Information criterion versus mean number of changepoints for 1×10^6 characters of simulated sequence (alphabet size = 4), drawn from a mixture of seven Dirichlet distributions. The size of the numerals is related to the deviance (-2ln L), where larger numbers indicate a larger deviance and hence a poorer fit.

simulated data (Fig. 2). For the simulated seven-class sequence, it appears as though the majority of reduction in all criteria had occurred by the sixth class (Fig. 3). Since there is no further reduction in information criterion or mean number of changepoints beyond the seven-class model, we would choose this model as optimal and indeed this was the true model. Note, a choice based on information criterion alone would be the six-class model, which turns out to be conservative as it is fewer classes than were simulated.

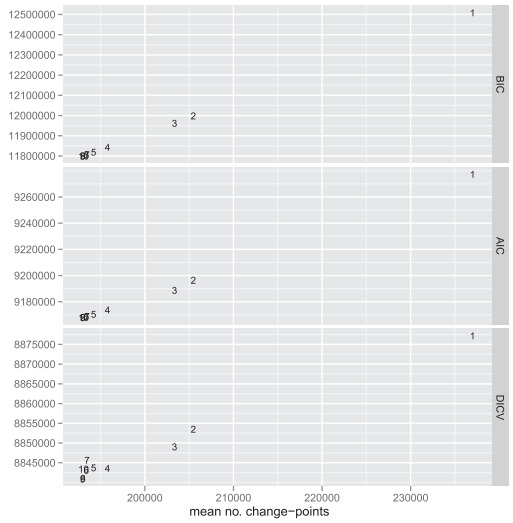

Fig. 4. Information criterion versus mean number of changepoints for the parsimony sequence generated from the five-species alignment (alphabet size = 4). The size of numerals does not reflect deviance in this figure. The information criteria appear to have levelled off and there is no further reduction in number of changepoints by six classes. This would suggest the six-class model is optimal.

For the seven-class data, the AIC and DICV behaved similarly. The number of estimated changepoints for fitted models with $7, 8, 9$ and 10 classes were essentially the same, and both AIC and DICV assigned essentially the same rankings to the models. However, the effect of penalty on higher class numbers could be seen for BIC: higher information criterion values were assigned in BIC for models of higher dimension. This is consistent with BIC being a conservative measure placing a higher weight on model complexity; the complexity term $[\bar{k} + c \times (a+1)]$ is multiplied by 2 in AIC, whereas for BIC the multiplicative factor is $\ln n$.

4.2 Segmenting primate genomes

For an alignment of five DNA sequences, the only possible parsimony scores given the tree relating the species are $\{0, 1, 2, 3\}$. The parsimony score input sequence was 61 656 42 characters in length and there were 109 271 fixed changepoints corresponding to alignment boundaries. The proportion of each of the four characters in this input sequence were 62.35, 36.11, 1.49 and 0.04%, respectively.

We calculated the three information criteria for each of the models, and plotted these against the posterior mean number of changepoints (Fig. 4). It is clear that a minimum value of all three information criteria is reached at six classes, and for more classes it can be seen that there are much smaller reductions in the posterior mean number of changepoints. It could be argued that the majority of reduction in all information criteria had occurred by the four-class model; however, the mixture proportions for the six-class model are all appreciable ($\pi = \{0.034, 0.064, 0.451, 0.035, 0.317, 0.097\}$), which suggests the additional classes may be real.

Figure 5 shows fitted densities of a proxy for conservation level for the five species alignment. The proxy is the proportion of columns in which no mutations occur, parametrized by α_0. The distributions of conservation levels are mixtures of beta

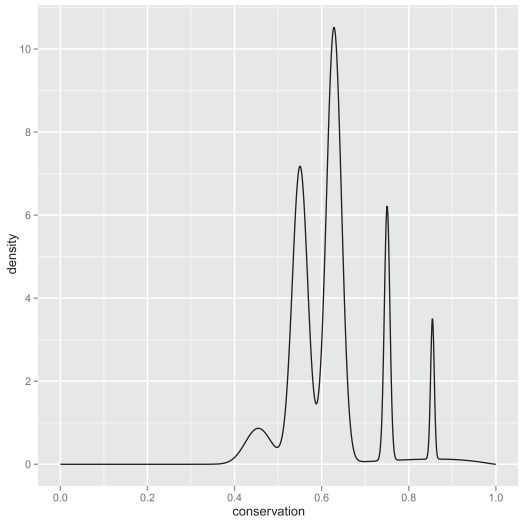

Fig. 5. The density of conservation levels for the five-species alignment. We used the expected proportion of alignment columns with no mutations as a proxy for sequence conservation, and this figure shows at least five well-separated classes. There is a small component with low conservation in the left tail of this distribution which is not visible on this plot (proportion = 0.9%).

distributions, with the mixture proportions estimated by taking posterior sample means and beta parameters estimated using the posterior medians (the medians were used in this case because a small number of large values had large effects on the means). The figure shows clearly that the modes are all well separated and correspond to class of sequence evolving at different rates among the five species.The analyses of the two additional multiple alignments were overwhelmingly similar, with the same optimal number of classes chosen. This is remarkable considering the difference in size of the input sequences for the three multiple alignment. A plot of the fitted conservation densities for all three alignments and the alignment of four species from Oldmeadow *et al.* (2010) is in Supplementary Figure S2. This shows the clear consistency in the location and relative proportions of the six classes of evolutionary rate. Of particular interest is the most slowly evolving class that occurs in all four alignments. We speculate this class could represent a highly conserved group of functional elements in primates, and is the focus of further research. The median length of the intervals between adjacent changepoints is given in Supplementary Table S1 for each of the identified classes for each analysis.

5 CONCLUSION

We have presented an overview of information criterion approaches to model selection in genomic segmentation, and presented three methods of approximating information criterion using posterior samples. We conclude that the approximate information criteria are reasonable for model selection. In the simulation study, we successfully chose the true number of components for all three scenarios, differing in number of components and also the underlying Dirichlet parameters. Although, these criteria are very blunt instruments, used on their own they may result in incorrect model choice; however, a better choice can be made when these criteria are used in conjunction with an examination of the other model parameters such as the mixture proportions, model deviance and number of changepoints. Hence, it is advised that they be used with care and intelligence.

Another important result is that the information criterion approach to model selection does not inflate the number of modes; if anything it is a conservative estimate of the number of modes. We applied the information criterion to determine the number of classes for an alignment of four primates and mouse and found six well-resolved classes. A previous analysis (Oldmeadow *et al.*, 2010) using a parsimony score mapping of columns from an alignment of three primates (human, chimp, Rhesus-monkey) and mouse found evidence of seven classes, and moreover a comparison of Figure 5 to Figure 5c in Oldmeadow *et al.* (2010) suggests that many of the classes identified in the two analyses correspond. The fact that such similar results were obtained is particularly intriguing because the estimated divergence times for the four primates (human, chimp, gorilla, orangutan) are much less than those of the three primates, yet there is still substantial evidence that there are multiple well-resolved classes of evolutionary rates. Hence, our methods are applicable to quite closely related species provided that a suitable outgroup is included.

Moreover, the results from the simulation study increase our confidence in the results reported in Oldmeadow *et al.* (2010). We have shown that the use of fixed changepoints does not result in an over-inflated estimate of the number of non-fixed changepoints, hence modelling natural boundaries such as alignment block boundaries as fixed changepoints apparently has no detectable influence.

There is still much work needed in the area of model selection and its application to these models. While information criterion approaches to estimating the number of classes in the BSCM appear to be justified, they are computationally expensive, even with high powered clusters of computers. Further investigation would be warranted in applying a GGS scheme to sample over the model space, and this is a direction of future research.

The ultimate goal of genomic segmentation is to attribute a specific functional class to each of the discovered segment classes; however, at this stage results suggest that the classes contain mixtures of functional classes (Oldmeadow *et al.*, 2010). Multiple sources of data indicative of function can be incorporated into the model by defining a suitable input sequence. In preliminary work by the authors, we have used these methods to identify a large number of classes when other information is accounted for in the input sequence, such as GC content, frequency of indels and also mutation types. This has led to the discovery of many novel functional elements in regions once thought of as largely not functional.

The work presented here shows that we can be confident that any classes identified through changepoint analysis are most likely real classes and not artefacts of the data, and the information criteria should be used with a degree of caution. Our results also show the model is not influenced by the number of fixed changepoints or the number of segment classes underlying the input sequence.

Funding: Australian Research Council (grants DP0879308 and DP1095849).

Conflict of Interest: none declared.

REFERENCES

Aitkin,M. and Rubin,D. (1985) Estimation and hypothesis testing in finite mixture models. *J. R. Stat. Soc. Ser. B* , **47**, 67–75.

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **19**, 716–723.

Biernacki,C. *et al.* (1998) Assessing a mixture model for clustering with the integrated classification likelihood. *Rapports de recherche- INRIA*. Vol. RR-3521, Available at http://hal.inria.fr/inria-00073163/PDF/RR-3521.pdf.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Celeux,G. *et al.* (2006) Deviance information criteria for missing data models. *Bayesian Anal.*, **1**, 651–674.

Collins,F. *et al.* (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Dempster,A. (1974) The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference, Aarhus, May 7-12, 1973.* Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, p. 335.

Fitch,W. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.

Gelman,A. *et al.* (2004) *Bayesian Data Analysis*. Champan and Hall/CRC, Boca Raton, FL.

Gilks,W. *et al.* (1993) Modelling complexity: applications of Gibbs sampling in medicine. *J. R. Stat. Soc. Ser. B*, **55**, 39–52.

Keith,J. *et al.* (2004) A generalized Markov sampler. *Methodol. Comput. Appl. Probab.*, **6**, 29–53.

Keith,J. *et al.* (2008) Delineating slowly and rapidly evolving fractions of the Drosophila genome. *J. Comput. Biol.*, **15**, 407–430.

Keith,J. (2006) Segmenting eukaryotic genomes with the generalized gibbs sampler. *J. Comput. Biol.*, **13**, 1369–1383.

Keribin,C. (2000) Consistent estimation of the order of mixture models. *Sankhy Indian J. Stat. Ser. A*, **62**, 49–66.

Kirkpatrick,S. *et al.* (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Kuhn,R. *et al.* (2009) The UCSC genome browser database: update 2009. *Nucleic Acids Res.*, **37**, D755.

Lindblad-Toh,K. *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819.

Mattick,J. (2005) The functional genomics of noncoding RNA. *Science*, **309**, 1527–1528.

Oldmeadow,C. *et al.* (2010) Multiple evolutionary rate classes in animal genome evolution. *Mol. Biol. Evol.*, **27**, 942.

Pang,K. *et al.* (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.*, **22**, 1–5.

Pheasant,M. and Mattick,J. (2007) Raising the estimate of functional human sequences. *Genome Res.*, **17**, 1245.

Pollard,K. *et al.* (2010) Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.

Raftery,A. *et al.* (2007) Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Stat.*, **8**, 1–45.

Richardson,S. and Green,P. (1997) On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. Ser. B*, **59**, 731–792.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Siepel,A. and Haussler,D. (2004) Combining phylogenetic and Hidden Markov Models in biosequence analysis. *J. Comput. Biol.*, **11**, 413–428.

Spiegelhalter,D. *et al.* (2002) Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* , **64**, 583–639.

Stephens,M. (2000) Dealing with label switching in mixture models. *J. R. Stat. Soc.*, **62**, 795–809.

Sturtz,S. *et al.* (2005) R2WinBUGS: a package for running WinBUGS from R. *J. Stat. Softw.*, **12**, 1–16.

Titterington,D. *et al.* (1985) *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons Inc., New York.

Waterston,R. *et al.* (2002) Initial sequencing and comparitive analysis of the mouse genome. *Nature*, **420**, 520–562.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.

Yang,Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.*, **11**, 367–372.