Approximate Bayesian Discrimination between Alternative DNA Mosaic Structures

Dirk Husmeier and Frank Wright
Biomathematics and Statistics Scotland (BioSS)
SCRI, Invergowrie, Dundee DD2 5DA, United Kingdom
Phone: ++44 1382 562731, email: dirk@bioss.ac.uk

Abstract: We derive an approximate Bayesian hypothesis test to discriminate between alternative mosaic structures of DNA sequence alignments, and test the viability of this approach on a set of synthetic and real-world DNA sequence alignments.

1 Introduction

There has recently been an increased interest in sporadic recombination as an important, and previously underestimated, source of genetic diversification in bacteria and viruses [6], [7]. The exploitable consequence of this process, in which DNA subsequences are exchanged between different strains or species, is that the DNA sequence alignment of the involved taxa has a mosaic structure, with different regions corresponding to different phylogenetic topologies. While several methods for identifying the nature and the breakpoints of this mosaic structure have been developed (e.g. [2], [3], [5]), they do not satisfactorily address the question of whether the found mosaic structure is statistically significant\(^1\).

The aim of this paper, therefore, is to devise a hypothesis test to discriminate between alternative candidate mosaic structures

2 Method

Let \( \mathcal{D} \) denote a DNA sequence alignment, and \( H \) a hypothesis about its mosaic structure. In the absence of prior knowledge we should discriminate between alternative mosaic structures \( H_\alpha \) and \( H_\beta \) on the basis of the Bayes factor \( P(\mathcal{D}|H_\beta)/P(\mathcal{D}|H_\alpha) \), which, if greater than 1, suggests rejecting \( H_\alpha \) in favour of \( H_\beta \). Bayesian hypothesis testing and model selection is thus based on the computation of the marginal log likelihood or evidence

\[
P(\mathcal{D}|H) = \int P(\mathcal{D}|\mathbf{q}, H)P(\mathbf{q})d\mathbf{q}
\]

where \( \mathbf{q} \) is the vector of all model parameters, which has the prior probability distribution \( P(\mathbf{q}) \).

While this integral is intractable, a partial factorisation of the posterior distribution \( P(\mathbf{q}|\mathcal{D}, H) \) and

\(^1\)In [2] and [5], a statistical significance test is developed for the signal detecting the mosaic structure rather than the mosaic structure itself. This can lead to erroneous conclusions, as seen from Figure 2 in [5], and Figure 4A in [2].
Figure 1: **Left:** Phylogenetic tree for the synthetic problem. DNA sequences (5000 bp long) were evolved along the tree, using the Kimura 2-parameter model (transition-transversion ratio = 2) of nucleotide substitution. Two recombination events, involving closely related (A01 ←→ A10) and distantly related (A00 ←→ B00) taxa, were simulated by swapping the indicated lineages. All branch lengths are equal (unit branch length $w$) except for the internal branch between A and B, whose length is $2w$. **Right:** Mosaic structures of DNA sequence alignments. The first two rows show the null hypothesis (top) and the true mosaic structure (second from the top). The remaining rows show alternative mosaic structures: 1) Segmentation of only the left recombinant region (A00 ←→ B00). 2) Segmentation of only the right recombinant region (A01 ←→ A10). 3) Subdivision of both recombinant regions. 4) Subdivision of the three non-recombinant regions. 5) Subdivision of all regions. 6) Correct segmentation with a slight misplacement of the breakpoints, shifted by 10 nucleotides to the right.

the Laplace method lead to the following approximation:

$$
\ln P(D|H) = U(H) + \sum_{k=1}^{K} S_{\psi_k}(H) + \frac{1}{2} \sum_{k=1}^{K} \ln \text{det } C_k - \frac{n}{2} \sum_{k=1}^{K} \ln N_k + \frac{m+n}{2} \ln(2\pi e)
$$

Here, $n$ is the number of branches in the respective phylogenetic tree, $m$ is the number of adapted parameters in the nucleotide substitution model, $K$ is the number of separate mosaic segments in the alignment, $N_k$ denotes the number of polymorphic sites in the $k$th segment of the alignment, $C_k$ is the posterior covariance matrix of the adapted nucleotide substitution parameters, obtained from the $k$th mosaic segment, $S_{\psi_k}(H) = -\sum_{\psi_k} P(\psi_k|D, H) \ln P(\psi_k|D, H)$ is the marginal entropy over tree topologies $\psi_k$, determined with Markov chain Monte Carlo (MCMC), and

$$U(H) = \frac{1}{T} \sum_{t=1}^{T} \left[ \ln P(D|q_t, H) + \ln P(q_t|H) \right]
$$

where $\{q_1, \ldots, q_T\}$ is a sample from the posterior distribution $P(q|D, H)$, obtained with MCMC. A detailed derivation of this equation can be found elsewhere\(^2\).

\(^2\)http://www.bioss.sari.ac.uk/~dirk/papers/BayesSegmentation.pdf
3 Test Datasets

Synthetic Data

We simulated sporadic recombination events in a synthetic population of 8 strains. The DNA sequences of the taxa were evolved down the branches of the phylogenetic tree of Figure 1, using the Kimura 2-parameter model of nucleotide substitution with a transition-transversion ratio of $\tau = 2$. Partial sequences were generated from different topologies, as indicated in the figure, and then spliced together. This simulates the exchange of DNA sequences between different strains of the population. We generated DNA sequence alignments for different values of the unit branch length, varying between $w = 0.1$ and $w = 0.01$.

Hepatitis B Virus

Hepatitis B is caused by a DNA virus with a short genome of 3200 bp. Evidence for recombination was first found in [1]. The sequences used in this paper include two recombinant strains (accession numbers D00329 and X68292), and eight nonrecombinant strains (accession numbers D00330, D00630, L27106, M32138, M54923, M57663, V00866, X01587). The sequences were aligned with ClustalW, using the default parameters. Columns with gaps were discarded, giving a total alignment length of 3049 nucleotides. The recombinant breakpoints found in [1] are at positions $3603, 1882, 2071, \text{and} 2238$.

Neisseria

We analysed a 787 bp DNA sequence alignment of the argF gene of eight strains of Neisseria with the following GenBank accession numbers: X64860, X64861, X64866, X64869, X64870, X64871, X64872, X64873. This data set was used in [8], where a recombinant region between positions 1 and 202 and a differently diverged region between nucleotides 508 and 538 was found.

4 Results

Figures 1 and 2 show different candidate mosaic structures of the DNA sequence alignments discussed in the previous section. These segmentations include the null hypothesis (one homogeneous region), the true mosaic structure, and various alternative mosaic structures in which the true sub-regions have either been partially merged or further sub-divided. The task is to find the true mosaic structure, and we test how reliable a selection criterion the approximate Bayesian evidence of (2) is. The MCMC simulations were carried out with the program package BAMBE, described in [4]. We started the simulations from two different initialisations – a random tree and a tree obtained with the neighbour-joining clustering algorithm – to ensure that the obtained results were consistent. The results thus obtained are shown in Table 1. The upper table shows the values of the relative evidence $\ln P(D|H) − \ln P(D|H_0)$, where $H_0$ is the null hypothesis (no segmentation). The bottom table shows the average log posterior scores, $U(H) − U(H_0)$. For a uniform prior on the parameters, which is assumed in our study – see [4] – $U$ reduces to the average log likelihood and is thus a selection criterion that one

---

$3$ The recombination breakpoints in the original data set prior to discarding columns with gaps were at positions 735, 2014, 2203, 2370.

$4$ The numbering scheme for the bases in [8] starts at 296bp and ends at 1083bp, so the location of the breakpoints have to be shifted by 295bp.
Figure 2: Mosaic structures of real-world DNA sequence alignments. The null hypothesis and the true mosaic structure are shown in the first two lines; the other lines show alternative mosaic structures. *Left (Neisseria):* 1) Only the left recombinant region resolved. 2) Only the right recombinant region resolved. 3) Subdividing the left non-recombinant region. 4) Subdividing the right non-recombinant region. 5) Subdividing both non-recombinant regions. *Middle (Hepatitis-B virus):* 1) Merging of the first three regions. 2) Merging of the first two regions. 3) Merging of the last three regions. 4) Merging of the last two regions. 5) Subdividing the largest region. 6) Subdividing the last region. 7) Subdividing both regions.

naively might be inclined to adopt. However, Table 1 shows that a selection based on $U$ identifies the correct mosaic structure only in a single case and usually prefers the finer tessellated alternatives. On the contrary, the approximate evidence is consistently maximised for the correct mosaic structure, and thus proves to be a reliable selection criterion.

5 Conclusion

This paper has proposed an approximate Bayesian method for the discrimination between alternative mosaic structures of DNA sequence alignments. The method has been applied to several synthetic DNA sequence alignments, where the true mosaic structure was consistently identified. In an application to two real DNA sequence alignments, the method indentified the putative mosaic structures predicted in the literature.
Evidence:\[\ln P(D|H) - \ln P(D|H_0)\]

<table>
<thead>
<tr>
<th>Data</th>
<th>True</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic, w = 0.1</td>
<td>1323</td>
<td>962</td>
<td>407</td>
<td>1236</td>
<td>1206</td>
<td>1119</td>
<td>1271</td>
<td></td>
</tr>
<tr>
<td>Synthetic, w = 0.075</td>
<td>1119</td>
<td>814</td>
<td>333</td>
<td>1048</td>
<td>1009</td>
<td>938</td>
<td>1093</td>
<td></td>
</tr>
<tr>
<td>Synthetic, w = 0.05</td>
<td>965</td>
<td>745</td>
<td>262</td>
<td>887</td>
<td>858</td>
<td>780</td>
<td>933</td>
<td></td>
</tr>
<tr>
<td>Synthetic, w = 0.025</td>
<td>641</td>
<td>517</td>
<td>175</td>
<td>583</td>
<td>550</td>
<td>492</td>
<td>639</td>
<td></td>
</tr>
<tr>
<td>Synthetic, w = 0.01</td>
<td>209</td>
<td>191</td>
<td>18</td>
<td>160</td>
<td>138</td>
<td>89</td>
<td>192</td>
<td></td>
</tr>
<tr>
<td>Neisseria</td>
<td>81.9</td>
<td>70.4</td>
<td>36.0</td>
<td>65.5</td>
<td>75.2</td>
<td>58.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>84.0</td>
<td>70.1</td>
<td>38.2</td>
<td>67.6</td>
<td>78.4</td>
<td>62.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hepatitis-B virus</td>
<td>247</td>
<td>104</td>
<td>198</td>
<td>94</td>
<td>192</td>
<td>242</td>
<td>232</td>
<td>227</td>
</tr>
<tr>
<td></td>
<td>249</td>
<td>102</td>
<td>200</td>
<td>94</td>
<td>193</td>
<td>245</td>
<td>234</td>
<td>230</td>
</tr>
</tbody>
</table>

Average log posterior: \[U(H) - U(H_0)\]

<table>
<thead>
<tr>
<th>Data</th>
<th>True</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic, w = 0.1</td>
<td>1506</td>
<td>1057</td>
<td>500</td>
<td>1500</td>
<td>1503</td>
<td>1496</td>
<td>1453</td>
<td></td>
</tr>
<tr>
<td>Synthetic, w = 0.075</td>
<td>1296</td>
<td>907</td>
<td>425</td>
<td>1302</td>
<td>1295</td>
<td>1301</td>
<td>1271</td>
<td></td>
</tr>
<tr>
<td>Synthetic, w = 0.05</td>
<td>1135</td>
<td>833</td>
<td>349</td>
<td>1128</td>
<td>1138</td>
<td>1130</td>
<td>1102</td>
<td></td>
</tr>
<tr>
<td>Synthetic, w = 0.025</td>
<td>796</td>
<td>596</td>
<td>255</td>
<td>799</td>
<td>796</td>
<td>799</td>
<td>793</td>
<td></td>
</tr>
<tr>
<td>Synthetic, w = 0.01</td>
<td>341</td>
<td>259</td>
<td>87</td>
<td>338</td>
<td>345</td>
<td>343</td>
<td>327</td>
<td></td>
</tr>
<tr>
<td>Neisseria</td>
<td>141.0</td>
<td>120.2</td>
<td>72.3</td>
<td>141.3</td>
<td>149.6</td>
<td>149.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>141.7</td>
<td>119.9</td>
<td>73.0</td>
<td>142.0</td>
<td>150.9</td>
<td>151.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hepatitis-B virus</td>
<td>369</td>
<td>167</td>
<td>288</td>
<td>168</td>
<td>293</td>
<td>396</td>
<td>383</td>
<td>410</td>
</tr>
<tr>
<td></td>
<td>371</td>
<td>166</td>
<td>290</td>
<td>168</td>
<td>294</td>
<td>399</td>
<td>385</td>
<td>413</td>
</tr>
</tbody>
</table>

Table 1: Results of the simulation study. The upper table shows, for each segmentation, the relative evidence \[\ln P(D|H) - \ln P(D|H_0)\], where \(H_0\) is the null hypothesis (no segmentation). The bottom table shows the average log posterior, \[U(H) - U(H_0)\]. For each data set, the highest score is printed in boldface letters. The segmentations are shown in Figures 1 and 2. The MCMC simulations on the real-world data were repeated with different initialisations (top: neighbour-joining tree, bottom: random tree).

References


