October 22, 2010

Review of:

Bayesian Analysis of Partitioned Data

By: B. Moore, J. McGuire, F. Ronquist, J. Huelsenbeck

Summary of contribution

To describe the nature of the contribution made by the authors in this work, I take a very specific example. In phylogenetic analysis using multiple genes, the question of applying a concatenated analysis (treating the entire alignment as a single big gene) or applying a separate analysis (in this example, this refers to the use of a distinct evolutionary models for each gene) arises often. Let us suppose we have such a dataset—a multiple gene alignment—which we denote as D, and that we are applying a separate analysis. The likelihood has the form of a product over all genes:

$$p(D \mid \theta) = \prod_{g=1}^{G} p(D^{(g)} \mid \theta^{(g)}), \tag{1}$$

where $D^{(g)}$ is the g^{th} gene in the dataset, and $\theta^{(g)}$ is the parameter vector associated with that gene; this parameter vector could specify, say, a GTR+Gamma model (we will suppose that the tree to topology and branch lengths are common to all genes). Now of course, if the number of genes is high, the number of distinct GTR+Gamma models instantiated also becomes high. One might wonder if it would not be better to proceed in more prudent fashion. Taking the homogeneous model as a starting point, we begin by (say) invoking a distinct set of nucleotide exchangeabilities for each gene, while all other parameters of the GTR+Gamma model are common to all genes. But even then, this still amount to $5 \times G$ degrees of freedom for nucleotide exchangeabilities, and one might wonder if an even more cautious stance might be warranted. One idea would be to group genes together, and have gene-groups share the same set of nucleotide exchangeabilities. To be concrete, let us suppose we have $G = 10$ genes, but would like to use only $K = 5$ sets of nucleotide exchangeabilities (20 additional degrees of freedom, relative to the homogeneous GTR+Gamma model). If no information is available to tell us how to group genes for this configuration, the likelihood takes the form of a weighted average of the likelihood under the 5 sets of nucleotide exchangeabilities:

$$p(D^{(g)} \mid \theta) = \sum_{k=1}^{K} w_k p(D^{(g)} \mid \theta, \varrho^{(k)}), \tag{2}$$

where $\varrho^{(k)} = (\varrho_{lm}^{(k)})_{1 \leq l,m \leq 4}$ is the k^{th} set of nucleotide exchangeabilities, $w = (w_k)_{1 \leq k \leq K}$ is a weight vector (which we may interpret as prior probabilities of affiliation to the components), and θ is now the rest of the parameters in the GTR+Gamma model (other than nucleotide exchangeabilities), common to all genes. This is a *finite mixture modeling* approach, not to be confused with the *mixed-modeling* approach, where affiliations are known, or at least fixed based on prior knowledge. (Incidentally, from what the authors have already accomplished, such a finite mixture model should be quite easy to implement using an allocation system analogous to that used to implement the Dirichlet process. Briefly, assuming an initial random allocation has been set, a Gibbs update for the allocation of a particular data element [here, $D^{(g)}$] is done as follows: first, decrease by 1 the count of the number of data-elements affiliated to the component [denoted η_k] before the update; then re-set the allocation variable to the kth component of the mixture with probability $\propto (\eta_k + 1)p(D^{(g)} \mid \theta, \varrho^{(k)})$. Updating in this way implicitly integrates over the weights in eqn. 2, and is equivalent to having a flat Dirichlet prior on them. I personally find this to be an interesting example of the trade-off that can be made between explicitly integrating over the allocations [via the weighted sum in eqn. 2], but then having to perform updates to the weights themselves, versus integrating over the weights implicitly, but having to perform updates to the allocation vector. Moreover, it seems to me that such a system becomes essential if one is interested in exploring finite mixtures over several parameters jointly, for which the likelihood function would have an overly-costly nested sum of likelihoods.)

While this approach provides an account of the fact that we don't know how to group G genes into $K < G$ groups, we still have the problem of determining the dimensionality of the mixture. In other words, how should we determine a value for K that gives a good account of the level heterogeneity, while maintaining a compact model? Of course, a brute-force approach might be taken, such as trying every possible value of K, and using likelihood methods to select the most appropriate model, but this would be very laborious in practice. The Dirichlet process enables one to model the fact that we consider genes as having been "generated" by one of several possible models—as having been generated by a mixture model—without knowing the form of the mixture. Rather, the form of the mixture (including the number of components) is controlled by a higher level parameterization (i.e., a base distribution, and a "granularity" parameter). Strictly speaking, the likelihood function would then be expressed as an integral over all possible mixture models conditional on these higher-level parameters defining the Dirichlet process. But in practice, demarginalization techniques, based on an auxiliary allocation variable, are used within MCMC samplers to handle this integration. The Dirichlet process thus allows one to "fluidify" mixture modeling efforts, and is often referred to as an *infinite mixture modeling* approach.

The work presented in the manuscript by Moore et al. takes this more general infinite mixture modeling approach, and would be directly applicable to this example. Moreover, the work goes well beyond this simple example, and is relevant to any manner of subdividing the available data (e.g., by codon-position, or by both gene and codon-position). This form of mixture modeling is novel (as far a I know), and such ideas are of high relevance to the readership of the Syst. Biol. I also find that the analyses presented are generally rigorous and insightful.

Suggested ameliorations

Although I much appreciate the work, and do not think that any additional calculations need be done, I do feel that the presentation of the material could be improved, and that some of the motivations given for the work need to be revised. Specifically, I think that laying out the progression from mixed-model to finite mixture models to infinite mixture models would better orient the readers as to what has been accomplished (even if the finite mixture model is not in fact explored). The type of mixture studied here differs from previous works, and unless this is made clearer, I suspect that the preconceptions of many readers will lead them to confusion. For instance, with regards to the example I give above, in eqn. 2, Pagel and Meade (2004) have proposed a somewhat similar mixture model in which the likelihood of each site (as opposed to each gene) is a weighted average. This amounts to "factoring-out" the heterogeneity completely (as opposed to only accounting for across-gene heterogeneity in the above example). I am not saying that this is necessarily a better modeling approach (having different objectives in mind), but I think that clarifying the distinction between past mixture modeling efforts and that explored here would be useful.

A main motivation for the work that is presented is an argument regarding the tediousness and potential problems associated with the mixed-modeling approach and its evaluation via Bayes factors. Here, I find the argument confused. I agree that it is overly tedious to try out alternative mixed-models, and compare them via Bayes factor, not to mention that one is likely to overlook some relevant mixed-model possibilities. However, I do not understand why the authors seem to not take seriously the problems with the method used to compute Bayes factors (the harmonic mean estimator). They are aware of these problems, stating that "[...] the harmonic mean estimator of Newton and Raftery (1994) may be biased toward the inclusion of superfluous parameters leading the the selection of over-partitioned composite models." The authors are also aware of existing reliable methods for computing Bayes factors, but nonetheless go on to state "that the application of Bayes factors to phylogenetic inference problems may not sufficiently penalize the inclusion of superfluous parameters [...]". On one hand, we are told that there are computational problems with the evaluation of Bayes factors (is such a way that richer models are over-scored with respect to simpler models), and on the other hand we are told that the Bayes factor itself may suffer from inappropriate statistical properties in model selection. Given that the harmonic mean estimator has been shown to over-evaluate marginal likelihoods, in manner that becomes all the more pronounced as the models get richer, one should expect a model-ranking study utilizing it to elect the richest model as the most appropriate. When reliable computational methods are used, the most appropriate model is not always the most complex one. This has been shown in the statistical literature, but also in the phylogenetics literature. Specifically, the work of Lartillot and Philippe (2006), which the authors cite, gives examples where the harmonic mean estimator would greatly favor the most complex model (way beyond the significance scales sometimes invoked for Bayes factors), but where the thermodynamic integration estimator favors a simpler model. Other recent works have re-iterated the problems with the harmonic mean estimator (Xie et al., 2010; Fan et al., 2010), but I suspect that it may be a while before this message is fully absorbed. I'm not sure why this is, but I suspect that when some researches see that the Bayes factor computed on the basis of the harmonic mean estimator is so markedly in favor of the richer model, they conclude that this model *must* be better than the simpler one even if there are problems with the computational method. They may think something like “even if the harmonic mean estimator is mis-calculating the marginal likelihood, it can’t be *that* bad, and so the richer model probably is better.” Unfortunately, as shown by Lartillot and Philippe (2006) and others, this turns out to be wrong.

That the authors do not appear to take the problems of the harmonic mean estimator seriously is also revealed in the Results and Discussion. In it, the authors speculate as to the differences between mixed-model inferences and those of the Dirichlet process and offer these potential alternatives: “1) Bayes factor selection based on the comparison of marginal likelihood estimated using the harmonic mean estimator may be biased toward the inclusion of superfluous parameters and [hence] selection of over-partitioned mixed models; 2) the Dirichlet process prior model has a tendency for clusters to self propagate, such that data partitions might be attracted to a smaller number of larger process partitions, which may lead to the selection of under-partitioned composite models; and 3) the candidate partition schemes evaluated by means of Bayes factors are separated by substantial gaps in dimensionality (owing to the limited sampling from the pool of all possible partition schemes), which may lead to the selection of over-partitioned composite models.” As I have said, I feel that point 1) can not be ignored. I also have doubts about the exploratory analysis that is presented as digging into point three: They compared the marginal likelihood (computed via the harmonic mean estimator) between the “mean partition scheme” inferred using the Dirichlet process apparatus and other partition schemes, adjacent to this mean partition scheme; these other partition schemes are thus of similar dimensionality to the mean partition scheme. They find that the marginal likelihood is greater for the mean partition scheme than for the alternatives. Ultimately, I think that this exploratory analysis should be excluded from the manuscript, simply because it is based on an unreliable computational approach. However, if the authors insist on keeping it, I think its claims should couched in a more cautious language; it should be stated as an *ad hoc* analysis. The reason for this is that the mean partition scheme identified by the Dirichlet process is subjected to a model-selection procedure as if it constituted a *bona fide* statistical model. The "model" has been defined for the data set under study based on a previous analysis of that data set (i.e., the data has been looked at twice). Also, I feel that to be complete we would need a comparison of the marginal likelihood between the mean partition scheme and the most parameter rich partition scheme. I suspect that if this were done using the harmonic mean estimator, the most parameter-rich partition scheme would be deemed more appropriate than the mean partition scheme; such a finding would give weight to point 1).

Also in this regard, the Dirichlet process approach studied by Moore et al. is presented as an *alternative* to computing Bayes factors. I agree that their approach is more practical than contrasting the Bayes factor for a few mixed-model (or even for a few finite mixture models), but there is nothing that prevents one from evaluating Bayes factor comparing their Dirichlet process approach against, say, a fully homogeneous model. I am not saying that the authors should consider evaluating such Bayes factors (and certainly not with the harmonic mean estimator), but I feel that their presentation of this motivation should be re-phrased to acknowledge this possibility. In particular, one could imagine applying the Dirichlet process to two classes of model formulations based on entirely different rationales, and wanting to rank these two distinct applications; for example, it might be interesting to compute the Bayes between a model that applies the Dirichlet process to the ω parameter of codon substitution models (Huelsenbeck et al., 2006), with one that applies the Dirichlet process to parameters of a codon substitution model that control amino acid fitness (Rodrigue et al., 2010).

Minor issues of potential relevance

Given that my suggestions involve a re-writing of significant parts of paper, I will not spend time on minor editorial slip-ups. I do mention for few points, however.

I wonder about the pertinence of laying out Bayes theorem (twice!) as well as the Metropolis-Hastings algorithm in detail, as done in the manuscript. I can appreciate the efforts to have a "self-contained" manuscript, but I suspect that the Syst. Biol. readership is now well accustomed to the underlying principles of Bayesian inference and MCMC sampling.

I find the author's use of K somewhat confusing. For instance, when the data sets are presented, the number of possible data subsets is given using K. But then, later in the paper, we are shown prior and posterior expectations of K. I would suggest using something like K_{max} in the presentation of the data (and a few other places), such that K would more clearly be associated with the number of components in the mixture.

Towards the end of the paper, the authors state that "as currently implemented, the Dirichlet process prior model requires that the biologist define the set of data partitions prior to the analysis." I understand that this is true for the focus taken in this study, but I don't see what would prevent the authors from setting what I have called K_{max} to be equal to the length of the alignment (which the authors refer to as L). Doing so would amount to a model that is analogous to previous applications of the Dirichlet process in phylogenetics (in reference to the example I have given above, the model would then fall back to a site-wise mixture model, as opposed to a gene-wise mixture model).

Altogether, if the authors were to implement the finite mixture model update mechanism I briefly described in passing, I feel that they have in hand a system that is extremely versatile: the mixed-models are obviously available to them (by fixing the allocation vector, and bypassing update operators on it), the finite mixture model would be available (by fixing the number of component of the mixture, and calling updates to the allocation vector as I've described), along with the infinite mixture model. Still in relation to the specific example that I opened with, both finite and infinite mixtures could formulated as either site-wise mixtures, or gene-wise mixtures, by setting $K_{max} = L$ or $K_{max} = G$. Finishing the paper with a description of these possibilities might provide a more resounding closure, mapping out future directions.

Summary of recommendation

I hope that my comments will not be perceived as overly critical, and emphasize once again that I am greatly impressed with what the authors have accomplished. I have tried to write my summary of the contribution as an example of how the work might be presented without having it gravitate to much around issues of Bayes factors. Specifically, I think the paper would hold up very well as being a study of the most general modeling strategy along the mixed-model, finite-mixture-model and infinite-mixture-model continuum. I have also mentioned that the form of the mixture model explored is novel (and should be clarified); this distinctiveness is also a worthy motivation for the work. I am certain that once issues of presentation have been addressed, the paper will constitute a very strong contribution, of high importance to researchers in the field.

It would be my pleasure to help-out in the review of a revised version of the paper.

Best regards,

Nicolas Rodrigue

References

Fan, Y., R. Wu, M.-H. Chen, L. Kuo, and P. Lewis. 2010. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. in press.

Huelsenbeck, J. P., S. Jain, S. W. D. Frost, and S. L. K. Pond. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc. Natl. Acad. Sci. U.S.A. 103:6263–6268.

Lartillot, N. and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.

Newton, M. A. and A. E. Raftery. 1994. Approximating Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. B 56:3–48.

Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:561–581.

Rodrigue, N., H. Philippe, and N. Lartillot. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc. Natl. Acad. Sci. U. S. A. 107:4629–4634.

Xie, W., P. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2010. Improving marginal likelihood estimation for Baysian phylogenetic model selection. Syst. Biol. in press.