T O P T A L E N T *2015*

Application form

1. Details of applicant(s)

Name: Richel Bilderbeek
E-mail: richel@richelbilderbeek.nl

2. Intended supervisor

Titles, initials, first name, surname: Prof. Dr. Rampal S. Etienne
Host institution: University of Groningen

3. Title of research project

Multi-step versus joint inference methods in estimating phylogenies and derived parameters.

4. Summary of research idea (max. 50 words!)

Parameters involved in phylogenesis of species communities are estimated by either a multi-step process or a joint-inference Bayesian analysis. The latter being a cleaner approach, it is computational more expensive. This research investigates if this expense is vital in parameter inference from phylogenies.

5. Brief description of research proposal **(maximum 1500 words for 5a to 5d; strive for ≤1200 words)**

Parameters of species communities can be estimated from molecular data in multiple ways. This research tries to find the simplest, but not simpler, way to confidently estimate these pararameters.

5a. Introduction and scientific background

Phylogenies are increasingly used as an instrument to estimate variables of species communities, for example speciation rate (Nee (2001), Etienne et al. (2014)). These variables, in turn, may be used by policy makers. Incorrect estimation might thus have consequences in nature conservation.

Estimating these parameters can be a multi-step process: first molecular data is aligned, then the best alignment is used to create a phylogeny. The best phylogeny then is used to estimate a parameter (for a review of this process, see Holder & Lewis (2003)). An alternative is doing a joint Bayesian inference, in which all steps are evaluated at the same time, never settling for a single best. These two methods are shown side by side in figure 1.

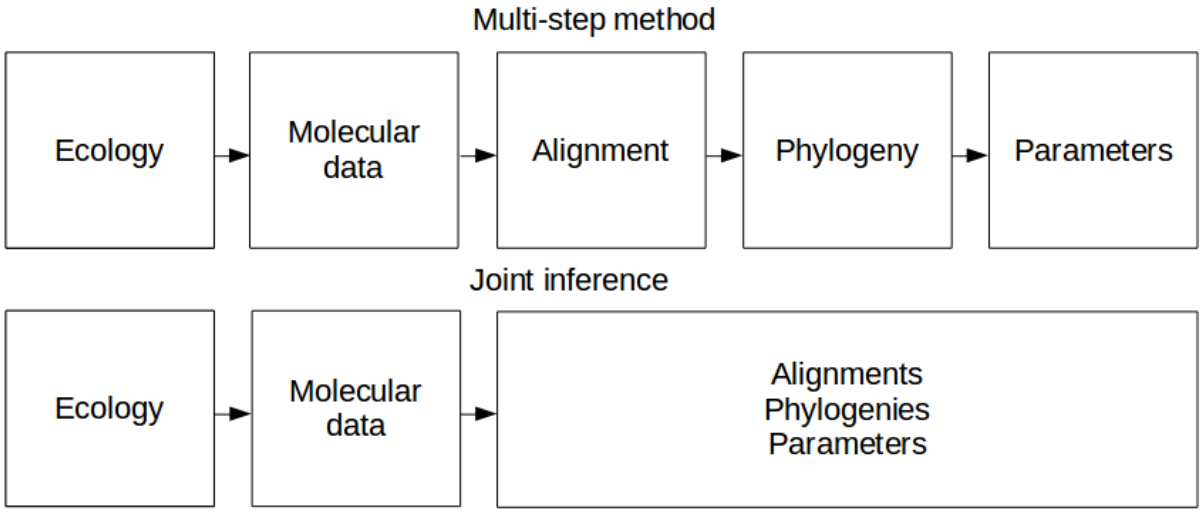

Figure 1: the two possible workflows to estimate parameters from an ecology: a multi-step method takes the best result of a previous step as the input of the next step, where a joint inference method estimates all unknowns concurrently.

It has already been shown by Wong et al. (2008) that discarding the uncertainty of an alignment results in different resulting best phylogenies. Doing a joint-inference from molecular data to both alignment and phylogeny can be done, but according to Drummond et al. (in preparation), the extra computational effort cannot be justified in most cases.

This research starts from an assumed best alignment, focusing on estimating a phylogeny and ecological parameters. Choosing a best alignment is an accepted practice, as discussed in the previous paragraph. **The preffered practice of phylogeny and parameter estimation is not yet known** (Etienne, personal communication). Not only the increased computational effort plays a role in this, also the development of algorithms does so: a multi-step approach does not need the development of new algorithms, where a joint inference does (for example, Lambert et al., 2015).

Simulated data, in which there is full knowledge of the system, is used to test the accuracy of parameter estimation. The way phylogenies are randomly created is not by an individual-based simulation, but by randomly creating a phylogentic structure directly (for example, Stadler & Bokma (2013) or Etienne & Pigot (in press)). Because these null phylogenies are vital to assessing a method's validity, also these must be investigated. Already in Pigot & Etienne (in press), where they propose a new null-model, the conclusions drawn from molecular data differ.

5b. Research questions

- Under which models does a joint inference estimate parameters significantly different?
- Which of these models has its assumptions met in real data sets?
- Of these real data sets, do the random phylogeny creation algorithms ?

5c. Approach
The FOSS tool BEAST2 (Bouckaert et al., 2014) is used multi-step and for joint inference workflows. In the multi-step model it can generate a consensus tree used for parameter estimation. In the joint inference model, the prior developed by Lembert et al. is lacking and needs to be implemented first. BEAST2 its software architecture is designed to be easily extendible, so no major software rewrite will be needed.

The two workflows will result in different phylogenies and parameter estimates. A protocol, similar to Wong et al. (2008) has to be developed to estimate if these differences are significant. To the best of my knowledge, it is not trivial to compare phylogenies and there are multiple ways to compare these. Wong et al. developed a measure of tree distance, that is one of the candidates to be used.

After the two workflows for the first models have been conclusively found indentical or different, this approach is used on other models, like a Birth model or Birth-Death model. These models are selected from the literature by popularity and connection to real-world systems, allowing for tests of the models against nature itself.

With a natural model system and a mathematical description of it, it can be tested if the null-phylogenies generated match the actual system. These random phylogenies should match the actual phylogeny with high probability, otherwise it can be concluded that either the theoretical model is wrong, or we learn that the natural system works under different assumptions as excepted.

5d. Innovative aspects and scientific/societal relevance
This research proposal is about a fundamental investigation. Albeit that a thorough check of an accepted workflow is not sexy, it might change that workflow and lead to different results. These results will result in better-informed policy makers, making better decisions. For us scientists, it will result in a workflow in which our estimations can be made with confidence.

5e. Literature references
1 Bouckaert, Remco, et al. "BEAST 2: a software platform for Bayesian evolutionary analysis." *PLoS computational biology* 10.4 (2014): e1003537.
2 Drummond, Alexei J., and Remco R. Bouckaert. "Bayesian evolutionary analysis."
3 Etienne, Rampal S., Hélène Morlon, and Amaury Lambert. "Estimating the duration of speciation from phylogenies." *Evolution* (2014).
4 Holder, Mark, and Paul O. Lewis. "Phylogeny estimation: traditional and Bayesian approaches." *Nature reviews genetics* 4.4 (2003): 275-284.
5 Lambert, Amaury, Hélène Morlon, and Rampal S. Etienne. "The reconstructed tree in the lineage-based model of protracted speciation." *Journal of mathematical biology* (2015): 70:367-397.
6 Nee, Sean. "Inferring speciation rates from phylogenies." *Evolution* 55.4 (2001): 661-668.
7 Pigot, Alex, and Etienne, Rampal S.. "A new dynamic null model for phylogenetic community structure" (in press)

8 Stadler, Tanja, and Folmer Bokma. "Estimating speciation and extinction rates for phylogenies of higher taxa." *Systematic biology* 62.2 (2013): 220-230.
9 Wong, Karen M., Marc A. Suchard, and John P. Huelsenbeck. "Alignment uncertainty and genomic analysis." *Science* 319.5862 (2008): 473-476.

5f. Time planning

- Week 1-10: add the algorithm of Lambert et al. (2015) in BEAST2
- Week 11-12: protocol development in comparing the two-step method of Pigot & Etienne with this new algorithm
- Week 13-18: protocol development to conclude if the results differ significantly
- Week 19-20: selection of other model to test
- Week 21-25: test models in the two different workflows
- Week 26-27: conclude if the results differ significantly
- Week 28-32: write first draft of paper
- Week 33-36: improve protocol, strengthen statistics
- Week 37-47: write paper

Indicate the total number of words (parts 5a to 5d):

795

6. Budget

	Costs (€)
Salary	180,000
Equipment	20,000
Consumables	0
Travel	0
TOTAL	200,000

[NB. Salary costs of a 4-year Dutch PhD project: €180,000.]

Specification:
Additional to a one-year salary of a PhD student, €20,000 is reserved for renting additional computing power.