GROEP BIOMEDISCHE WETENSCHAPPEN
DOCTORAL SCHOOL BIOMEDICAL SCIENCES
CAMPUS GASTHUISBERG
O & N2, BUS 700
HERESTRAAT 49
BE-3000 LEUVEN

KU LEUVEN
DOCTORAL SCHOOL
BIOMEDICAL SCIENCES

Evaluation form PhD thesis

Please return by e-mail to peter.hoet@med.kuleuven.be before 17/12/2014

Name PhD student: Filip Bielejec
Title thesis: “Continous-time markov chain models for pathogen phylodynamics: model extension, simulation and visualisation”
Name referee: Stijn Vansteelandt

Evaluation score: please indicate in the table below your final score

	Decision		Implication
□	Accepted without revision		Permission for public defense is granted immediately.
X	Accepted with minor revision	I **do not want to review** the revised version	Permission for public defense will be granted via e-mail procedure.
□	Accepted with minor revision	I **want to review** the revised version	
□	Major revision required		Decisions 4 and 5 need to be confirmed by a formal meeting of the examining committee (external members are exempted from attending, but will be consulted via e-mail)
□	Not accepted	New version of manuscript has to be submitted	

Evaluation

General Remarks:

This doctoral thesis makes important contributions to field of phylogenetic inference:

- It provides a methodological framework for improving the biological plausibility of current models of evolution by allowing the substitution rate matrices to be different in different pre-specified epochs. This builds upon earlier ideas by Goode et al. (2008), but is more advanced in that it, for instance, allows for inferring the evolutionary tree. As in nearly all phylogenetic inference, the computational demands are enormous because of the need to integrate over the unknown ancestral history. In view of this, Bayesian Markov Chain Monte Carlo algorithms are employed in combination with a massively parallel computing approach. Simulation studies confirm the adequate performance of the approach, and data analyses confirm the evidence for variability in the evolutionary parameters across epochs. I much welcome this contribution as more realistic evolutionary models are likely to translate into more accurate phylogenetic inferences. It is a pity that evidence for this is lacking from the manuscript (see my later comments), but I do not view this as prohibiting.
- It provides a software application, SPREAD, to analyze and visualize phylogeographic reconstruction. Although this makes up a relatively small contribution to the printed manuscript (in terms of number of pages), it is the result of a significant effort and will make an extremely valuable contribution to the field as it will make these highly demanding methods accessible to the end user. Not surprisingly, this work has also been well picked up by the scientific community.
- It provides a GUI to simulate sequence alignments under a wide variety of evolutionary models. Again, I believe this is the result of a significant effort that I view as a major advance, because the lack of tools to simulate data under a given evolutionary model also prohibits the evaluation of the performance of statistical methods to infer evolutionary parameters and trees.

The doctoral thesis is advanced at various levels: the statistical MCMC technology, the state-of-the-art model evaluation through Bayes factors, the use of a massively parallel computing approach, the visualization of spatially-mapped trees (and associated uncertainty), ... It is especially strong in its valorization component and will undoubtedly prove very useful to a very broad group of researchers in the field. In summary, I see no major issues and recommend that the candidate be permitted to proceed to the oral defense. My later comments are merely suggestions for further clarification.

Major comments:

- The thesis starts with a very nice introduction, which I think is very insightful for the reader who is not familiar with phylogenetic and/or Bayesian inference. In spite of this, I believe that the readability of the later chapters would be further improved if the introduction:
    - briefly described the evolutionary models, such as GY94, that are later used;
    - briefly described how to accommodate time-stamped sequences, which is now considered in the analyses in Chapter 2, but not really discussed.
    - gave a bit more detail on phylogeography, e.g. how do the models look like and what is the connection between phylogeography and epoch models (some connection is described in Chapter 2, but remains a bit vague).

  I will leave it to the author to decide whether or not to incorporate this.
- It is sometimes unclear in chapter 2 whether the analysis is performed under a fixed evolutionary tree or not. This is especially so in the evaluation of Bayes Factors in Table 2.3. It would also have been nice to understand better whether the use of the epoch model (versus not) has an essential impact on tree reconstruction. Somewhat related, it would have been nice to learn from the simulation study how mislead one could be if the data were generated under an epoch model, but analyzed under a simpler model. I am not asking for additional analyses here, but merely for additional clarification or insight.

Minor comments:

P6, lines 5-7: this is not correctly phrased as the parameter values do not have an associated probability in the frequentist likelihood framework.

Expression (1.12) is confusing in that indices i and j appear in the lefthand side, but not in the righthand side. Likewise in expression (1.13).

P20, 2nd display: an equality sign is missing on the 2nd line.

I would find it useful to see a brief discussion on the plausibility of (1.22).

If I am right that matrix P(t) on the bottom of page 22 is four-dimensional, then why not write it in full, which will be less confusing?

P25, the statement at the end of the penultimate paragraph (`most plausible hypothesis for the data’) is vague.

P28, I understand what T(n) is, but would nevertheless find it helpful if it were stated more explicitly.

P31, If the phrase `proper prior’ is used, then perhaps it should be defined.

P32, Since the author assumes that the reader may have no prior knowledge of Bayesian inference, I believe the meaning of Figure 1.11 should be explained in more detail.

P33,. I think the notation \theta[t] has not been defined.

There is a typo in the lefthand side of (1.37).

I would find it useful to get a bit more detail how a model for the likelihood of Y and for the prior of \Phi might look like.

P43. What is compositional heterogeneity?

P44. I would find it useful to receive a bit more guidance concerning the connection to phylogeographic inference. Also in Section 2.4.3, the description of the proposal by Bahl et al. (2011) is too brief for me to appreciate why the epoch model is more appropriate.

Display (2.5): I believe this display implicitly assumes that the time T1 is a priori known. It would be good to state this explicitly.

P57. The sentence `To determine the posterior odds, ... of that model’ is vague.

P85. Riemannian -> Riemann

P86. The last sentence of Section 5.1.3 is not sufficiently clear. Please make clear why the use of mixed effects models could allow this.

P88. It would be useful to receive a bit more detail what is meant by `infinite hidden Markov Model’.