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: David Posada

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

	Decision		Implication
☐	Accepted without revision		Permission for public defense is granted immediately.
☒	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	

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

Evaluation

General Remarks:

This thesis describes a niece piece of work aiming to improve current models and strategies for the analysis and interpretation of sequence data from pathogens, in particular fast-evolving RNA viruses. The thesis is well written and self-contained. The objectives of the thesis are original and relevant. The introduction is well-balanced in general and offers a good review of the different ideas and methods used in the thesis. The next three chapters describe the three main contributions of the thesis, corresponding to three different publications. Chapter 2 describes the implementation of a sophisticated n-epoch model very suitable for pathogen phylodynamics. Chapter 3 describes very succinctly the implementation of a useful visualization tool that should facilitate the interpretation of the phylogeographic analysis of not only virus but also of other organisms. Chapter 4 describes a evolutionary simulation tool that is not specially new but that can be very useful due to the strong connection with BEAST. Finally, Chapter 5 offers mainly an overview of potential extensions of this work. Throughout this work the candidate demonstrates a good understanding of molecular evolution, demographic and phylogeographic concepts. The objectives of the thesis have been appropriately achieved in time. The accomplishments are mainly technical, but nevertheless, they will be very useful for other researchers aiming to study viral evolution. No particularly new theoretical or biological insights seem to have been produced, but this is not to diminish its relevance. Importantly, the thesis has already resulted in three publications in peer reviewed journals.

Major comments:

1) I missed a brief primer of Bayesian model selection in the introductory chapter. Techniques such Bayes factors, path sampling or stepping-stone sampling are used throughout but never explained. This should be very easy to fix.

2) In the simulations in Chapter 2 you assume you know the true epoch times, but this information is rarely available in real life. Simulation 1 addresses the impact of model misspecification partially, as the true process does not have epochs. It could be a good idea to further check the impact of model misspecification, in particular simulating a number of epochs that do not correspond with the assumed ones (you later mention this idea in the final chapter but do not end up testing its potential biasing effect). Also, simulating under a non-epoch model –for example where different process can apply to

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

contemporaneous branches, would be fundamental to understand the robustness of the model. This does not need to be done before the defense, but it could be a nice extension for the future.

3) In general, there is some repetitive text, sometimes with identical phrases/paragraphs, between the introduction and some of the other chapters. On the other hand, some relevant aspects of the research are discussed in the general discussion in Chapter 5, instead of in the corresponding 'research' chapter. This is true in particular for Chapter 3, were a very concise description of SPREAD is provided. I understand these redundancies and fragmentations are a byproduct of reproducing papers as chapters, but however these issues have been corrected in some occasions. For example, chapter 2 is not identical to the SysBio paper. I suggest you try a bit to further consolidate the different chapters.

4) In chapter 3 I missed the inclusion of some biological examples illustrating the novel or faster insights we could obtain with SPREAD.

5) The simulation software piBUSS described in Chapter 4 can be very useful, in particular in regards to its connectivity with BEAST. The possibility of simulating discrete phylogeographic traits is very convenient. Here it would have been interesting to see, for the sake of comparison and interpretation, the correspondence between simulated and estimated tMRCAs when the true model has one epoch with omega=0.5 (or even better, omega equal to the weighted mean of the n-epoch model). This does not need to be done before the defense, but it could be a nice extension for the future.

Minor comments:

1) Page 14: I would say that the term 'Big Data' does not apply to sequence alignments. It might do for NGS files etc., but not for the type of data that BEAST uses as input. Not yet. Moreover, you focus on HPC architectures like GPU and multicores, which are in some senses opposite to Big Data architectures (i.e., Hadoop Map Reduce).

2) Page 25, 31: likelihood should be $P(\mathbf{X}|\Theta)$, not $P(\Theta|\mathbf{X})$.

3) Page 34: on the left-term of equation 1.37, the denominator should be $P(\theta[t]|\mathbf{X})\, q(\theta^*|\theta[t])$, not $P(\theta[t]|\mathbf{X})\, q(\theta[t]|\theta^*)$.

4) Page 52: according to the legend in Figure 2.3, a two-epoch plot is missed.

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

5) The bibliography would have been be easier to consult if last names came first than the initials, as usual.

6) Page 137: Figure B.12 should be in color. Otherwise it is very difficult to grasp.

7) Page 140: Figure B.13 should be in color. Otherwise it is very difficult to grasp.

8) Page 143: Figure B.15 should be in color. Otherwise it is very difficult to grasp.

9) Unless I missed it, the appendices are not cited in the text.

Vigo, Spain, 11 January 2015

David Posada