1 November, 2020

Dear Thesis Committee

The thesis submitted by Thibault Latrille addresses a number of important topics in phylogenetic analysis. In broad terms, these projects involve developing a unified perspective that connects population genetics and phylogenetics, with the aim of being able to use phylogenetic analyses to understand how the evolutionary process occurs at a genetic level in terms of mutation, selection, and drift. After a thorough and well written introduction, three different projects are described. The first project involves creating a more complicated codon model that includes a more accurate characterisation of selection, based upon a mean field approach. This model does better at identifying mutation bias and average mean fixation probability than standard simpler codon models. The second project involves an ambitious attempt to include changes in population size in substitution rate models. Although the population size changes and differences did not seem to be estimated accurately, there are interesting reasons why this may be the case that were explored and may provide insight and directions for future work. The third project was a more theoretical study of the role of population size and expression levels in protein evolution, including the effect of changes in population size. Such projects can provide clarity in terms of how these parameters affect the substitution process, and what aspects need to be included in a successful treatment.

The thesis is, in general, well written and well organised. The work is interesting and solid, and limitations of the work are described honestly and completely. I had a few concerns and some minor quibbles, as described below, but I believe that the work makes a significant contribution to the field.

## The Introduction

The introduction is one of the most complete and comprehensive introductions that I have come across in a PhD thesis. Having this clarity at the beginning provides a very welcome context for reading the rest of the thesis. I have a few minor issues:

When talking about the maximum mutation rate and its relationship to genome size, I might mention the theory of mutational meltdown, the physics-based argument about what happens when the mutation rate exceeds a certain amount. In addition, if one looks at the rate of mutation is a function of genome size, much of the range of the of these parameters is defined by viruses going from viroids to RNA viruses to DNA viruses. Of course, viruses have limited amounts of transposable elements, suggesting that the purging of transposable elements in larger populations is not the appropriate mechanism over most of the range of genome sizes.

I think care has to be taken when talking about evolution of mutation rates and exploration, as ability to explore cannot be directly selected. This goes back to all of the debates of the evolution

UCL Department of Infection & Immunity
Cruciform Building, 90 Gower Street
University College London, WC1E 6BT
Tel: +44 (0)20 3109 2206

of evolvability. To me, the most reasonable picture is to see the ability to explore as something that hitchhikes on the advantageous mutations that result.

When discussing the selection acting on proteins, it is important to remember that a large number of proteins are not folded under physiological conditions, or may have large regions that are not folded. It is easy to get the impression that proteins tend to be folded because the structures that we see of proteins are generally of proteins that do fold. Likewise our image of proteins tends to be of globular proteins. When talking about aggregation and the tendency for the exterior amino acids to be hydrophilic, it is important to remember that these tenancies may not hold for the many membrane proteins. Finally, it is not true that the standard genetic code is used almost universally by all organisms. There are many organisms that use other genetic codes, and in addition many of our genes are mitochondrial, which also use a different genetic code.

The thesis tends to use the term 'positive selection' as synonymous with diversifying selection. It is important to remember that there are other types of positive selection, some of which may be more important than diversifying selection, which may be less often considered simply because it is harder to model.

On page 40, the thesis states that 'a default of non-synonymous substitutions, leading to omega less than one, mean the protein is globally under purifying selection.' This is not true. One can have many regions undergoing positive selection where the overall omega is still less than one.

I have trouble understanding section 3.4.3. If the change in the landscape results in lower fitness of this current sequence on average, then you are not talking about epistasis within a protein but between proteins, as the current sequence includes the entire sequence of the protein. There is also a problem of time reversibility. If substitutions are on average adaptive, does that mean that the fitness of the organism keeps increasing in a monotonic way? Or that the sequence would become increasingly maladaptive in the absence of such substitutions? This might be the case for antagonistic co-evolution, but this is a specific circumstance. I am suspicious of terms such as 'which can only' which seems to imply that the fitness of the target amino acid could never ever increase.

If one is talking about epistasis and entrenchment (pages 51 to 52), one really should reference Goldstein and Pollock 2012.

I think there is a serious misconception about the approach and use of the Miyazawa Jernigan mean field approximation for pairwise contacts. Their approach was never meant to be used to calculate detailed protein thermodynamical properties. It was originally derived in order to estimate the magnitude of the hydrophobic interaction in general for proteins, rather than to provide insight into any specific protein. It was then used with varying degrees of success for a specific goal, that is protein structure prediction. But I do not imagine anyone would use it to, for instance, estimate the thermodynamic consequences of an amino acid change.

When discussing the relationship of stability and fitness, I would look at and reference the work of Tawfik. You make a claim that because of translation errors, proteins are more stable than they would be without such errors. Such a claim requires either evidence or a reference.

When talking about protein thermodynamics, it is important to remember that proximity can exist even between amino acids far apart in the folded structure, inasmuch as they may be in contact in unfolded structures.

## Studies: Chapter 7

In order for analyses to be tractable, models generally have to be greatly simplified. An important question then arises, how much do these simplifications affect the results? A powerful approach to this problem is through simulations, where data is generated with a more complicated model and then analysed with the simpler model, and the results of the simpler model analysis are compared with the model that was used to generate the data. Chapter 7 describes an example of this, where a site-specific mutation selection model is used to generate the data which is then analysed using a simple codon model that is in common use. In particular, the thesis deals with the question of distribution of nucleotides at different codon positions, and how these distributions could be different even if the mutation process is the same. The simulation results clearly exhibits differences in nucleotide frequencies of the three sites due to selection acting at the amino acid level. As the parameters and the simulation can be varied, it is possible to explore the impact of different mutation biases and selection strengths on the observed results. It appears that the MG model does a good job at estimating omega although the mutation bias is significantly underestimated.

A more complicated model was then described that represents the average substitution rate between any two adjacent codons, requiring an expanded but not unreasonable number of adjustable parameters. The model was fit using maximum likelihood, providing estimates of the various parameters. This model does better at estimating the mutational bias as well as the true mean fixation probability.

I have some questions about the mean field derivation which computes averaged rates for the substitutions between different codons. Although it seems to yield good estimates of a number of important parameters, it is important to mention that the data involve substitutions that do or do not occur, and the average probability of a given substitution along a finite branch is not equal to the exponential of the average rate. It would be a good idea to expand a bit on what exactly the mean field model is, what is being averaged, and how this relates to modelling the data on a tree. I am curious about the modelling of the influenza data, specifically the assumption that multiple mutations cannot occur. There is some work by Nick Goldman and his group indicating that this assumption may be poor one for RNA viruses, inasmuch as there is diversity within a single host allowing for a second mutation to occur in a variant, resulting in a double mutation.

## Studies: Chapter 8

In addition to reconstruction of the phylogenetic relationship between different organisms, it would be interesting if we could reconstruct other aspects of the evolutionary process, especially aspects that relate to biological and ecological parameters. Chapter 8 describes the construction and testing of models that incorporate changing selection and mutation rates, the former in terms of changing effective population size. These models are first analysed using simulated data, and then applied to biological data. The analysis with a simulated data demonstrates that branch lengths can be accurately reconstructed, although patterns of change of the effective population size is problematic in the presence of epistatic interactions. Analysis of biological data demonstrated promise, although the magnitude of the changes in population size seemed unrealistically small.

As a small point, I do not think it necessary to classify individual substitutions into different categories, as all substitutions occur with a mixture of mutation, selection, and genetic drift. In terms of the literature, I might make a connection to the extensive research on the rate of evolution and effective population size, as there is a direct connection between the rate of evolution and the dN/dS ratio. It is not true that site-specific amino acid fitness profiles require aggregation into different categories of sites; see, for example, Tamuri and Goldstein 2009.

Looking at the results of the placental mammals dataset, it did raise a question in my mind that mutation rate and effective population size was so correlated. While there is a possible explanation in the effect of size on generation time, it did make me worry about crosstalk between these two terms. For that reason I was very interested in seeing the relationship between mutation rate and effective population size for the isopod species. I would have liked to see how mutation rates varied throughout this tree, and in particular whether there was also a relationship between mutation rate and effective population size.

In general I would have liked to see more data presented, more in the style of the data presented for the placental mammals. It is always important to distinguish between statistically significant correlations and substantive correlations, as statistical significance can represent amount of data more than size of the effect. It would have been nice to know, for instance, how much the variation in effective population size is explainable by habitat or pigmentation or ocular structure.

Concerning the difficulty of reconstructing effective population sizes with this model, an important question is why the method based on classical codon model seem to perform better. It would seem that many of the explanations given for this difficulty with the current model would also be true of the classical codon model. It would be good to specifically address the question, what aspects of the biology would affect the described method more than the classical codon model, as a way of better identifying the current limitations and how they could be overcome.

## Studies: Chapter 9

Chapter 9 describes a nice piece of work looking at the dependence of evolution rates on population size and expression level. By simplifying the model so that the stability is sitewise

additive but the fitness is a non-linear function of the stability, it is possible to generate analytical expressions even if they cannot be exactly solved. In particular, this chapter describes the susceptibility defined by how the substitution rate changes with the log of the effective population size. One nice result is how similarly expression levels and population size affect the rate of evolution. I also appreciate the comparison with more empirical data, although a susceptibility of 0.02 in primates still seems very shallow.

One of the aspects that induces a dependence of the evolutionary rate on population size is running out of acceptable amino acids. I would be curious to see how the results change and whether the susceptibility is larger if the number of acceptable amino acids changes from twenty to four or five, where the other amino acids were just considered unviable. In addition, as the thesis makes clear, there are many aspects to fitness beyond protein stability. Sites that have functional importance, for example, might also have a much-reduced number of acceptable amino acids compared with what one would expect based on stability criteria alone. These factors might act to increase the susceptibility closer to that which seems to be observed in biological sequences.

I am curious, based on equation 9.14, it would seem that there would be a dependence on temperature so that organisms live at higher temperature would have a greater susceptibility than those that live in colder conditions. I realise the temperature difference is small in absolute units, but maybe it is possible to compare similar proteins in thermophiles and psychrophiles. I am also curious whether virus proteins might give different results. There is some evidence (Tokuriki et al. 2008) that virus proteins have somewhat different biophysical properties which might translate to a different robustness to substitutions.

The results for an increase or decrease in population size should be compared with that described in Goldstein 2013.

In the discussion of the connection between evolutionary biology and physics, I would consider it important to note the role of diffusion models in population genetics, such as in the derivation of fixation probabilities.

One small point, absolute temperature is measured in Kelvin, not in degree Kelvin.

## Conclusions

I am unclear about the discussion of mutation limited evolution rates on page 145. The expected waiting time until the next mutation occurs is equal to the mutation rate times the population size, which is much faster than implied by the hundred- to thousand-million year timescale. In addition, even without epistasis, mutation selection balance would imply that again only a few substitutions would be required for a new equilibrium to be achieved. It is not true that each side and the sequence has to adapt. In the language of chapter 9, only sufficient substitutions would have to occur to change X to the new equilibrium value.

I think the distinction between phenomenological models and mechanistic models is an important one, and it is useful to frame different models in this context. I would say, however, that there is

no such thing as a pure phenomenological or pure mechanistic model. The most phenomenological models are still based on the overall process of sequence change, and the fact that there is, for instance, a difference between synonymous and non-synonymous substitutions. The most mechanistic models are still highly coarse-grained, abstracting away a certain amount of biological, chemical, and physical phenomena and replacing it with descriptions of how things are observed to behave at a higher level.

Yours sincerely,

Richard A. Goldstein
Chair, Pathogen Evolution