Evaluating the effect of sequencing error on estimating phylogenies

Eleni Georgoulopoulos *egeo2509@uni.sydney.edu.au*
Associate Professor Michael Charleston
School of IT
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

INTRODUCTION

Phylogenetics

•Phylogenetics involves utilising various forms of biological data so as to ascertain the relatedness of all organisms. There is a plethora of methods available to estimate phylogenies. Sequencing, a process of determining the genetic composition of an organic sample, is often carried out first. The sequences obtained from the organisms of interest are compared to each other as a measure of similarity as part of a process known as alignment. Phylogenies are then inferred; some of the existing methods include maximum likelihood, parsimony analysis, and the relatively more recent, and highly disputed [1; 2; 3], molecular clock methods. While such methods have existed for some decades and methodologies include scrupulous attempts to reduce error it is virtually impossible to guarantee this; there is no "correct" phylogeny against which to compare and ensure validity. Effects of sequencing error are not as well understood compared to other areas of the process, particularly in regard to the manner in which they affect phylogenetic estimation.

Significance

•In 2008 the platypus genome was sequenced and annotated resulting in an intimate understanding of the platypus genome [4]. Later than year Sang and Blecha showed that cathelicidins, compounds produced by platypuses as part of their immune response, are known to be highly active against a broad spectrum of bacteria and fungi including highly resistant strains of *Staphylococcus aureas* [5], an ever-growing concern, particularly in hospitals. Furthermore these compounds do not harm eukaryotic cells such as our own. Pharmaceutical companies are currently developing antibiotics from these cathelicidins as a direct result of this research. This is a recent example of how such research can be beneficial for society in general but can also be financially rewarding.

EXPERIMENTAL DESIGN

•The aim was to produce a controlled experiment in which the effect of error on conserved and variable alleles, and on long and short sequences, could be observed (see Figure 1.1). Two types of error were observed:
•indels, i.e. insertions/deletions
•base-call error, in the form of point mutations

The experimental design emulates the process of inferring phylogenies. It is an automated process written in the python programming language [6] and can process sequence data or data in tree format. Tree format data must first be processed by Seq-Gen [7] which will produce sequences for the trees using Monte Carlo simulation [8]. These sequences, as well as sequences from real data, such as the published data sets that were used for this experiment, are then passed to READSEQ [9] for conversion to FASTA format bringing the sequencing phase to a conclusion. Next indel error is added to one duplicate of the data and base-call error to another using a custom python script. Six different error rates (.001%, .01%, .1%, 1%, 3%, 5%) are used for each. Alignment is then carried out by MUSCLE [10] Next, dnaml, dnamlk and dnpars of the PHYLIP phylogenetic inference package [11] are run on each of the data streams. Trees are then displayed using DensiTree [12] (see Figure 1.2 in Results section for example). Consensus trees are produced using PHYLIP's CONSENSE method [11], then passed back to the respective programs which produced the phylogeny for each stream with its sequence dataset to add branch lengths to this consensus tree.

- Finally trees are grouped using a python script and unrooted trees produced by dnaml and dnapars are rooted using PHYLIP's retree [11]. Analysis using python's DendroPy package [13] was then carried out to investigate any changes to the tree topology or branch lengths.

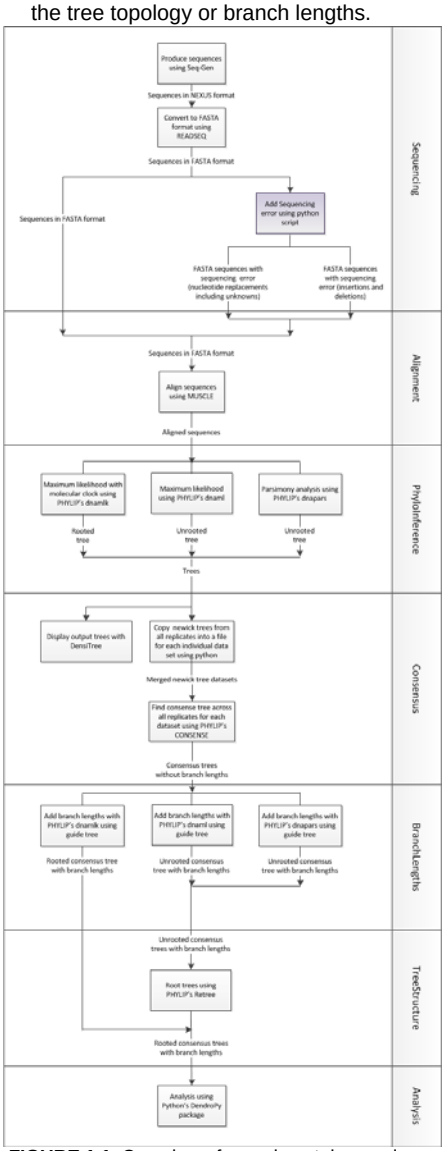

FIGURE 1.1: Overview of experimental procedure demonstrating the treatment of data streams.

RESULTS

As demonstrated in Figure 1.2, erroneous data was found to have a consistent and significant effect on branch length and in rare instances also alters the tree topology. While the different phylogenetic inference methods gave fairly consistent and similar results for "correct" data sets the methods varied widely in their interpretation of data containing sequencing error as seen in Figure 2.1.

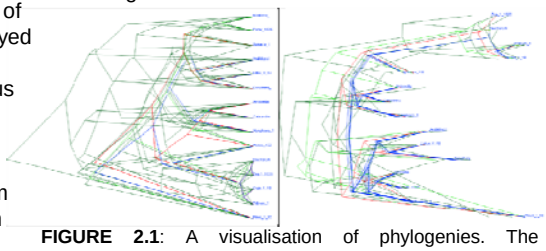

FIGURE 2.1: A visualisation of phylogenies. The visualisation on the left was given by dnamlk; indel and base error trees are overlaid on the phylogeny given by the "correct" data set. The visualisation on the right was given by dnaml with indel and base error overlays. The original data set used for both visualisations was the same well-conserved plant data set. The same error rates were also used for both visualisations. Colours of overlaid trees change to indicate frequently-occurring and well-supported trees.

•It was also found that conserved sequences are more susceptible to the effects of sequencing error until a threshold is passed in terms of the error rate. At this point the phylogenetic inference programs tested produce very exaggerated branches or deviant topologies from that of the "correct" phylogeny. Indel error was also found to be particularly pernicious to the estimation of the phylogeny.

FURTHER WORK

•This research examined indel and base-call error in isolation. Interesting trends may arise when both are present together.
•Clearly there is a great difference between the effect of error on one phylogeny program to the interpretation given by others. Reasons for this occurrence would be worth further study.
•Introducing bootstrapping, which has been documented to smooth the effect of error, would be another worthwhile variation on this experiment.

REFERENCES

[1] D. Graur and W. Martin, "Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision," *Trends in Genetics*, vol. 20, pp. 80–86, Feb 2004.
[2] M. van Tuinen and E. A. Hadly, "Error in estimation of rate and time inferred from the early amniote fossil record and avian molecular clocks," *Journal of Molecular Evolution*, vol. 59, pp. 267–276, Aug 2004
[3] F. Rodriguez-Trelles, R. Tarrio, and F. J. Ayala, "A methodological bias toward overestimation of molecular evolutionary time scales," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 8112–8115, Jun 2002
[4] W. C. Warren, *et al.*, "Genome analysis of the platypus reveals unique signatures of evolution," *Nature*, vol. 453, pp. 175–U1, May 2008
[5] Y. Sang and F. Blecha, "Antimicrobial peptides and bacteriocins: alternatives to traditional antibiotics," *Animal Health Research Reviews*, vol. 9, pp. 227–235, Dec 2008.
[6] P. S. Foundation, "Python programming language ¨U official website," August 2010.
[7] A. Rambaut, "Seq-gen," July 2010.
[8] A. Rambaut and N. C. Grassly, "Seq-gen: An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees," *Computer Applications in the Biosciences*, vol. 13, pp. 235–238, Jun
[9] D. Gilbert, "Readseq: Biosequence conversion tool," July 2010.
[10] R. Edgar, "Muscle: Protein multiple sequence alignment software," August 2010.
[11] J. Felsenstein, "Phylip home page," July 2010.
[12] R. Bouckaert, "Densitree development page," October 2010.
[13] J. Sukumaran and M. T. Holder, "Dendropy phylogenetic computing library," August 2010.