Differentiating between error and information in DNA sequence data

Taika von Königslöw°, Daniel Ashlock* and Paul Hebert
Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada, N1G 2W1
*Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario Canada, N1G 2W1
°E-mail: tvonkoni@uoguelph.ca

Introduction

DNA sequencing error estimates

- There are multiple steps in the process of obtaining a DNA sequence from a specimen, **Figure 1.**

Figure 1: An illustration of some steps in the process of obtaining a DNA sequence from a specimen.

- In theory, error can be introduced at any stage in the DNA sequencing process (Kunkel, 2004).
- In practice, we will focus only on error introduced in the determination, interpretation and editing of DNA sequences and assume all other sources of error negligible.
- Raw DNA sequence data often includes an electropherogram of the sequence and unbiased position-specific error estimates produced by the program PHRED (Ewing et al., 1998).
- DNA sequence data requires interpretation and editing before returning a 'finished' sequence suitable for analysis. It is common to sequence both the forward and reverse complement of a segment of DNA in order to reduce error, **Figure 2**.

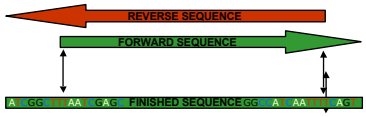

Figure 2: An illustration of complementary DNA sequences aligned to help resolve ambiguous bases and build a finished sequence for use in sequence analysis. The area between the arrows is the overlapping region in which the two sequences can be used to resolve random error introduced in the process of DNA sequencing.

DNA sequence editing

- The quality, clarity and spacing of the peaks from which bases are called in a sequence electropherogram can vary dramatically, **Figure 3**.

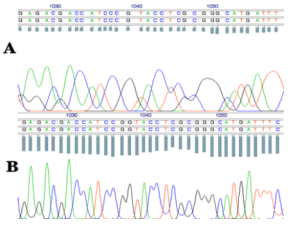

Figure 3: **A)** An example of the electropherogram trace of a DNA sequence with poorly resolved peaks; **B)** An example of the electropherogram trace of a DNA sequence with well resolved peaks. Above both electropherograms are quality scores calculated from error estimates (e.g.*Phred* scores), which can be used to influence decisions in the editing process.

- Quality can vary across a sequence and so it is possible that random and systematic error can be incorporated into a finished sequence, **Figures 4** and **5** respectively.

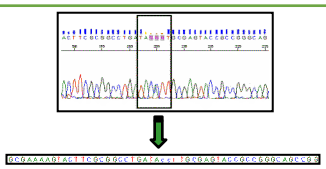

Figure 4: An instance in which a random error could be incorporated into a finished DNA sequence. The bars above the sequence base calls represent the quality score as determined by the calculated *Phred* scores, high bars represent high quality base calls and low bars represent low quality base calls. The boxed region in the middle highlights the occurrence of a low quality score that could be edited to include an erroneous base call.

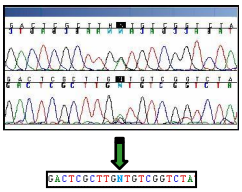

Figure 5: An instance in which a systematic error could be incorporated into a finished DNA sequence. The base highlighted in black on both sequences marks the position of an uncalled base in both the forward and reverse sequence.

Methods

Model

- We propose to address the problem of post-editing sequence error by building a *profile hidden Markov model*.
- This model would complement the error probability estimates assigned to the raw sequence data (*Phred* scores). Thus, providing another level of error detection and increasing confidence in sequences with low sequence variation where an error is more likely to influence results in analysis.

Detecting DNA sequence error

- The steps to detecting error in a set of DNA sequences are as follows:
  1. Detect shallow sequence divergences in a set of sequences
  2. Remove the sequence(s) of interest
  3. Build a pHMM from the remaining sequences
  4. Evaluate the probability of nucleotide variation at all sites in the sequence(s) of interest
  5. Sites detected are then recommended for re-examination

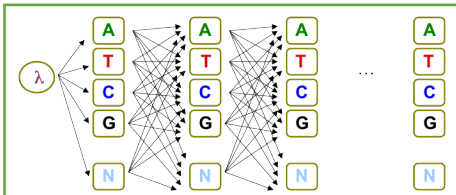

Figure 6: The diagram above is a possible design for the profile hidden Markov model to be built for discriminating systematic and random error from true variant sites. λ represents the initial state prior to the first observed state in the data set. Each arrow represents a possible transition path between states and each square represent a possible emission character which are the 4 DNA nucleotides Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) and a fifth character (N) representing an unresolved site. There are no insertion or deletion states introduced in this model since they are not necessary for modeling the sequences on which this model will be built, but can they be included for other data sets.

Testing the Model

Testing

- The accuracy of the model can be estimated directly by its ability to detect error introduced into a set of DNA sequences.

Outcomes

- Testing the model produces probability estimates for all types of situations in which DNA sequencing error can mask the true level of variation in a set of sequences.

Significance

- Since error can be incorporated into a data set post sequence editing, as shown in **Figure 3 - 5**, another level of error detecting is important.
- In particular, evolutionary analysis focused on rare events may be greatly disturbed by even very low error rates due to the low signal-to-noise ratio (Clark & Whittam, 1992).

Other Applications

- Other applications include determining the suitability of sequences for analysis of sequence variability, **Figure 7**.

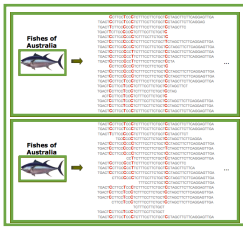

Figure 7: **A)** A set of sequences that are suitable for analyzing sequence variability; **B)** A set of sequences that are not suitable for analyzing sequence variability.

- This model can also be used to identifying patterns of informative sites, **Figure 8**.

Figure 8: An illustration of the detection of patterns in sequence variability.

Selected References

- Ewing, B., *et al.* (1998) *Genome Res.* **8**: 175- 185.
- Clark, A. & Whittam, T. (1992) *Mol. Biol. Evol.* **9**: 744- 752.
- Kunkel, T. (2004) *J. Biol. Chem.* **279**: 16895- 16898.
- Ward, R.,*et al.* (2005) *Phil. Trans. R. Soc. B.* **360**: 1847- 1857.

Acknowledgments

I would like to thank my Thesis Advisory Committee members and the grad students in the Hebert Laboratory for editing and advice.