Differentiating between error and information in DNA sequence data

Taika von Königslöw
M. Sc. Candidate, Department of Integrative Biology
University of Guelph, Ontario, Canada

Thesis Advisors:
Dr. Daniel Ashlock
Dr. Paul Hebert

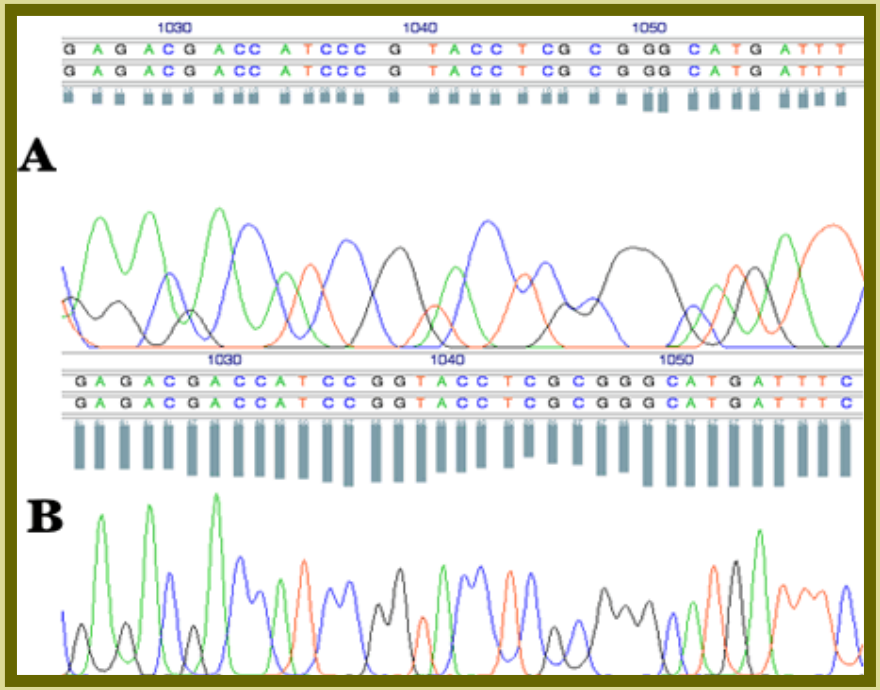
Committee Members:
Dr. Alex Smith
Dr. Gary Umphrey

Introduction / Background

- DNA sequencing

Introduction / Background

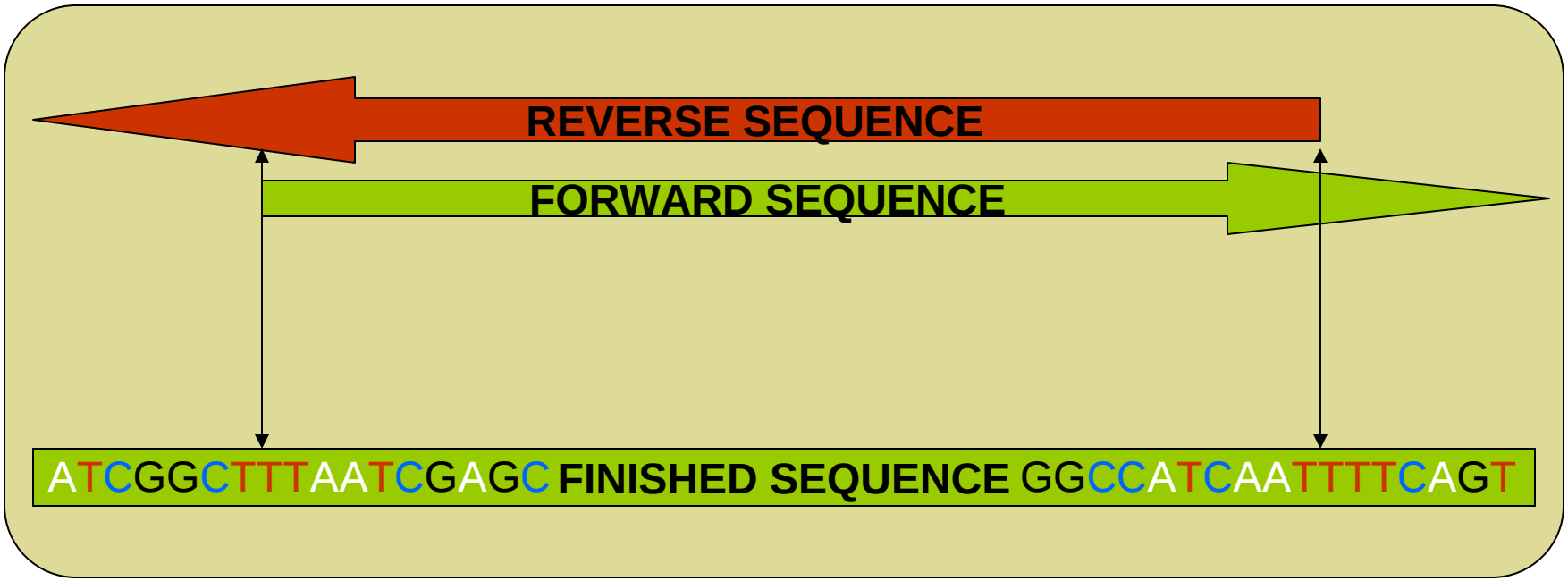
- Phred scores



Source: http://www.nucleics.com/images/site_images/CounterTrace-before-after.gif

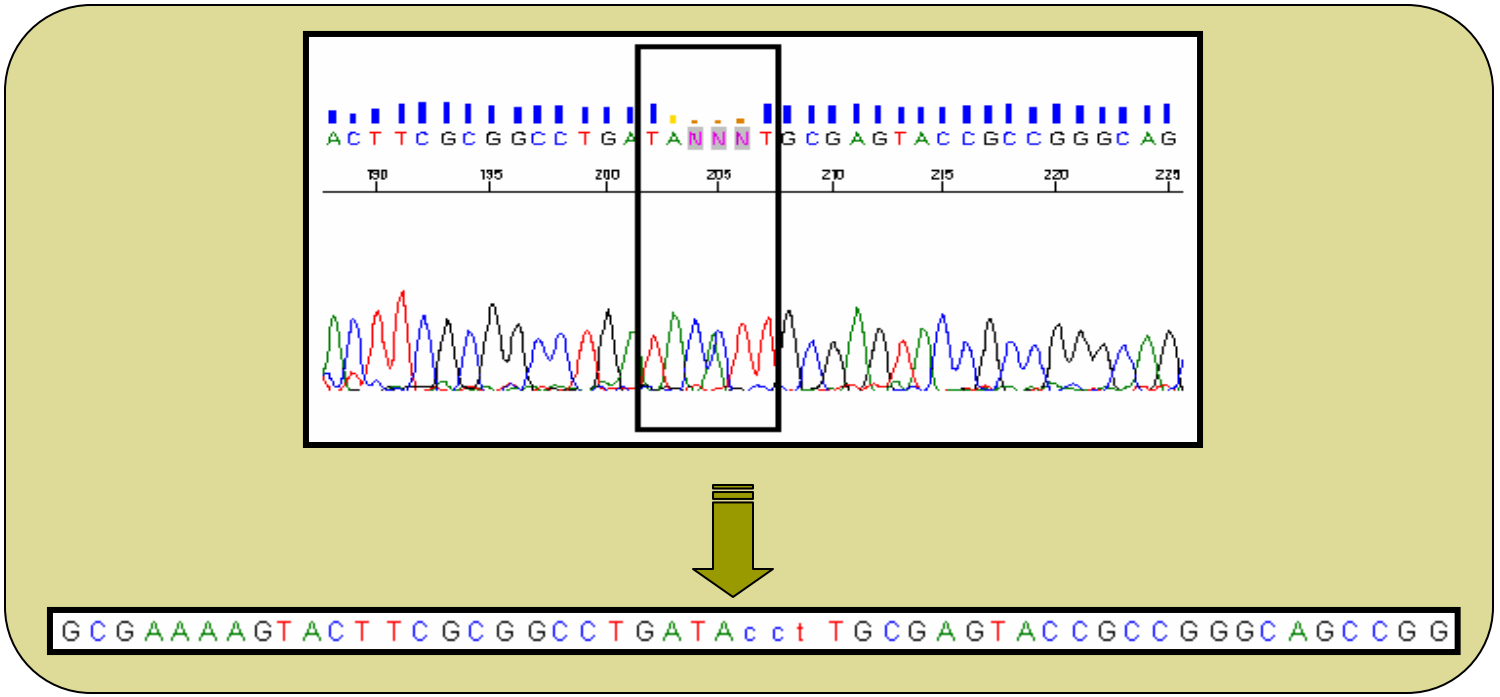
Introduction / Background

- Problem being addressed



Introduction / Background

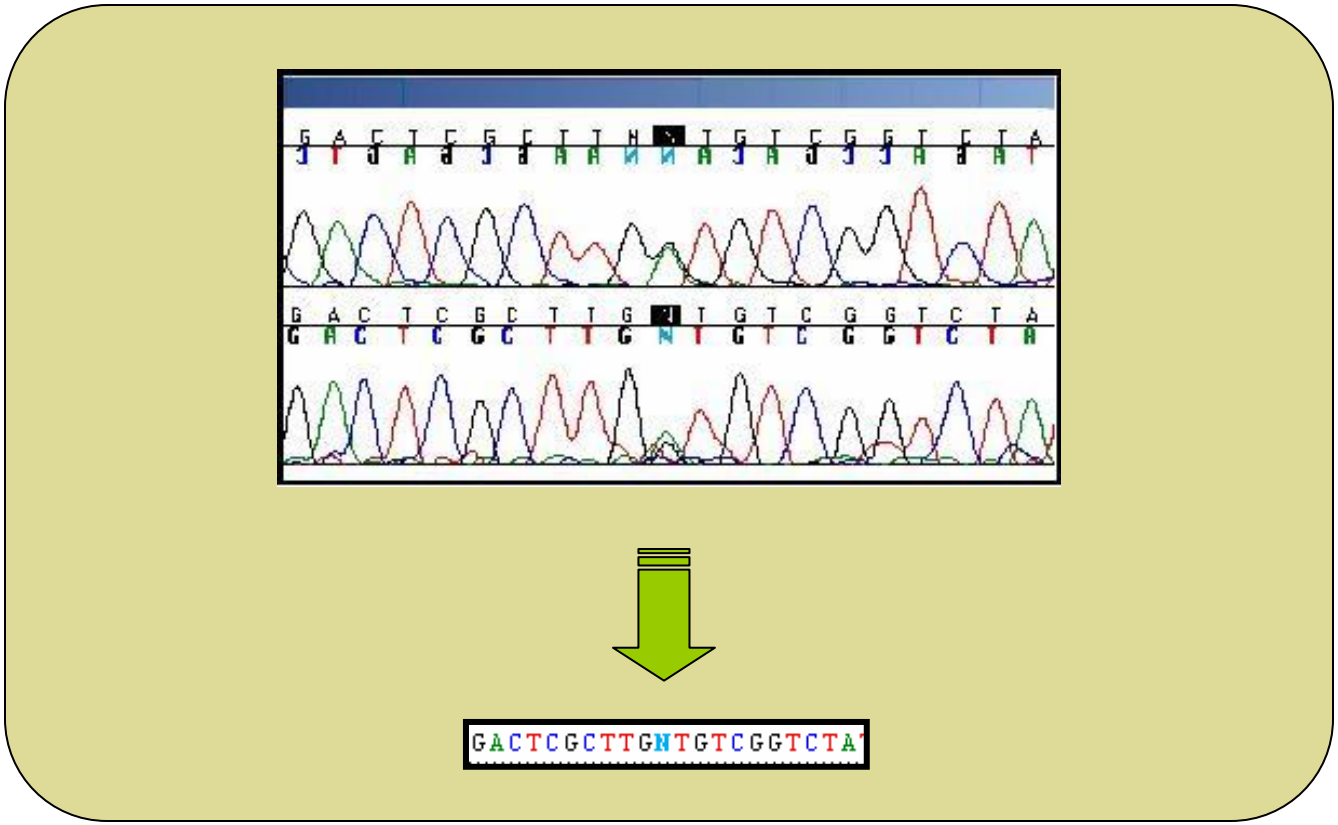
- Problem being addressed



Source: Applied Biosystems Sequence Scanner Software v. 1.0

Introduction / Background

- Problem being addressed



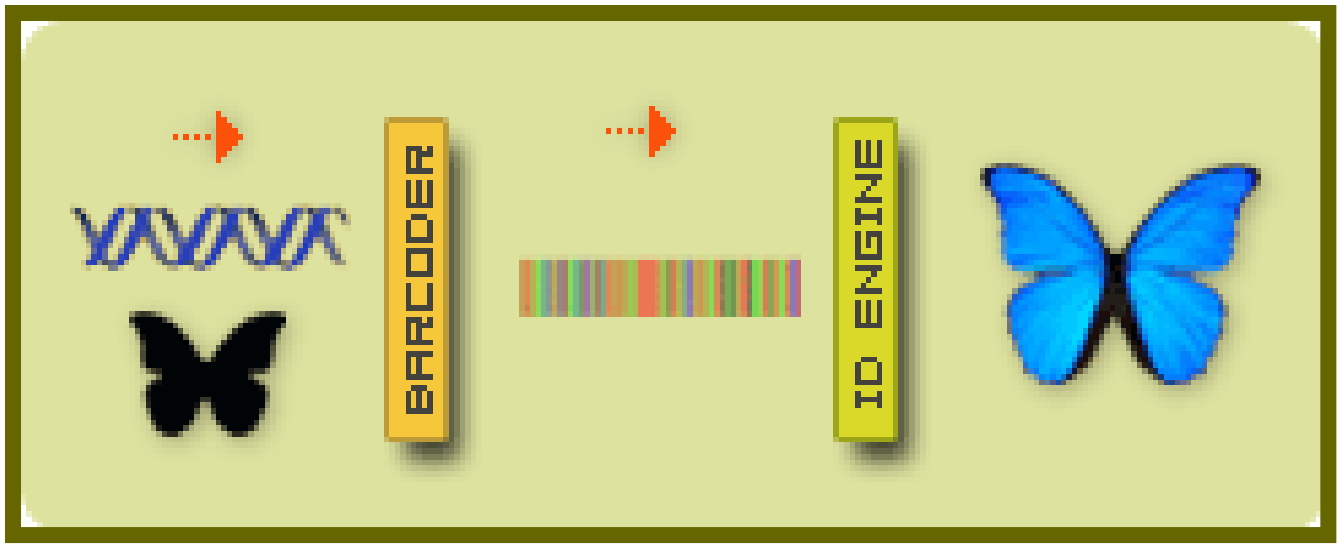
Source: Gene Codes Corporation Sequencher Software v. 4.5

Introduction / Background

- Why is this important?
  - Sequences with low variability are particularly influenced by sequencing errors in analyses of sequence variation due to their low signal-to noise ratio

Introduction / Background

- DNA Barcoding



Source: BOLD http://www.barcodinglife.org/views/login.php

Introduction / Background

- DNA Barcoding
  - *Astraptes fulgerator*

P.N.A.S. (2004) **101**: 14812 – 14817

Model

- DNA sequence data

Bats of Guyana

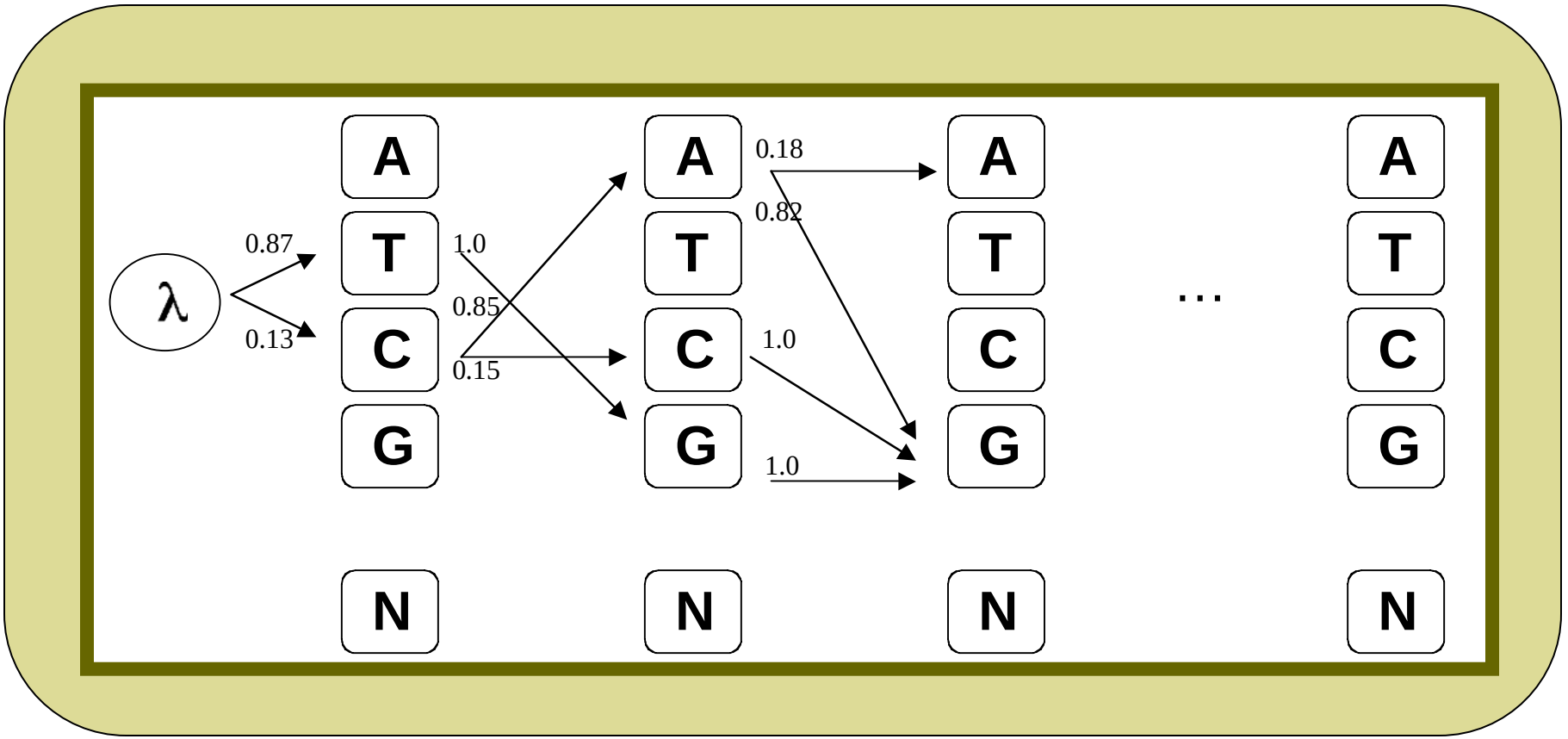
CAGTTGAAGCTGGTGTTGGGACTGGTTGAACCGTATATCCAC
CAGTTGAAGCTGGTGTTGGAACTGGTTGAACCGTATATCCAC
CCGTCGAAGCAGGGGTAGGTACTGGTTGAACCGTGTACCCCC
CCGTCGAAGCAGGGGTAGGTACTGGTTGAACCGTGTACCCCC
TGGTAGAAGCTGGGGCTGGTACCGGATGGACAGTATACCCAC
TGGTAGAAGCTGGGGCTGGTACCGGATGGACAGTATACCCAC
CAGTCGAAGCCGGAGTAGGAACTGGCTGAACTGTTTATCCCC
CAGTCGAAGCCGGAGTAGGAACTGGCTGAACTGTTTATCCCC
CAGTCGAGGCTGGAGTAGGGACTGGCTGAACTGTTTATCCCC
CAGTCGAGGCTGGAGTAGGGACTGGCTGAACTGTTTATCCCC
CAGTCGAAGCAGGAGTAGGTACCGGCTGAACAGTATACCCAC
CAGTCGAAGCAGGAGTAGGTACCGGCTGAACAGTATACCCGC
CAGTTGAGGCTGGAGTAGGTACAGGCTGAACAGTCTACCCTC
CAATTGAAGCAGGCGTTGGCACCGGCTGAACCGTCTACCCCC
CAATTGAAGCAGGCGTTGGCACCGGCTGAACCGTCTACCCCC

...

M. E. Notes (2007) **7**: 184 – 190

Model

- Representing DNA sequence data using a *profile Hidden Markov Model (pHMM)*



Model

- Probability of other sequence given the model

- Ex 1: *Ametrida centurio*

➡ **CAGTTGAAGCTGGTGTTGGGACTGGTTGAACCGTATATCCACC**

➡ **$P(O|\lambda) = 3.461 \times 10^{-9}$**

- Ex 2: *Boophis rappiodes*

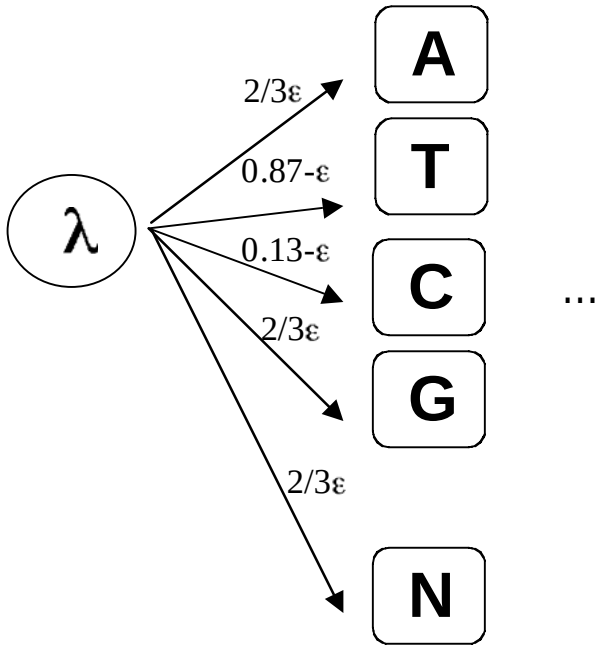
➡ **CCACCTTCAATGACCCAGTACCAAACCCCTCTTTTTGTGTGATC**

➡ **$P(O|\lambda) = 0$**

Model

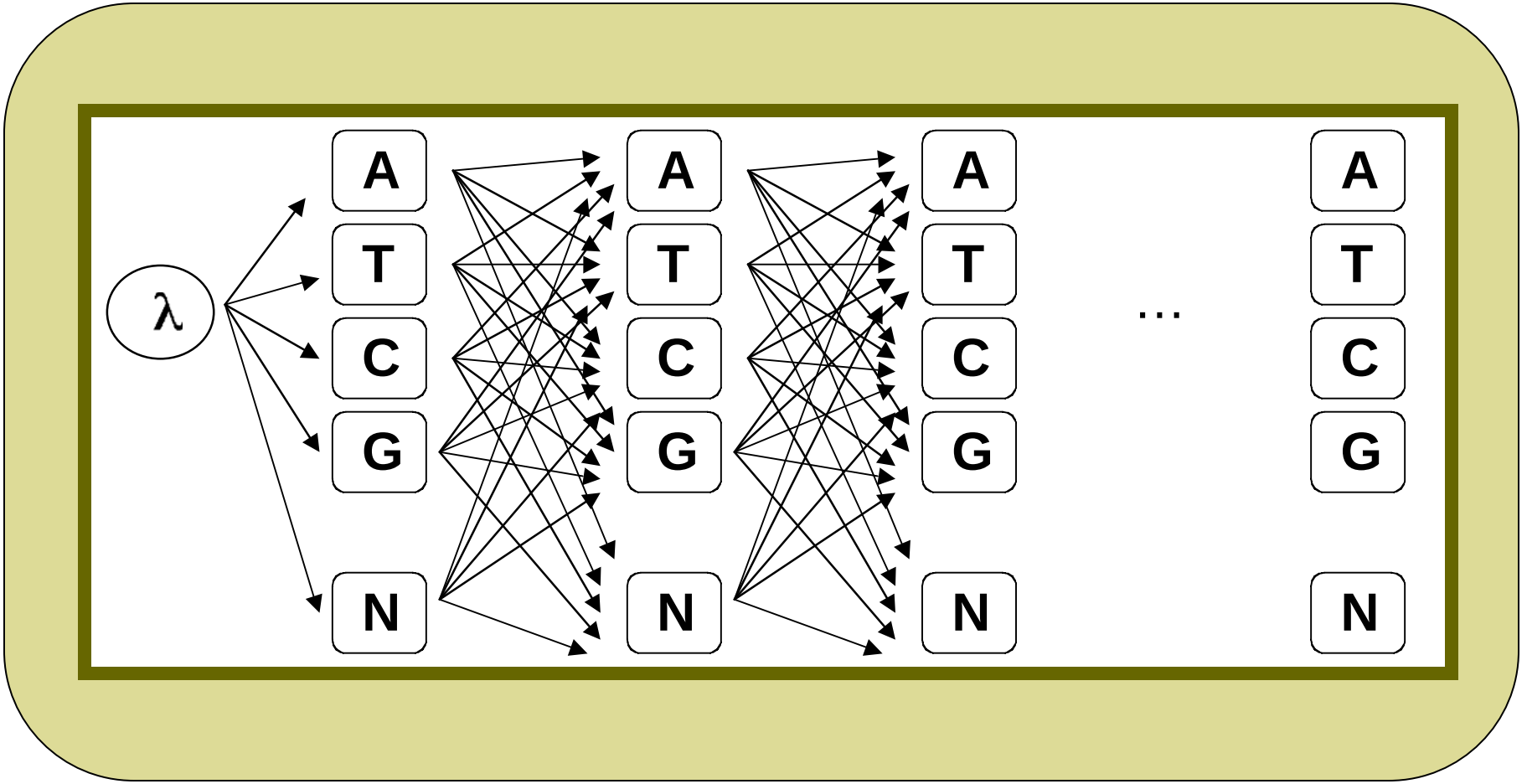
- Modifications to the model

- ε-padding
  - To allow for all possible sequences to have a positive probability according to the model
  - Can be adjusted to increase or decrease the sensitivity of the model to external sequences



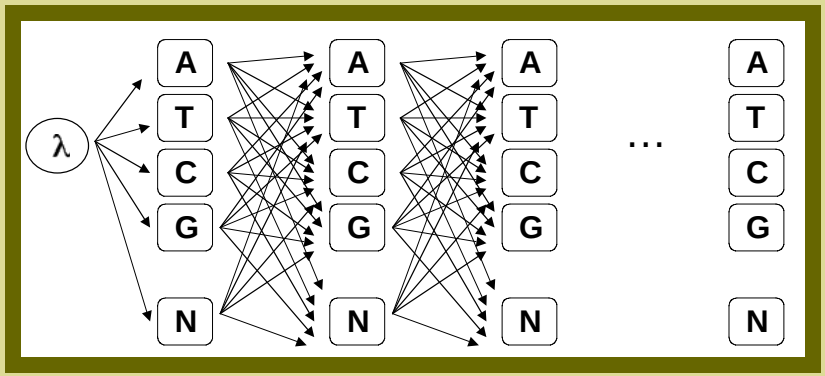
Model

- Profile Hidden Markov Models (pHMM)



Testing the Model

- Test the accuracy of the pHMM
  - Estimate accuracy directly by the pHMM's ability to detect error introduced into a set of DNA sequences

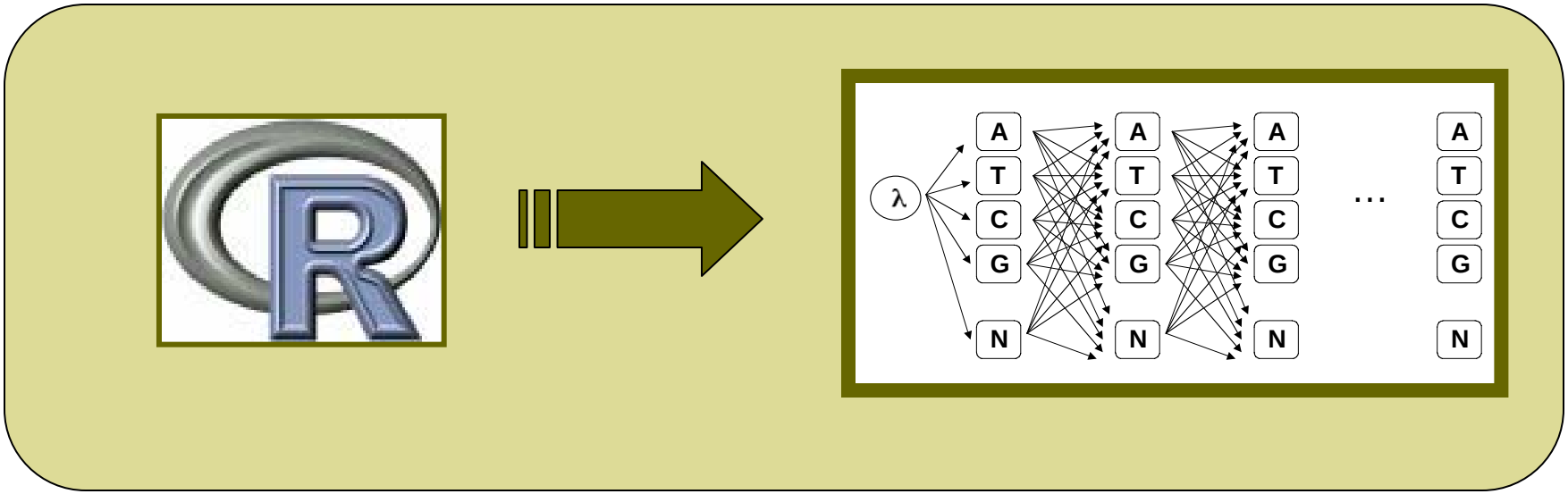


Methods

1. Detection of shallow sequence divergences in a set of sequences
2. Removal of the sequence(s) of interest
3. Construction of 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

Methods

- The R programming environment was selected in which to build the pHMM of sequence data



Methods

- Example

```
$Rho
                [,1]        [,2]        [,3]        [,4]
[1,] 1.762655e-09 0.0001610240 7.992953e-01 8.425386e-01
[2,] 5.732607e-01 0.1424176100 8.142005e-09 3.906765e-13
[3,] 1.432232e-01 0.2814073574 6.207791e-04 7.020717e-06
[4,] 2.835161e-01 0.5760140085 2.000839e-01 1.574544e-01

$tpm
                [,1]        [,2]        [,3]        [,4]
[1,] 9.628128e-01 3.718670e-02 5.260536e-07 2.401726e-10
[2,] 1.831845e-01 8.166392e-01 1.395701e-04 3.666461e-05
[3,] 1.565513e-04 8.277518e-04 7.974409e-01 2.015748e-01
[4,] 2.625085e-06 6.299285e-05 5.919335e-02 9.407410e-01
```

```
$ispd
[1] 0.74884766 0.15199885 0.02250537 0.07664812

$log.like
[1] -13.20095
```

Significance

- Another level of error detection post - sequence editing
  - This level takes sequence history into consideration after the unbiased error estimates of the program PHRED

Other Applications

- Determining the suitability of sequences for analysis of sequence variability

Fishes of Australia

TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTTCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTTCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTTCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA

...

Phil. Trans. R. Soc. B. (2005) **360**: 1847– 1857

Other Applications

- Determining the suitability of sequences for analysis of sequence variability

Fishes of Australia

CCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAG
TGACTTCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTC
TGACTTCTTCCCCCCTCTTTCCTTCTGCTC
TGACTCCTTCCCCCCTCTTTCCTTCTGCTC
TGACTCCTTCCCCCCTCTTTCCTTCTGCTTCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTA
CCTTCCCCCCCTCTTTCCTTCTGCTC
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCT
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAG
ACTCCTTCCTCCCTCTTTCCTTCTGCTC
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA

...

Phil. Trans. R. Soc. B. (2005) **360**: 1847– 1857

Other Applications

- Determining the suitability of sequences for analysis of sequence variability

Fishes of Australia

TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTTCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCT
TCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTTCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
CCTTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTC
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTCTTCA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
CTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
CTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TCTTTCCTTCTGCT
TGACTCCTTCCTCCCTCTTTCCTTCTGCT
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA

...

Phil. Trans. R. Soc. B. (2005) **360**: 1847– 1857

Other Applications

- Search for patterns of informative sites in sequences

Fishes of Australia

TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTTCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTTCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTTCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCTTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCCCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA
TGACTCCTTCCTCCCTCTTTCCTTCTGCTCCTAGCTTCTTCAGGAGTTGA

...

Phil. Trans. R. Soc. B. (2005) **360**: 1847– 1857

Acknowledgements

Laboratory

The Hebert Lab
Including:
Members of the CCDB and BIO

Funding & Support

The Gordon and Betty Moore Foundation, Genome Canada and NSERC of Canada

Committee

Thesis Advisors:
Dr. Paul Hebert
Dr. Dan Ashlock

Committee Members:
Dr. Alex Smith
Dr. Gary Umphrey