

DNA Assembly With Gaps (Dawg)

Simulating Sequence Evolution

Reed A. Cartwright

Department of Genetics
University of Georgia

11/7/2005



1 Simulating Evolution

- Why Simulate Phylogenies?
- DNA Assembly with Gaps
- Estimating Indel Rate



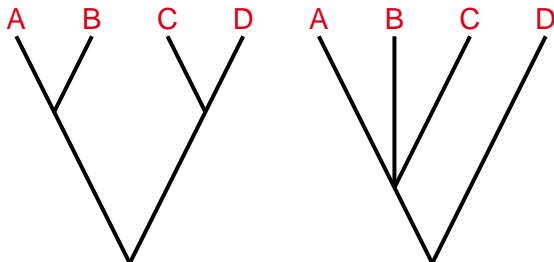
1 Simulating Evolution

- Why Simulate Phylogenies?
- DNA Assembly with Gaps
- Estimating Indel Rate



What is a Phylogeny?

- A phylogeny is an evolutionary tree.
- It shows how individuals and/or groups are related to one another.



Why Simulate Phylogenies?

- Biologists use many techniques to reconstruct phylogenies based on biological data
- However, true phylogenies are unknown, save a few instances.
- How then can we test the accuracy of these reconstruction methods?
- Solution: Use simulations.



Why Simulate Phylogenies?

- Techniques are often based on certain models of evolution.
- Simulating sequence evolution based on these models produces an ideal situation to test the techniques.
- Using other models can test how robust a technique is.



Basic Testing Procedure

- 1 Specify a guide tree for biological, theoretical, and/or practical reasons.
- 2 Simulate sequence sets using this guide tree and an evolutionary model.
- 3 Estimate trees from simulated sequences.
- 4 Compare the estimated trees to the guide tree.



- Tree reconstruction
- Branch length reconstruction
- Model parameters
- Alignments
- Recombination events



Simulating Evolution

- Proper simulation of molecular evolution should include both substitutions and indels.
- However, existing programs either do not include indels or use an unjustified model of indel formation.
- I created a program to address this gap.



1 Simulating Evolution

- Why Simulate Phylogenies?
- DNA Assembly with Gaps
- Estimating Indel Rate



DNA Assembly with Gaps

- This program is called **DNA Assembly with Gaps**.



DNA Assembly with Gaps

- This program is called **DNA Assembly with Gaps**.

Dawg



What is Dawg?

- A portable and robust command-line program for simulating molecular evolution.
- Distinguished by the inclusion of a novel, continuous time indel model along with the GTR+ Γ +I substitution model.
- Development Website: <http://scit.us/projects/dawg/>



Comparing Software

Feature	Seq-Gen	Evolver	Rose	Dawg
Indels			Yes	Yes
Indel Parameter Estimator				Yes
Recombination	Yes			Yes
Substitution	GTR	GTR	PAM	GTR
Rate Heterogeneity	$\Gamma+I$	Γ	$\Gamma+I$	$\Gamma+I$
Input Format	Switch	File	File	File
Unix	Yes	Yes	Yes	Yes
Mac OS X	Yes	Yes	Yes	Yes
Win32	Yes	Yes		Yes



Sample Input File

```
Tree = ((A1:0.001359,A2:0.001359):0.084512,  
        (B1:0.006116,B2:0.006116):0.079756);  
Model = "GTR"  
Params = {1.08031, 2.45581, 0.44452,  
          1.09145, 4.06519, 1.00000}  
Freqs = {0.353470, 0.143681, 0.178206, 0.324643}  
Length = 300  
Lambda = 0.143120  
GapModel = "NB"  
GapParams = {1, 0.753247}  
Format = "Clustal"  
File = "example.aln"  
Seed = 1981
```



Sample Output

CLUSTAL multiple sequence alignment (Created by dawg current-r267)

```
A1      TTCAAAAATATGCTAGGACTGAATATGAATTCTTAAAGTTAAGAAAGATAAAGAAAAACA
A2      TTCAAAAATATGCTAGGACTGAATATGAATTCTTAAAGTTAAGAAAGATAAAGAAAAACA
B1      TTCGAAAATATGTTAGTACTCAATATGAATTCTTTGAGTTAAGAAAGATAAAGCAAA--A
B2      TTCGAAAATATGTTAGTACTCAATATGAATTCTTTGAGTTAAAAAAGATAAAGCAAA--A
```

```
A1      GTACATAATGTAAA----TTATTGCAA-----AAAACGGCTAACAAATTAGACGATT
A2      GTACATAATGTAAA----TTATTGCAA-----AAAACGGCTAACAAATTAGACGATT
B1      ATACATAATGTGATTTCAATATTTCCAATTACCTAACAAATACGGCTATCAATTAAACGATT
B2      ATACATAATGTGATTTCAATATTTCCAATTACCTAACAAATACGGCTATCAATTAAACGATT
```

```
A1      TTAGGATTACGCTGACAAATATTAGGATGATATTAATTTA-----TCTTGTATTTAGAT
A2      TTAGGATTACGCTGACAAATATTAGGATGATATTAATTTA-----TCTTGTATTTAGAT
B1      TTAGGATTACACCGACAAATATTAGGCCGATATGAATTTACCATCATGTTGTATTTAGAT
B2      TTAGGATTACACCGACAAATATTAGGCCGATATGAATTTAACATCATGTTGTATTTAGAT
```

```
A1      GCTGTCTTTTTATCAACATTCATCACTAGATATTGGAACCTATTGCATCTAAGAAGTACAT
A2      GCTGTCTTTTTATCAACATTCATCACTAGATATTGGAACCTATTGCATCTAAGAAGTACAT
B1      GCTGTCTTTTTATTAACATTCATCATTTAAAT-TTGGAACCTTTTGTATTTTAAGAAGTACAT
B2      GCTGTCTTTTTATTAACATTCATCATTTAAAT-TTGGAACCTTTTGCATTTAAGAAGTACAT
```

```
A1      GTTTAAATAGGGTT-AAAACATATATGAAGTCGATTATAAGGAATTTCTATAAAATGTAGC
A2      GTTTAAATAGGGTT-AAAACATATATGAAGTCGATTATAAGGAATTTCTATAAAATGTAGC
B1      GTTTAAATAGTGTTTATAA-TATATATGAAATGATCGTAAGGA---TCTATAAAATGCAGT
B2      GTTTAAATAGTGTTTAAAA-TATATATGAAATGATCATAAGGA---TCTATAAAATGCCGT
```

```
A1      TCTTCAATTTCTTA
A2      TCTTCAATTTCTTA
B1      TCTTCAATTTCTTG
B2      TCTTCAATTTCTTG
```

1 Simulating Evolution

- Why Simulate Phylogenies?
- DNA Assembly with Gaps
- Estimating Indel Rate



Estimating Indel Rate

- Dawg would be of little benefit if biologists could not estimate parameters of indel formation from real data.
- Dawg's indel model allows such estimation, which is implemented in a Perl script, lambda.pl.
- Lambda.pl estimates the rate of indel formation from an alignment and tree as $\hat{\lambda} = n/\bar{L}T$.

n number of unique gaps

\bar{L} average sequence length

T Total branch length of the tree



Confidence Interval of Indel Rate

Example Usage

- I aligned the sequences of chloroplast trnK introns from two Hibiscus and two Prunus species.
- Using Paup*, I estimated the phylogeny and substitution parameters.
- Using lambda.pl, I estimated the indel formation parameters.

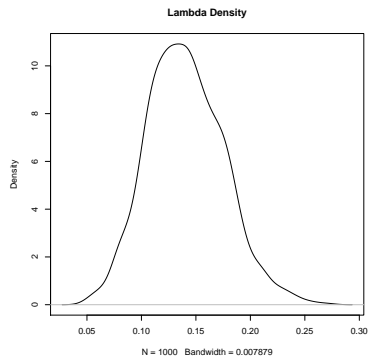


Example Usage

- From these estimated parameters of evolution, I constructed an input file for Dawg.
- From the input file Dawg produced a thousand simulated sequence sets.
- The rate of indel formation was estimated for each of the simulated sequences.



- The estimated rate of indel formation was 0.143120.
- Bootstrapping gave a 95% CI of 0.078530 to 0.213560.
- Biologically this is 8 to 21 indels per 100 substitutions.



Dawg is not (yet) perfect

- Real life is more complex than Dawg's indel model.
- Many indels occur in hotspots and repetitive DNA.
- Continued development on Dawg may eventually take such things into account (hopefully).



Summary

- Sequence simulation is important.
- Dawg is a new sequence simulation program.
- Dawg can be used to bootstrap evolutionary estimates.



Thanks

- Marjorie Asmussen
- Wyatt Anderson
- Paul Schliekelman
- Ron Pulliam
- Jim Hamrick
- John Wares
- Jessica Kissinger
- John Avise
- Anderson Lab
- Asmussen Lab
- Jeff Ross-Ibarra
- Douglas Theobald

