

# Estimating the Rate and Distribution of Indels

Reed A. Cartwright

Postdoctoral Research Associate  
JL Thorne Lab  
Bioinformatics Research Center  
Department of Genetics  
North Carolina State University

June 26, 2007

# Sequence Alignments

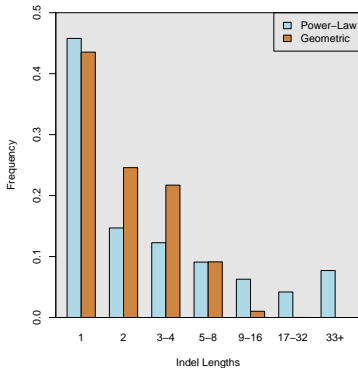
```
Taxa_1  TTTTAGCAAGCATAGCG---TTGACATCCCTCCA GACGAGGGTAGGACCGGATCATGATGATTC---GGAGCCCATGCTAGCA CGAA GATAGT---  
Taxa_2  ATATACGTACCCT-GCGTAGCTCACGTCCT- GAT-----CCATATCG--TTCATTCCA CGAAGCCCATGCTTGCACTCAGATATTTAG
```

- Sequence alignments are traditionally modeled by assuming that indel lengths follow a geometric distribution, i.e. affine gap model (GOTOH 1982).
- This distribution has nice features that make it highly tractable: a single parameter, efficient algorithms, partial biological justification, etc.
- However, everyone agrees that it is not a perfect model for the indel length distribution.
- With the speed of modern computers, can we use a more realistic distribution?

# The Power-Law Distribution

- Several empirical studies have found that indel length distributions obey a power law (GONNET ET AL. 1992, BENNER ET AL. 1993, GU AND LI 1995, ZHANG AND GERSTEIN 2003, CHANG AND BENNER 2004).
- In a power law,  
 $\log f(x) = -z \log x + c$
- This distribution is very different than a geometric.
- The power-law puts more mass on short and long indels.
- However, they both have the same number of parameters.

Difference Between Geometric and Power-Law with Same Medians

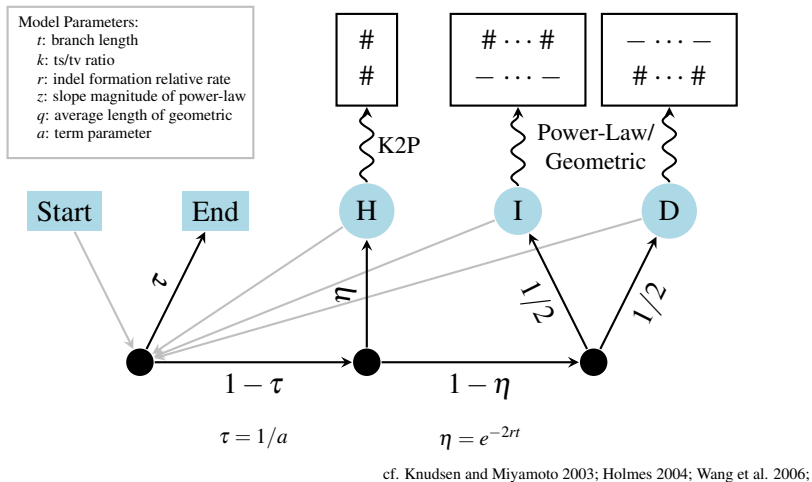


- Power-Law (Zeta) Distribution:  
 $f(k) = k^{-z} / \zeta(z)$
- Geometric Distribution:  
 $f(k) = \frac{1}{q} (1 - \frac{1}{q})^{k-1}$

- The amount and length distribution of indels are estimated from alignments.
- However, alignments are estimated using algorithms with implicit assumptions about rates and distributions of indels.
- Particularly important is the assumption of a geometric distribution of indel lengths.
- To solve this circularity, we need to estimate the indel formation model without assuming any particular alignment.

- If we average over all possible alignments, we do not need to assume that any alignment is true.
- Weighting alignments based on their fit to sequence data allows for alignment ambiguity to improve estimation.
- In order to assign alignments weights, we need a statistical model of alignments.
- This is implemented as a generalized pair hidden Markov model.

# Our Generalized Pair-HMM



# Measuring Indels

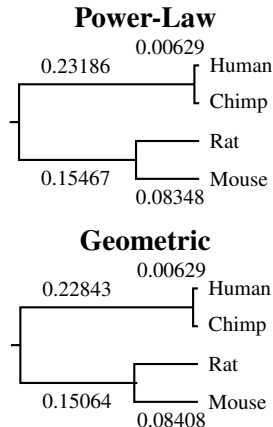
## The Missing Data

- The path through an HMM represents missing data, e.g. the alignment.
- The expectation-maximization algorithm can be used to handle the missing data.
- This allows us to find maximum likelihood estimates for the model parameters.

- Our EM-HMM algorithm is suited for pairwise comparisons.
- Because it can handle multiple sequence pairs from the same individual/species, it is great for genomic comparisons.
- Selection can confound our estimates, so neutrally evolving sequences are needed.
- We don't need aligned sequences, only homologous sequences.
- Therefore, sets of homologous introns (identified by exon homology) from multiple genomes are the best data for this algorithm.

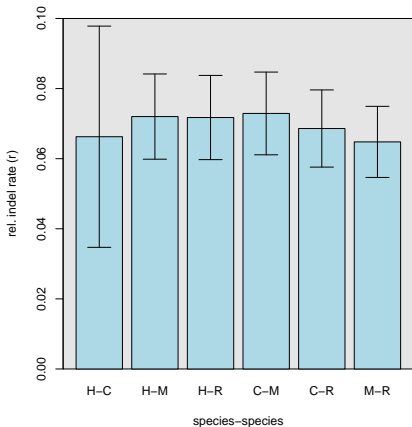
- The dataset consists of introns from several serpin (serine protease inhibitor) genes in four genomes.
  - Human, Chimp, House Mouse, and Brown Rat.
  - Gathered via the HomoloGene database at NCBI
- Due to computational concerns, introns greater than 2500 residues were not included in the analysis.
- The final dataset has 11 sets of introns.
- Each genome is represented by approximately 12 thousand residues.

- The EM-HMM can give us an MLE for the evolutionary distance,  $t$ , between each pair of species under both models.
- These distances can be used to construct phylogenies.
- These trees were produced via rooted least squares by program KITSCH in the Phylip package.
- They are consistent and reasonable given what we know about these species.

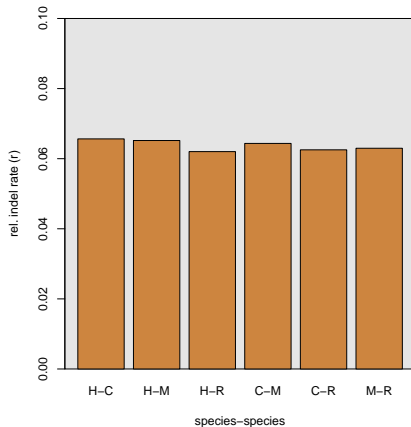


# Relative Rates of Indel Formation

## Power-Law



## Geometric

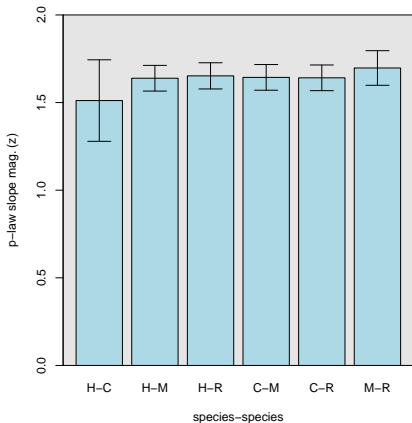


- There is a consistent estimate of 6–8 insertions and 6–8 deletions per 100 substitutions.

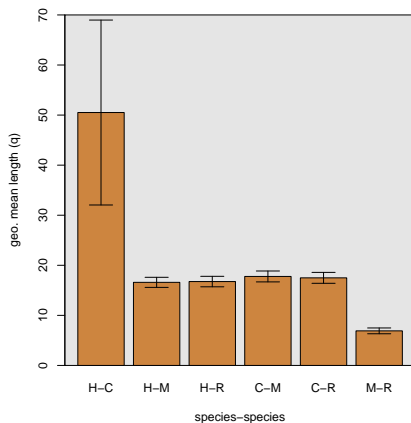
# Distribution Parameters

- Estimated parameters of indel-length distribution depend on species pair for geometric but not for power-law.

## Power-Law



## Geometric



## EM-HMM Log Likelihoods

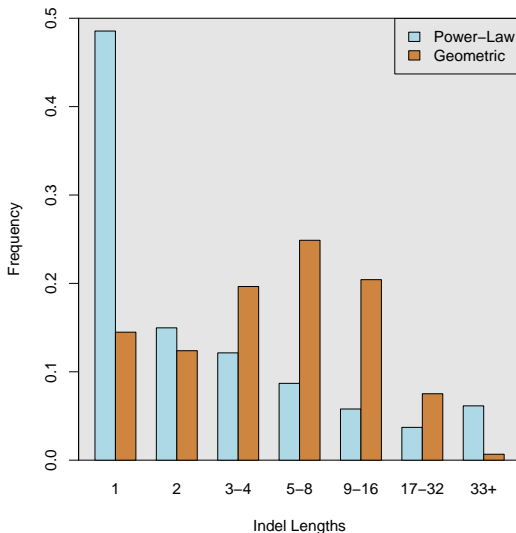
	Power-Law	Geometric	Difference
Human–Chimp	-19456	-19490	35
Human–Mouse	-32381	-32513	132
Human–Rat	-31601	-31736	135
Chimp–Mouse	-32958	-33110	151
Chimp–Rat	-32176	-32319	143
Mouse–Rat	-24543	-24621	79

- How different are the estimated power-law and geometric distributions?

# The difference between model distributions

## Mouse-Rat Comparison

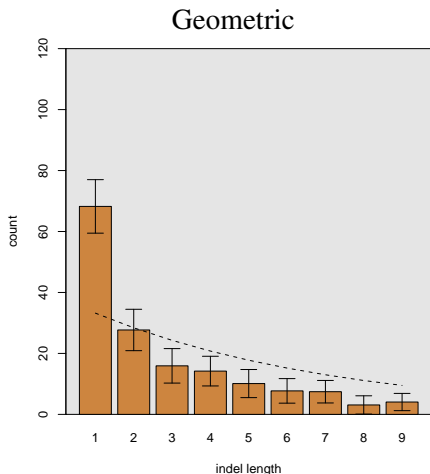
### Estimated Indel Distributions



# Indel Length Distributions

## Mouse–Rat Comparison

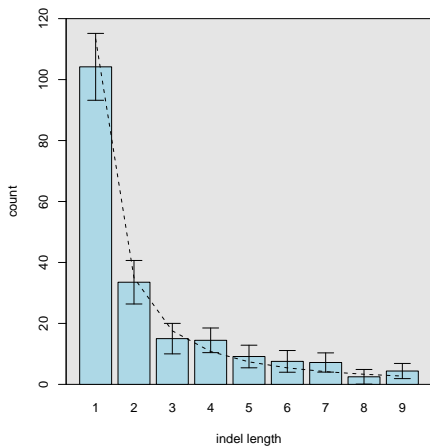
- For each length, we estimated the expected number of indels of that length given the data and given the maximum likelihood parameter estimates.



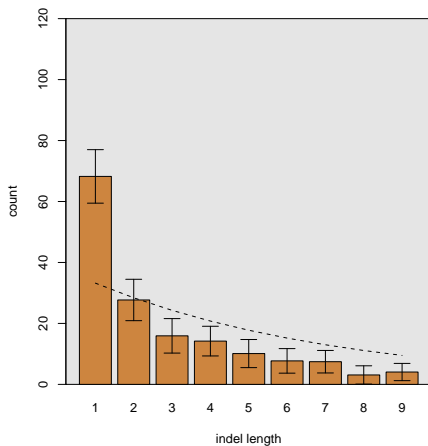
# Indel Length Distributions

## Mouse-Rat Comparison

### Power-Law



### Geometric



- It is possible to estimate neutral models of indel formation in a non-circular manner with an EM-HMM.
- The power-law distribution fits the sequences better than the traditional geometric distribution, yet still has a single parameter.
- Importantly, the power-law parameters are conserved between rodents and primates.
- Are gap length distributions in neutrally evolving portions of genomes conserved throughout large parts of the tree-of-life?
- Thanks—Jeff Thorne, Asger Hobolth, Ben Redelings, Jeff Ross-Ibarra, Sang-Chul Choi
- Support—NIH Grant GM070806