Evolutionary Inference via the Poisson Indel Process

Bouchard-Coté, Jordan

PNAS, 2013

Review by Michael Nute

2/18/16

Summary

- Paper proposes a model of sequence evolution where insertions/deletions occur as part of a Poisson Process:
  - Independent of sequence length
  - Indel lengths would be roughly Poisson distributed.
  - Otherwise borrows from TKF91 [1] model
- Benefits:
  - Uses a statistical model if insertions/deletions (like TKF91) but also allows likelihood to be calculated in linear time.
- Criticism:
  - Entire model is virtually identical to BAli-Phy [2] which also uses the TKF91 model and has been implemented for 8 years (as of 2013 publication).
  - No simulation to show benefit of this over BAli-Phy.
  - No effort by authors to seriously implement the method.
    - No maximum likelihood computation, even though we can now purportedly calculate likelihood.
    - Cursory implementation with off-the-shelf Metropolis-Hastings algorithm was used.
  - Evaluation studies are essentially non-existent.

Model

TKF91 Model:

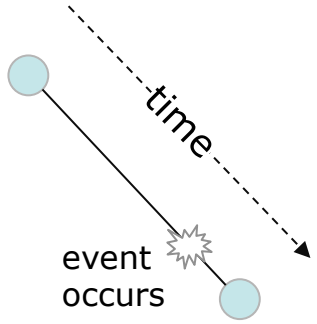
- "Events" occur in time according to a Poisson Process.
  - I.e. in any interval of time t:
    1. The expected number of events is proportional to t, and
    2. The number of events is independent for disjoint intervals
- Theorem: multiple Poisson processes acting at once is equivalent to a single Poisson process with events of different "type".
  - In this case, "type" is a) Insertion, b) Deletion, and c) Mutation.
  - **Insertion:** it is a single character in length and the value is drawn from the stable distribution
  - **Deletion:** single character in length
  - **Mutation:** value drawn from substitution matrix
  - Rate parameters proportional to sequence length

Bouchard-Coté, Jordan:

- Same as TKF91 except Insertions have a single parameter that does not depend on length and when they occur their length is Poisson distributed.

Comments:

- This is essentially the GTR model but with a new process occurring in the background governing insertions and deletions.



Substitution Matrix:

$A=e\uparrow Qt$

where

$Q=[\blacksquare q\downarrow aa \&q\downarrow ac \&q\downarrow ag \&q\downarrow at @q\downarrow ca \&\cdots$

Evaluation

Evaluation Protocol (and comments):

- Used 100 simulated trees with 7 taxa each.
  - *7 is way too small to be useful in 2013 and does not allow parameter space to vary at all.*
- Compared four methods:
  1. Baseline: Clustal-Omega [3] alignment and Phyml Tree (i.e. not using PIP)
  2. PIP for tree: keeping MSA fixed, varying trees with PIP
  3. PIP for MSA: keeping tree fixed, varying MSA with PIP
  4. PIP for Both
  - *This won't do much to actually evaluate the method. If the likelihood function is helpful at all it will show some improvement.*
  - *Clustal-Omega is a progressive alignment method, far from the best and especially on such small data.*
  - *No effort to compare to other methods (e.g. BAli-Phy) that do* exactly *the same thing.*

Figures from the Paper:

Table 1. PIP results on simulated data

	Reconstruction accuracy			
Tree resampled?	No	Yes	No	Yes
MSA resampled?	No	No	Yes	Yes
Edge recall (SP)	**0.25**	—	0.22	0.24
Edge Precision	0.22	—	0.56	**0.58**
Edge F1	0.23	—	0.31	**0.32**
Partition Metric	0.24	0.22	—	**0.19**
Robinson-Foulds	0.45	0.38	—	**0.33**

Reconstruction accuracy using five different metrics. The bold font highlights the best-performing combination of resampling for each row.

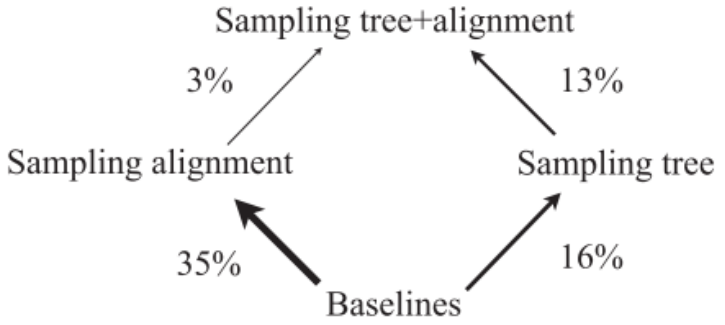


Fig. 4. Relative improvements for enabling each component of the sampler. Arrows on the left are relative alignment improvements, and arrows on the right are relative tree improvements.

References

[1] Thorne, J. L., Kishino, H., & Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, *33*(2), 114–124. http://doi.org/10.1007/BF02193625

[2] Suchard, M. A., & Redelings, B. D. (2006). BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, *22*(16), 2047–2048. http://doi.org/10.1093/bioinformatics/btl175

[3] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539. http://doi.org/10.1038/msb.2011.75