Notes On Papers For Thesis

Siddharth Reed, 400034828

October 17, 2018

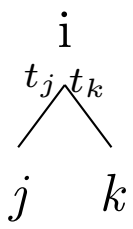
Estimation Of Gene InDel Rates With Missing DataDang et al., 2016

Intro

- Can infer indel rates from P/A of genes in closely related species
- Parsimony underestimates indel rates
- Can compare different gene trees as well
- Need whole sequence to be sure rearrangements dont mask homologs
- ML methods can now account for missing data
  - Missing can mean non-whole genomes
  - Can also be genome reduction beyond normal flux (pathogen deletion)
- Evo. rates can vary across lineages, allows for this to some extent

Methods

- Use P/A markov chain
- Assumes indels are independant, at constant rate
- Consider families to avoid paralog issues (cluster above > BLAST % Ident)
- Rate matrix $Q = \begin{bmatrix} -\mu & \mu \\ v & -v \end{bmatrix}$ with insertion,deletion $= v, \mu$ respectively
- $P(P_d^i | P_a^i, t) = (\mu + v)^{-1} \cdot (v + \mu e^{-(\mu+v)t})$
- using liklihood to accomodate missing data
  - $L^i(g_i)$ =1 if g_i observed at node i, 0 otherwise
  - $L^i(A) = (\delta, 1)' \implies$ prob of gene missing even if truly present is $(\delta, 1)$
  - $L^i(P) = (1 - \delta, o)' \implies$ prob of gene present even if truly absent is $(1 - \delta, 0)$
  - $\delta = (\delta_1 \dots \delta_s)$ is the proportion of missing data for all tree members $1 \to s$ where $\delta \in [0, 1]$



For a given gene and the above tree

$$L_i(g_i) = [\sum_{g_j} p_{g_i g_{[}j}(t_j) L_j(g_j)] \times [\sum_{g_k} p_{g_i g_k}(t_k) L_k(g_k)]$$

		Observed	
		0	1
True	0	1	0
	1	δ_i	$1-\delta_i$

- Pr(observed P/A of g_i) $= f(x_h) = \sum_{x_0} \pi_{x_0} L_0(x_0)$
- Log-lkl of P/A pattern for a set of genes (Θ) is $l(\Theta) = \sum_{h=1}^{N} \pi_{x_0} L_0(x_0)$
- correction for genes never observed (lost over time) $L_+^h = \frac{l^h}{1-L_-^h}$ (L_-^h is Pr(gene h absent in all taxa, computed bu calculating liklihood of all 0 vector on the tree)
- Assumed all genes are equally likely to be missing
- Four models used
  - $\mu = v$
  - $\mu = v, \delta > 0$
  - $\mu \neq v$ or $\mu \neq v$
  - $\mu \neq v or \mu \neq v, \delta > 0$
- pick model params with DT or BIC

Results

- BIC is able to recover μ and v params effectively within 100 runs on a set of 5000 genes in 5 taxa in homogenous indel rates
- Deletion rates are artificially inflated if missing data not accounted for
- simulation procedure:
  - generate tree with taxa
  - branch lengths estimated fro beta dist.(1:4,8), with scaling factor
  - 500 rnd samples of 5000 phyletic patterns
  - patterns simulated with $\mu \in [0.625, 1.167]$ and $v \in [0.875, 2]$,
- assumed at least 3 taxa have (no δ)
- for *Troy* OTUs, estimated up to ~ 3 indel event per base sub. by Model 1
- estimating missing data has large effect on param estimates
- can cluster branches by length
- more stuff

Discussion

- Dont overparameterize (no δ for each tip)
- these methods are best for trees of closely related taxa with short branch lengths
- extend with GMMs, Γ rate var for gene fams

Inferring Horizontal Gene Transfer Ravenhall et al., 2015

Intro

- Transformation, conjugation, transduction as HGT methods
- Methods of detection
  - Nuc. composition analysis
    - GC content
    - Codon Usage
  - only need 1 genome
  - vanishes over time via mutation
  - need to account for intragenmic variation
  - Phylogenetic methods
    - Gene vs Species distance (low,high $\implies$ HGT)
    - Look for ILS between gene/species tree (close genes,farther species)
  - need a few genomes
  - genearlly better than parametric methods
  - can infer donor and time of transfer
  - generally only applied to coding sequences
  - issues with duplication $\rightarrow$ gene loss vs. HGT
- Combining methods can improve results, but also increase FP rate

Parametric

- Need HGT candidates to be sig. diff from host signature and insig. diff from donor signature
- cant detect ancient transfers as well, mutation
- signatures include
  - nuc. composition
  - kmer frequency
  - codon usage bias
  - structural features
  - genomic islands

Phylogenetic

- Nodes that look like ILS can indicate HGT
- explicit methods
  - need strong BS support on these nodes
  - paralogy can lead to false HGT detection
  - test if gene/species trees are sig. diff. (KH test), if no resonable explanation infer HGT
  - spectral approach, compare discordance of gene/species subtrees
  - create these bipartitions only at strong BS branches, can also use quartet decomposition

  - If pruning/regrafting can resolve gene/species tree diff., that edit path can imply donor and recipient of HGT
  - can also use reconciliation methods to map possible HGT events
- Genes that are similar in highly diff. species
- Implicit methods
  - bunch of distance metrics for stuff
  - also assumptions and issues

Fate of Laterally Transferred Genes Hao and Golding, 2006

Intro

- indels vary between species
- inferring gene indels hard
- Parsimony underestimates evo. events
- assumed insertion = deletion rate for each branch, but branches can vary
- suggests many LGT genes are quickly deleted after transfer
- LGT rates increase at tree tips
- LGT rates generaly $\geq$ than nuc. sub rates

Results

- α, β were selected using ML, assuming $\alpha = \beta$ and not
- α much lower than β
- Based on Bacillus dat, could seee up to 5 gene indels ber nuc. sub
- more gene movements in closer vs more distant organisms (i.e. at tips)
- Assuming genes cannot be regained has no sig. effect on param estimates
- External branches have higher InDel rates than internal $\rightarrow$ higher indels at tips
- Strain specific genes evolve faster than ancestral genes in other taxa
- K_s and K_ns both elevated in strain specific genes (i.e. more recently transfered genes)

Discussion

- Want complete genomes to eliminate hidden paralogs or homolog masking
- Removed all predicted ORFs with no BLAST homologs
- High indel rates not explained by highly mobile genetic elements (patho islands), most such ORFs were excluded

Methods

- –

References

Dang, Utkarsh J. et al. (2016). "Estimation of Gene Insertion/Deletion Rates with Missing Data". In: *Genetics* 204.2, pp. 513–529. ISSN: 0016-6731. DOI: 10.1534/genetics.116.191973. eprint: http://www.genetics.org/content/204/2/513.full.pdf. URL: http://www.genetics.org/content/204/2/513.

Hao, W. and G. B. Golding (2006). "The Fate Of Laterally Transferred Genes: Life And Death In The Fast Lane". In: *Genome Research* 16.5, pp. 636–643. DOI: 10.1101/gr.4746406. URL: Genetic://genome.cshlp.org/content/16/5/636.long.

Ravenhall, Matt et al. (2015). "Inferring Horizontal Gene Transfer". In: *PLOS Computational Biology* 11.5, pp. 1–16. DOI: 10.1371/journal.pcbi.1004095. URL: https://doi.org/10.1371/journal.pcbi.1004095.