Extracting evolutionary signal from patchy genomic data and gene content

Part I: Partial taxon coverage.
Joint work with:

Mike Steel

Michelle McMahon

Michael Sanderson

SMBE, Lyon, June 6, 2010

Part I: Partial taxon coverage


Taxon coverage matrix (by genus) for cluster set at subtree whose root is 'Cactaceae'

Group	Taxa	Loci	% Missing	Citation
Metazoa	77	150	55	Dunn et al. 2008
Papilionoid legumes	2228	39	96	McMahon and Sanderson 2006
Asterales	4954	5	91	Smith et al. 2009
Eukaryotes	73060	13	92	Goloboff et al. 2009

Taxon coverage pattern

	Gene1	Gene2
a	x	x
b	x	x
c	x	x
d	x	
e		x

Two taxon sets: {*a*,*b*,*c*,*d*} and {*a*,*b*,*c*,*e*}

$\mathbf{S}$ = {{*a*,*b*,*c*,*d*},{*a*,*b*,*c*,*e*}}

Note: this does not commit to any *particular* gene tree topologies, just the taxon sets.

Decisiveness

Definition: Let $\mathbf{S}$ be a collection of taxon sets. Then:

- $\mathbf{S}$ is *decisive for a phylogenetic tree T* provided *T* is uniquely determined by how it resolves each of the taxon sets in S.

- $\mathbf{S}$ is *phylogenetically decisive for all trees* if this holds for all choices of *T*.

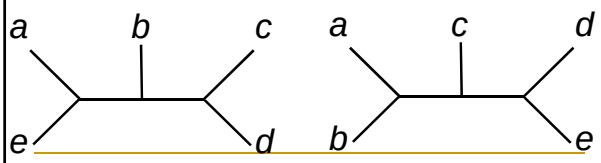
Not phylogenetically decisive (for all trees)

Of the 15 possible binary unrooted trees for this data set...

a	x	x
b	x	x
c	x	x
d	x	
e		x

There are 6, like that below, where the taxon coverage is decisive

...and 9, like that below, where it is not decisive



Phylogenetically decisive (for all trees)

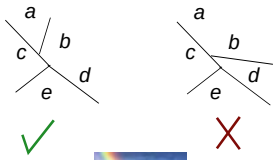
a	x	x	x	x
b	x	x	x	
c	x	x		x
d	x		x	x
e		x	x	x

Testing whether taxon coverage is phylogenetically decisive (for all trees)

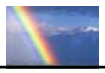
Theorem [S+Sanderson, 2010]

- ***S*** is phylogenetically decisive for all trees **if and only if**:
  For each 4-way partition of the full taxon set, there are four taxa in one of the taxon sets in *S* that intersect each block of the partition.

a	x	x
b	x	x
c	x	x
d	x	
e		x



Complexity? Manuel Bidirsky's `No rainbow colouring' problem



7

A lower bound on number of loci for decisiveness for all trees

Theorem:
If a collection ***S*** of taxon sets of size $n_1, n_2, \ldots, n_k$ is phylogenetically decisive for all trees with n leaves then:

$$\sum_{j=1}^{k} n_j(n_j-1)(n_j-2) \geq n(n-1)(n-2)$$

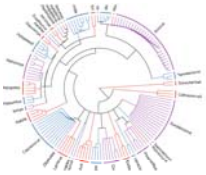
So, if each taxon set in ***S*** has size at most *m* then: $k \geq \frac{n(n-1)(n-2)}{m(m-1)(m-2)} \geq (n/m)^3$

Examples:

max col density (m/n)	"min" num loci (k)
33%	27
80%	2

8

Modelling random taxon coverage



UC Uniform coverage:

$p(x) = p$ for all taxa x in X

(also Variable coverage)

- Next gen whole genome sequencing at "low" coverage that generates partial assemblies
- Even complete genome sequences at deep phylogenetic depths where loss of homology is an issue in assembling data sets

Theorem

★ *For any rooted binary tree T with n leaves, with coverage probability p, the probability that a set* ***S*** *of k (random) taxon sets is phylogenetically decisive for T is at least* 1 - ε *if*

$$k \geq \frac{\log((n-2)/\varepsilon)}{-\log(1-p^3)}$$

★ *Moreover, if k is much less than this, then* ***S*** *is not (w.p. 1- ε) phylogenetically decisive*

Decisiveness in real data (random sampling)

Taxon	Source	Taxa	Loci	Coverage Density	Min loci	Max loci
Cactaceae	PhyLoTA [a]	488	18	0.14	2932	23891
Papilionoid legumes	McMahon (2006) [b]	1794	72	0.16	2306	16230
Metazoa	Hejnol et al., 2009	94	1487	0.18	1095	7148
Rice	Cranston et al. (2010)	10	9481	0.48	35	91

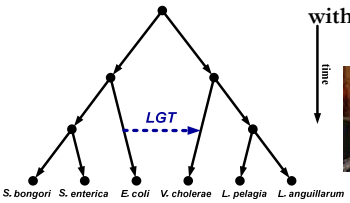
[b] Their "sparse analysis"
[a] PhyLoTA database

n *p*

11

Part II

- Quantifying LGT
  - Avoiding a `Genome of Eden'

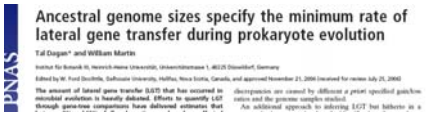


Joint work with:


Leo van Iersel

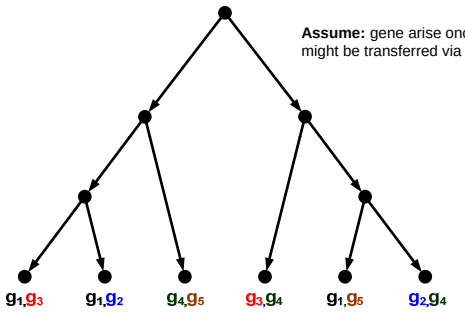

Charles Semple

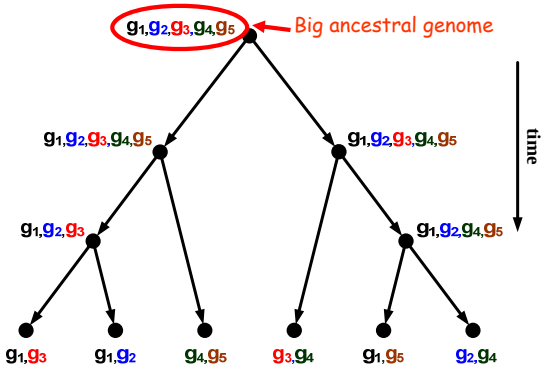
Quantifying LGT


Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution

- Tal Dagan and Bill Martin, 2007
  - Used presence-absence patterns of genes on tree
  - Heuristic approach to estimate the number of LGTs necessary to avoid large ancestral genomes (> 5x larger than present day with no LGT)

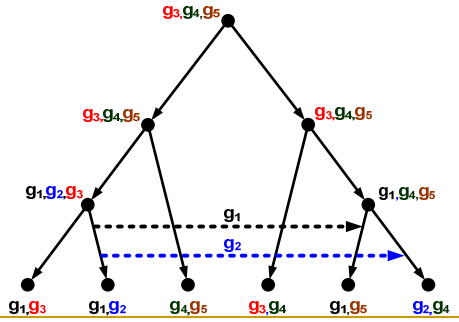
Presence-absence pattern of genes





"Genome of Eden"
(Doolittle et al. 2003)

Explain the pattern, using LGTs to reduce the largest ancestral genome size



Without LGT, is there a tree that avoids a 'genome of eden'?

More precisely:

Given genomes is there a **tree** that can explain the pattern of presence/absence of genes **without any LGT** and insisting that no ancestral sequences are larger than *k*?

Theorem

This problem is already (NP) hard!

Even when |G(x)|=2 for all x.

What if the tree+LGT arcs (network) is given?

- Given genome assignment to the leaves of a LGT network can we determine if ancestral genomes need to be larger than *k*?

- **Theorem**

This problem is also (NP) hard!

Is there any good news?

1. Bounds:

For each species a collection of genes, max genome size *k*, and a tree *T*.

Find: Upper and lower bounds on the minimal number l(*T*,*G*,*k*) of LGT transfer events* needed.

Theorem

$$\ell(T,G,k) \geq \sqrt{\tfrac{2}{3}|\{v \in V : n(v) > k\}|}.$$

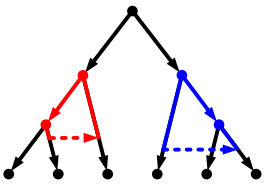
$$\ell(T,G,k) \leq \lceil \tfrac{|G|-k}{k} \rceil \cdot (|\mathcal{X}|+1).$$

$n(v)$:= number of genes g for which v is a mrca of two leaf taxa having g in their genomes

*Each arc can transfer several genes!

2. An algorithm:

There is an efficient method for determining whether ancestral genomes of size >k can be avoided with at most m LGTs that form non-overlapping cycles



Further details

- Van Iersel, L., Semple, C. and Steel, M. (2010). Quantifying the extent of lateral gene transfer required to avert a `Genome of Eden'. *Bulletin of Mathematical Biology* (in press).

FOR EARLIER PART OF TALK

- Sanderson, M.J., McMahon, M.M. and Steel, M. (2010). Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evolutionary Biology* 10: 155
- Steel, M. and Sanderson, M.J. (2010). Characterizing phylogenetically decisive taxon coverage. *Applied Mathematics Letters* 23, 82-86.

Leigh Sawmill 2011

The Annual New Zealand Phylogenetics Meeting

Sunday 6th - Friday 11th February, 2011

http://www.math.canterbury.ac.nz/bio/events/leigh2011/