genealogical sorting

bsci471/biol671

2013
expected learning outcomes

you should be able to

• describe the basic divergence process in terms of lineage sorting
• define monophyly, paraphyly and polyphyly
• describe the genealogical sorting index: what it measures and how it is calculated
• recall examples of how the genealogical sorting index is used
the species divergence process

• irrespective of speciation mechanism or species concept, divergence of species should be reflected in gene genealogies
• initially, when two populations begin to diverge from a single population the gene copies within both descendent populations will share many ancestors
• over time genetic drift will lead to sorting of the gene lineages in the absence of genetic exchange between these populations
• as some gene lineages proliferate and others go extinct, patterns of exclusive ancestry within each population evolve
Hennigian view of lineage divergence

Species 2 (or 1)

Species 3

Ancestral Species/
Most recent common ancestor

Phylogenetic relationships

Tokogenic relationships

Tokogenic relationships
de Queiroz view of lineage divergence
from a gene genealogical perspective
gene tree–species tree mismatches

• result chiefly from shared ancestral polymorphism or introgressive hybridization
• are most likely in groups that are actively diverging or that have diverged recently
• are commonly observed in studies of closely related species
• present problems for inferring species relationships from gene genealogies
relationships from a phylogenetic perspective

monophyly

polyphyly

monophyly/paraphyly

paraphyly/polyphyly
probability of relationships

redrawn from Tajima (1983)
probability of reciprocal monophyly

- mt or cp
- 1 locus
- 5 loci
- 15 loci

redrawn from Hudson and Coyne (2002)
standard categorical phylogenetic concepts

- polyphyly, paraphyly, and monophyly qualitatively describe relationships
- monophyly represents an endpoint in the divergence process
- monophyly as a binary condition (i.e., a group is either monophyletic or not monophyletic) is ill-suited for quantifying divergence
- categorical concepts fail to quantify divergence along a continuum
genealogical sorting index
the genealogical sorting index

- the genealogical sorting index (gsi), is a statistic that quantifies exclusive ancestry of individuals in labeled groups on a rooted tree
- the gsi is a normalized measure of relative exclusive ancestry
some differences ...

- phylogenetics
  - topological relationships are emphasized
  - absolute categorical relationships (e.g., monophyletic, not monophyletic)
- coalescent theory
  - topology is a nuisance parameter
  - coalescent times are emphasized and continuously distributed
- the genealogical sorting index
  - topology is used to quantify coalescent events uniting a group
  - exclusivity of groups is continuously distributed
the genealogical sorting index

• genealogical sorting (gs)
  • the minimum number of nodes (coalescent events) on a fully resolved tree required to unite a group, divided by the number of nodes (estimated coalescent events) actually uniting the group
  • thus, the numerator represents the fully exclusive case (i.e., monophyly), and the denominator represents the observed amount of exclusivity
  • the quotient is a measure of relative exclusivity

• the genealogical sorting index (gsi)
  • gs is normalized to quantify the amount of genealogical sorting along the unit interval, [0, 1], which facilitates comparisons among groups and among trees
genealogical sorting (gs)

\[ gs = \frac{n}{\sum_{u=1}^{U} (d_u - 2)} \]

where \( n \) is the minimum number of nodes (coalescent events) required to unite a group of size \( n + 1 \) through a most recent common ancestor.

d is the degree of node \( u \) of \( U \) total nodes uniting a group (estimated coalescent events) through a most recent common ancestor.

the degree is the number of branch (edge) ends at a node (vertex).
genealogical sorting index (gsi)

gsi = \frac{\text{observed gs} - \min(gs)}{\max(gs) - \min(gs)}

the normalized statistic quantifies genealogical sorting along the unit interval, [0, 1]
maximum possible gs value

\[ \max(gs) = 1 \]

max(gs) is reached when a group is monophyletic
minimum possible gs value

\[ \min(gs) = \frac{n}{\sum_{i=1}^{l} (d_i - 2)} \]

where \( i \) is one of \( l \) total nodes on the tree

thus the minimum value results if all nodes on a tree are required to unite a group (i.e., \( U = l \))
hypotheses

• $H_0$: the amount of exclusive ancestry observed is that which might be observed at random
  • in other words, labeled groups form a single group of mixed genealogical ancestry
• $H_A$: the amount of exclusive ancestry observed is greater than that which might be observed at random
statistical significance via permutation

• hold the tree constant and permute the group labels assigned to the tips of the tree, thus randomizing the common ancestry of members of the groups
• calculate the gsi value for a random labeling of the tree tips
• by generating many such permuted labelings and determining the gsi value for each, we obtain a distribution of values from which we can determine the frequency of gsi values equal to or greater than that which we observed for the original labeling
• this frequency is the P-value: the probability of observing by chance alone gsi values equal to or greater than the observed gsi value under the null hypothesis
example calculations
simple tree: calculations for group \( a \)

\[
gs = \frac{n}{\sum_{u=1}^{U} (d_u - 2)}
\]

numerator
\( n = 3 \)
denominator
\( (3 - 2) + (3 - 2) + (3 - 2) = 3 \)
\( gs = 3/3 = 1 \)
\( gsi = (1 - (3/7))/(1 - (3/7)) = 1 \)
\( P = 0.0121 \)
simple tree: calculations for group $b$

$$gs = \frac{n}{\sum_{u=1}^{U} (d_u - 2)}$$

**numerator**

$n = 3$

**denominator**

$$(3 - 2) + (3 - 2) + (3 - 2) + (3 - 2) = 4$$

$gs = 3/4$

$gsi = ((3/4) - (3/7))/(1 - (3/7)) = 0.563$

$P = 0.0130$
Minimum possible gs value for group \( a \)

\[ gs = \frac{n}{\sum_{u=1}^{U} (d_u - 2)} \]

\[ \min(gs) = \frac{n}{\sum_{i=1}^{U} (d_i - 2)} \]

Numerator

\( n = 3 \)

Denominator

\( (3 - 2) + (3 - 2) + (3 - 2) + (3 - 2) + (3 - 2) + (3 - 2) + (3 - 2) = 7 \)

\( gs = \frac{3}{7} \)

\( gsi = \frac{((3/7) - (3/7))/(1 - (3/7))}{(1 - (3/7))} = 0 \)
tree with polytomy: calculations for group a

\[ gs = \frac{n}{U \sum_{u=1}^{U} (d_u - 2)} \]

numerator
\[ n = 3 \]
denominator
\[ 5 - 2 = 3 \]
\[ gs = \frac{3}{3} = 1 \]
\[ gsi = \frac{(1 - (3/7))}{(1 - (3/7))} = 1 \]
\[ P = 0.0145 \]
the gsi and lack of topological resolution

• the gsi value of a group is affected by polytomies only if the polytomies affect exclusive ancestry of the group with respect to other groups
  • because resolutions of polytomies involving only one group have no influence on the extent to which individuals of that group share ancestry with other groups on a tree
• polytomies involving more than one group increase the number of ancestors over the minimum possible and will lower gsi values from their maximum
integrating across an ensemble of genealogies

- when analyzing multiple loci
- when accounting for uncertainty of inferred relationships
  - bootstrap: topologies from the bootstrap replicates
  - Bayesian analysis: topologies from the credible interval
- gsi\(_T\): weighted sum of gsi across genealogies
- advantages of the ensemble statistic
  - appropriately accounts for all available information
  - expected value does not change
  - reduced variance through averaging
integrating across an ensemble of genealogies

\[ gsi_T = \sum_{t=1}^{T} gsi_t P_t \]

where \( T \) is the total number of topologies in the ensemble
\( gsi_t \) is the gsi value based on topology \( t \)
and \( P_t \) is the probability of the topology \( t \)
the probability of each topology is its proportional representation in the sample of genealogies
results from coalescent simulations
the divergence process: diagrammatically
gsi increases rapidly with divergence
statistical power

• power is defined as the probability of rejecting the null hypothesis when it is false
• gsi: power is the proportion of permutation tests with $P < 0.05$ for the observed value of gsi across the genealogies
• monophyly: power is the proportion of genealogies in which at least one of the two group is monophyletic
power of the gsi is greater than monophyly

![Graph showing the comparison between gsi and monophyly power over time (Ne generations).]
power of the gsi is insensitive to number of loci
power of the gsi is insensitive to number of loci
• the Fixation index statistic $F_{ST}$ is a comparative measure of the genetic divergence among subpopulations
• $S$ represents subpopulations, and $T$ represents total population
• it is a single measure across all subpopulations included in the analysis
gsi compared to $F_{ST}$
power of the gsi is greater than $F_{ST}$
application to empirical data
delimiting two putative dolphin species

*Lagenorhynchus obliquidens*  
Pacific white-sided dolphin

*Lagenorhynchus obscurus*  
dusky dolphin
data and reanalysis

• DNA sequence data from Hare, Cipriano and Palumbi (2002)
  • four nuclear gene introns
    • ACT, actin
    • BTM, Butyrophilin
    • CAMK, calcium calmodulin-dependent kinase
    • HEXB, lysosomal beta-hexosaminidase
  • calculated gsi\textsubscript{T} integrating across 2000 bootstrap replicates
bootstrap consensus trees

ACT

BTM

CAMK

HEXB

L. obliquidens 143
L. obliquidens 143b
L. obliquidens 701
L. obliquidens 1495
L. obscurus NZ8
L. obscurus NZ8b
L. obscurus MFB178
L. obscurus NZ11
L. obscurus MFB178
L. acutus 9300

0.52
0.51
0.62
0.57
0.63
0.63
0.64
0.86
0.63
0.87
0.63
0.63
0.63
0.63
0.51

L. obliquidens 143
L. obliquidens 1137
L. obliquidens 1289
L. obliquidens 1454
L. obscurus NZ8
L. obscurus NZ11
L. obscurus K99
L. obscurus 178
L. obscurus 236
L. obliquidens 701b
L. obliquidens 1137
L. obliquidens 1477
L. obliquidens 1477b
L. obliquidens 1495
L. obscurus NZ8
L. obscurus NZ8b
L. obscurus MFB178
L. obscurus NZ11
L. obscurus MFB178
L. obscurus 236b
L. acutus 9300

0.54
0.51
0.52
0.51

ACT CAMK HEXB
<table>
<thead>
<tr>
<th>gene</th>
<th>$\text{gsi}_T$</th>
<th>$P$</th>
<th>$\text{gsi}_T$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>0.3802</td>
<td>0.0089</td>
<td>0.764</td>
<td>0.0006</td>
</tr>
<tr>
<td>BTM</td>
<td>0.2885</td>
<td>0.0001</td>
<td>0.1213</td>
<td>0.0173</td>
</tr>
<tr>
<td>CAMK</td>
<td>0.3531</td>
<td>0.0002</td>
<td>0.2474</td>
<td>0.0015</td>
</tr>
<tr>
<td>HEXB</td>
<td>0.0353</td>
<td>0.0112</td>
<td>0.1085</td>
<td>0.0002</td>
</tr>
<tr>
<td>all (gsi$_T$)</td>
<td>0.2642</td>
<td>0.0001</td>
<td>0.3103</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

L. obliquidens

L. obscurus

$\text{gsi}_T$: bootstrap integration

L. obliquidens

L. obscurus
genetics of divergence in *Anopheles gambiae*
Genetics of divergence in *Anopheles gambiae*

- Two “molecular forms” of *Anopheles gambiae* exist
- Thought to represent incipient species
- Defined by fixed sequence differences in the X-linked ribosomal DNA locus
- Different, but overlapping, geographical distribution in West Africa
- Appear to be some differences in larval habitat
- Previous research showed some differential $F_{ST}$ values in regions of the X chromosome
data and reanalysis

• data from Stump et al. (2005)
• M form (15 individuals) and S form (10 individuals) sampled from areas of sympatry
• intron sequences from 22 loci along X chromosome
• calculated gsiT integrating across 1000 bootstrap replicates
gsi and P-values along the X chromosome

<table>
<thead>
<tr>
<th>M</th>
<th>S</th>
<th>M</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>gsi</td>
<td>gsi</td>
<td>P-value</td>
<td>P-value</td>
</tr>
<tr>
<td>0.054</td>
<td>0.074</td>
<td>0.826</td>
<td>0.295</td>
</tr>
<tr>
<td>0.058</td>
<td>0.098</td>
<td>0.050</td>
<td>0.146</td>
</tr>
<tr>
<td>0.071</td>
<td>0.027</td>
<td>0.026</td>
<td>0.568</td>
</tr>
<tr>
<td>&lt; 0.001</td>
<td>0.017</td>
<td>0.381</td>
<td>0.211</td>
</tr>
<tr>
<td>0.063</td>
<td>0.087</td>
<td>0.449</td>
<td>0.094</td>
</tr>
<tr>
<td>0.077</td>
<td>0.016</td>
<td>0.346</td>
<td>0.998</td>
</tr>
<tr>
<td>0.071</td>
<td>0.057</td>
<td>0.615</td>
<td>0.608</td>
</tr>
<tr>
<td>0.087</td>
<td>0.162</td>
<td>0.122</td>
<td>0.005</td>
</tr>
<tr>
<td>0.236</td>
<td>0.854</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>0.022</td>
<td>0.092</td>
<td>0.003</td>
<td>0.596</td>
</tr>
<tr>
<td>0.050</td>
<td>0.109</td>
<td>0.596</td>
<td>0.066</td>
</tr>
<tr>
<td>0.566</td>
<td>0.515</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>0.113</td>
<td>0.130</td>
<td>0.407</td>
<td>0.155</td>
</tr>
<tr>
<td>0.670</td>
<td>0.385</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>0.005</td>
<td>0.856</td>
<td>0.196</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>0.137</td>
<td>0.187</td>
<td>0.194</td>
<td>0.011</td>
</tr>
<tr>
<td>0.111</td>
<td>0.098</td>
<td>0.414</td>
<td>0.454</td>
</tr>
<tr>
<td>0.761</td>
<td>0.240</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>0.990</td>
<td>0.959</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>0.998</td>
<td>0.997</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>0.039</td>
<td>0.895</td>
<td>0.042</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>0.986</td>
<td>0.996</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
attributes of the genealogical sorting index

- intuitive
- simple and easily calculated
- normalized to facilitate comparisons among groups and trees
- applicable to each group separately rather than a composite measure for all groups
- applicable to trees with polytomies
- accommodating to uncertainty in phylogenetic relationships as determined by bootstrap or Bayesian analyses
- quantifiable on any tree regardless of how it is generated