

3502-470 Plant Genetic Resources

Prof. Karl Schmid SS 2025

Institute of Plant Breeding, Seed Science and Population Genetics University of Hohenheim

1/24

Recall from last time

- Measures of genetic diversity
- Single marker: Gene diversity
- Multiple markers: Average gene diversity, haplotype diversity
- $\cdot\,$ DNA sequences: Nucleotide polymorphism, nucleotide diversity

2 / 24

Maize land races of Peru





Where do the differences come from?

- $\cdot\,$ Phenotypes are controlled by genetic information
- Differences between individuals: due to environment (not really inheritable) or genetic differences (heritable)
- \cdot Genetic differences: due to mutation, recombination
- Trace back the ancestry!
- Easy case: If genetic data always traces back to <u>one</u> ancestor (e.g. chloroplast or mitochondrial DNA) \Rightarrow Tree

4 / 24

Darwin thinks





A very general setting

Goal: Group entities that are more similar than others by reconstructing the ancestral tree

Definitions:

- Operational taxonomic unit, OTU: Things we want to cluster
- Examples: Species, different land races, different individuals

7 / 24

Key concepts of phylogenetic analysis

- Phylogenetic tree: Ancestral tree of the OTUs
- Species with a shared evolutionary history are more similar
- · Reconstruct the ancestral tree using present data
- Ancestral versus derived traits (alleles): Can we differentiate?
- Should we use phenotypic or genotypic variation for phylogenetic reconstruction?

8/24

Application of phylogenetics in plant breeding

- Reconstruct domestication history
- \cdot Identify extended gene pools (GP 1, 2, 3 according to Harlan)
- \cdot Identify introgressed genes
- Identify gene flow between:
 - Wild and domesticated species
 - Modern varieties and land races
 - Conventional cultivars and GMOs
- Find out what your competitors are doing!

Information about relatedness may help to assess how much new variation crossing partner bring, e.g. interesting for heterosis



- Rooted: Tree has direction, genealogy traces back to the most recent common ancestor (MRCA)
- To trace back to the true MRCA: Need knowledge of ancestral and derived alleles
- Usual approach: Ancestral allele = outgroup allele (of sister taxa)

Anatomy of a phylogenetic tree



- Tips of tree: OTUs, also called leaves
- Branches: Line segments, connect 2 nodes or node + OTU
- Branches are internal if they do not connect an OTU, split OTUs in 2 groups
- Clade (only rooted tree): set of all OTUs connected to an internal branch on the side not connected to the MRCA.
- Ancestor of the clade: Node connected on this side of the branch

Phylogeny of barley with focus on Quingke barley in Tibet



Hpu: H. pubiflorum, Hsp: H. spontaneum, Hag: H. agriocrithon, western: landraces in clade/cluster I, eastern: landraces in clade/cluster II, TWB: Tibetan weedy barley

Methods for calculating/reconstructing phylogenetic trees

- Distance-based
- Parsimony-based
- Maximum likelihood
- Bayesian methods

13 / 24

Distance-based methods

Two-step approach:

- \cdot Calculate a pairwise distance matrix between OTUs
- \cdot Cluster OTUs according to distance matrix

Which distance matrix to use?

e.g.,

• Hamming distance: Count number of differences, distances are not units of time. Sometimes divided by sequence length.

SEQ1 GATCTATCTACTAC SEQ2 ..G....C...A

14 / 24

Hamming distance = 3

Distances based on models of sequence evolution

- Estimate the time needed to produce the observed differences between sequences.
- $\cdot\,$ Depends on the rate for mutations from one nucleotide to the other
- All rates equal: Jukes-Cantor
- K80 (from Kimura): rate for transversions $\{A, G\} \leftrightarrow \{C, T\}$ different from rate for transitions $A \leftrightarrow G, C \leftrightarrow T$



- 2. Create ancestor node (i, j), connect it to ij. The 2 new branches have lengths s.t. (i, j) has distance $dist_{ij}/2$ to all OTUs connected.
- 3. Calculate the distances between (i, j) and each non-connected node k

$$dist_{k,(i,j)} = \frac{1}{2}dist_{k,i} + \frac{1}{2}dist_{k,j}.$$

4. Repeat until (artificial) root reached



Maximum Parsimony (MP)

- $\cdot\,$ Based on Occam's Razor: The simpler model is the preferred one
- Find tree that matches the data and is constructed with the least number of substitutions

Approach

- Determine all possible trees
- Calculate the number of substitutions
- \cdot Select the tree with the least number of substitutions

Further alternative: Based on a model of sequence evolution, search the tree (from all trees) that produces the data with highest likelihood $$_{\rm 18/24}$$

Bo	Bootstrapping: Confidence in reconstructed trees																		
			1	2	3	4	5	6	7	8		1	2	5	4	5	6	1	8
		А	G	С	А	G	Т	А	С	Т	А	G	С	Т	G	Т	А	G	Т
		В	G	Т	А	G	Т	А	С	Т	В	G	Т	Т	G	Т	А	G	Т
		С	Α	С	А	А	Т	А	С	С	С	A	С	Т	А	Т	А	А	С
		D	A	С	А	А	С	А	С	Т	D	A	С	С	А	С	А	А	Т
		÷									÷								
	Appro	oacl	h																
	۰Rá	and	oml	ly d	raw	ı ali	gnr	ner	nt p	osit	ions	wi	th r	epl	ace	me	nt		
	۰Cr	eat	e a	pse	eud	0-Si	amp	ole											
	• Ar	naly	se t	the	dat	ta (e	e.g.,	clu	iste	r w	ith L	IPG	MA)						

- Repeat many times (>1000 times)
- Ask how often a particular clade/split is observed

19 / 24

 $\cdot\,$ Bootstrap value of 95% for clade \Rightarrow Strong support

UPGMA-reconstructed phylogeny of genus Secale





Some recommendations

- Distance-based: UPGMA is not optimal (strict molecular clock cannot vary rate of molecular changes on different branches), use similar method neighbor-joining which corrects for this
- Model-based: Despite long running times and increased complexity is quite accurate and allows many futher analyses. Widely used:
 - RAxML https://cme.h-its.org/exelixis/web/software/raxml/
 - · BEAST/SNAPP http://www.beast2.org/
- Always bootstrap!

22 / 24

Further reading

- $\cdot\,$ (Hartl and Clark, 2007): Principles of Population Genetics, Chapter 7.8
- Barry Hall: Phylogenetics Trees made easy, 3rd edition. Sinauer Associates – A How-To manual of phylogenetics
- Joseph Felsenstein: Inferring Phylogenies. Sinauer Associates The standard book on phylogenetic methods.

23 / 24

References i

Hartl D, Clark A (2007) Principles of Population Genetics, 4th edn. Sinauer Associates