Intraspecific gene genealogies: trees grafting into networks

by David Posada & Keith A. Crandall

Kessy Abarenkov
Tartu, 2004

Article describes:

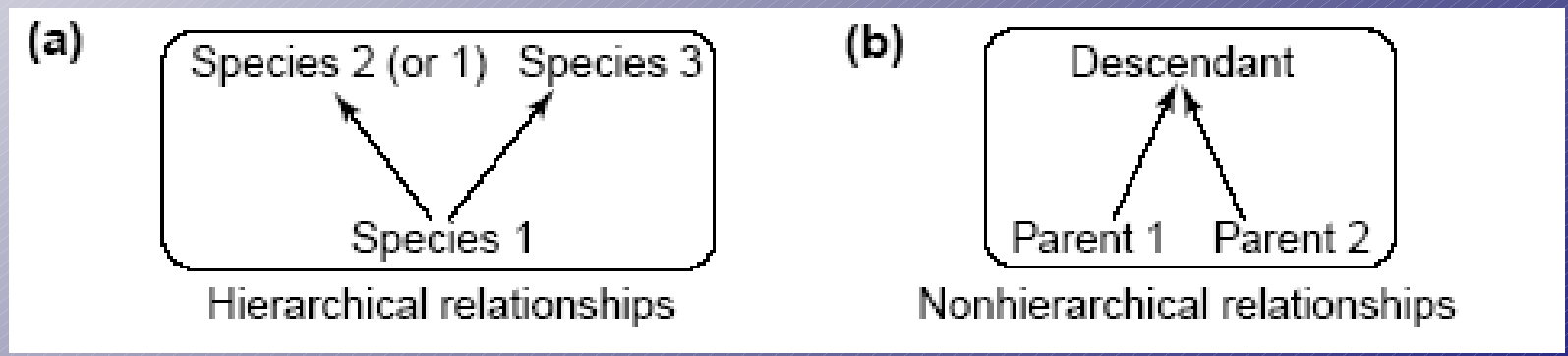
- Population genetics principles
- Intraspecific genetic variation – why networks?
- Available methods and software

Population genetics principles

- Coalescent event – the time inverse of a DNA replication event; that is, the event leading to the common ancestor of two sequences looking back in time
- Haplotype space – the collection of points representing the possible different haplotypes. The dimension of this space is the number of characters (L)
- Homoplasy – a similarity that is not a result of common history. It is caused by parallel, convergent or reverse mutations
- Tokogeny – nonhierarchical genetic relationships among individuals. Arising by sexual reproduction

Problems with interspecific methods at the intraspecific level

- Interspecific relationships are hierarchical
  - reproductive isolation, population fission over longer timescales
  - mutation + population divergence = fixation of different alleles
  - nonoverlapping gene pools
- intraspecific are not
  - result of sexual reproduction,
  - smaller numbers of relatively recent mutations,
  - frequently recombination



Problems with interspecific methods at the intraspecific level

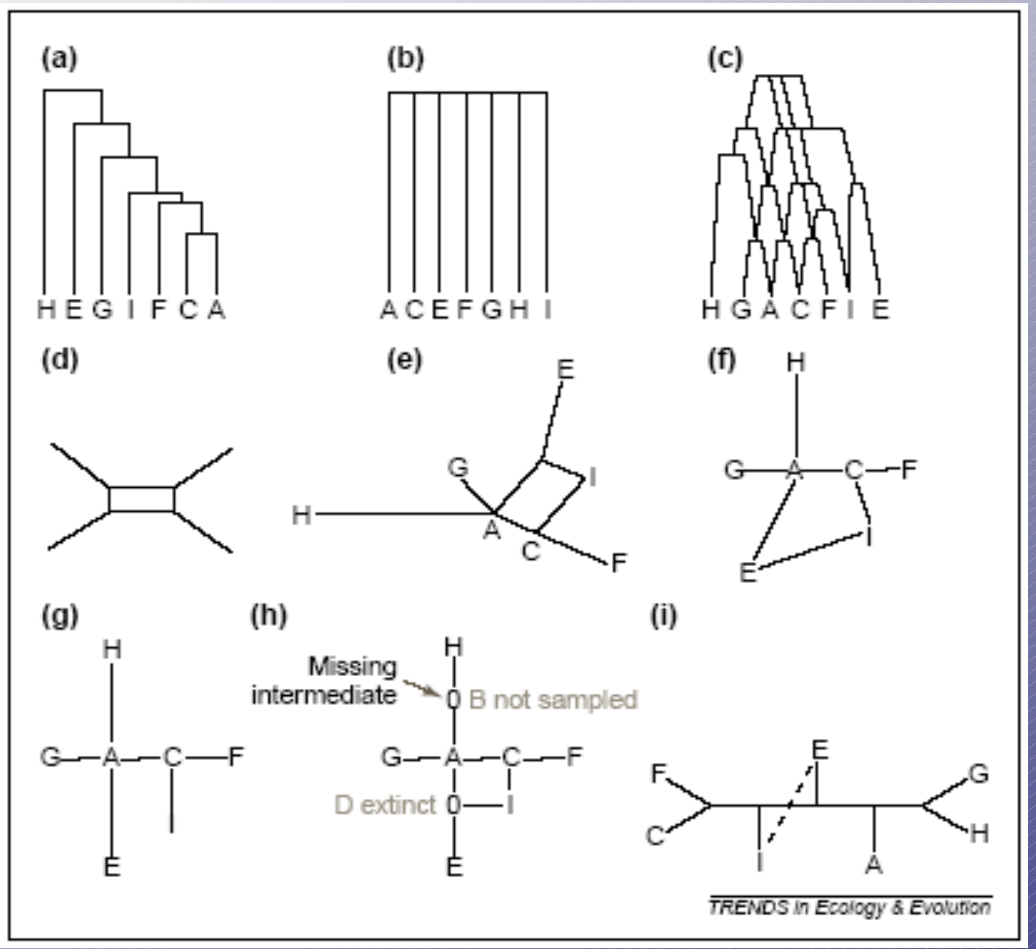
- More traditional methods (ML, MP, ME) cannot take account of the fact that, at the populational level, several phenomena violate some of them assumptions
  - low divergence -> fewer characters for phylogenetic analysis
  - extant ancestral nodes (ancestral haplotypes are expected to persist in the population and to be sampled together with their descendants)
  - multifurcations (single ancestral haplotype will often give rise to multiple descendant haplotypes, yielding a haplotype tree with true multifications)
  - reticulation caused by recombination between genes, hybridization between lineages, and homoplasy
  - large sample sizes

Solution: network methods

- Networks
  - can account effectively for processes acting at the species level
  - allow for persistent ancestral nodes, multifurcations and reticulations
  - provide a way of representing more of the phylogenetic information present in a data set (loops -> recombination, reverse or parallel mutations)
- Can and have been used for:
  - detecting recombination
  - delimiting species
  - inferring models of speciation
  - partitioning population history and structure
  - studying genotype and phenotype associations

Network algorithms

(a) UPGAM
(b) Maximum parsimony
(c) Pyramid
(d) Statistical geometry
(e) Split decomposition
(f) Minimum spanning network
(g) Median-joining network
(h) Statistical parsimony
(i) Reticulogram

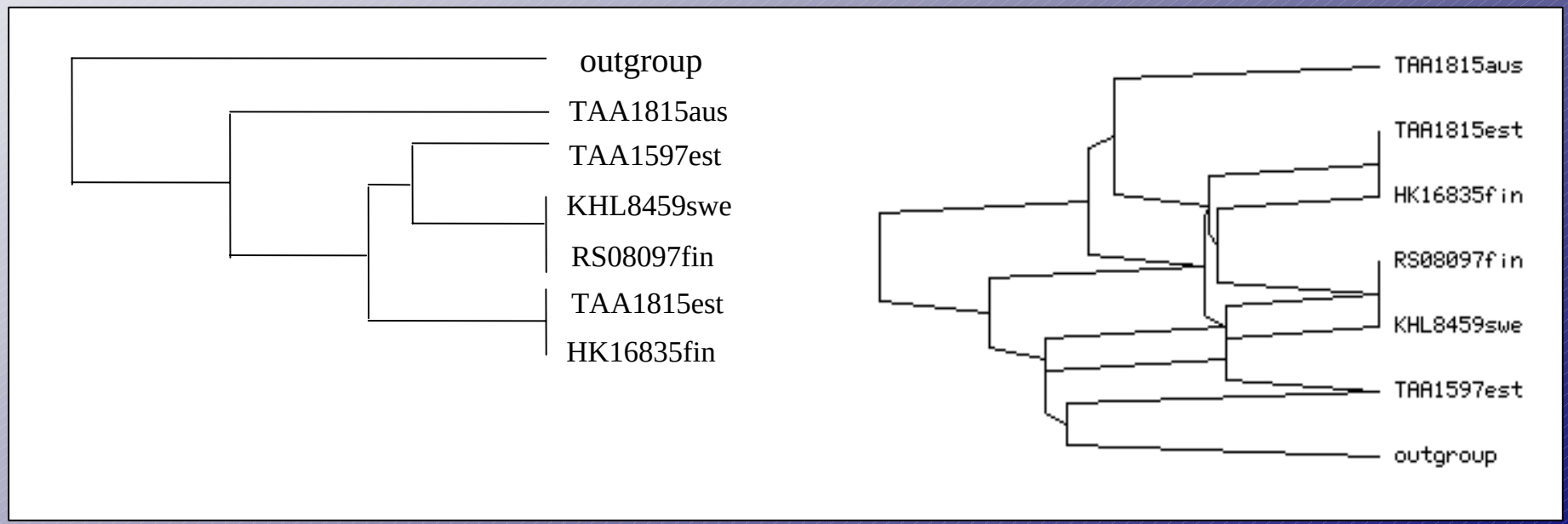


Methods	Category[a]	Software	Speed	Input data	Model of evolution	Reticulations	Statistical assessment
Pyramids	Distance	Pyramids	Fast	Distances	Yes	Yes	No
Statistical geometry	Distance invariants	Geometry, Statgeom	Fast	Multistate	Yes	Yes	Yes
Split decomposition	Distance parsimony	SplitsTree	Fast	Multistate	Yes	Yes	Yes
Median networks	Distance	No	Slow	Binary	No	Yes	No
Median-joining networks	Distance	Network	Very fast	Multistate	No	No	No
Statistical parsimony	Distance	TCS	Fast	Multistate	No	Yes	Yes
Molecular-variance parsimony	Distance	Arlequin	Fast	Multistate	Yes	Yes	Yes
Netting	Distance	No	Slow	Multistate	No	Yes	No
Likelihood network	Likelihood	PAL	Slow	Multistate	Yes	Yes	Yes
Reticulogram	Least squares	T-rex	Fast	Distances	No	Yes	Yes
Reticulate phylogeny	Least squares	No	Slow	Distances[b]	Yes	Yes	Yes

[a]Details of software programs given in Box 3; [b]Distances estimated from gene frequency data.

Pyramids

- An extension of the hierarchical clustering framework
- Represent a set of clades that can overlap
- Can be used to represent reticulate events (among terminal nodes that are sister taxa)
- http://bioweb.pasteur.fr/seqanal/interfaces/pyramids.html



Statistical geometry

- haplotypes are considered as geometric configurations in the haplotype space
- does not offer an estimate of the sequence genealogy
- incorporates a model of nucleotide substitution through the estimation of haplotype distances and allows a reliable assessment of the derived conclusions

- Geometry
- Statgeom

Split decomposition

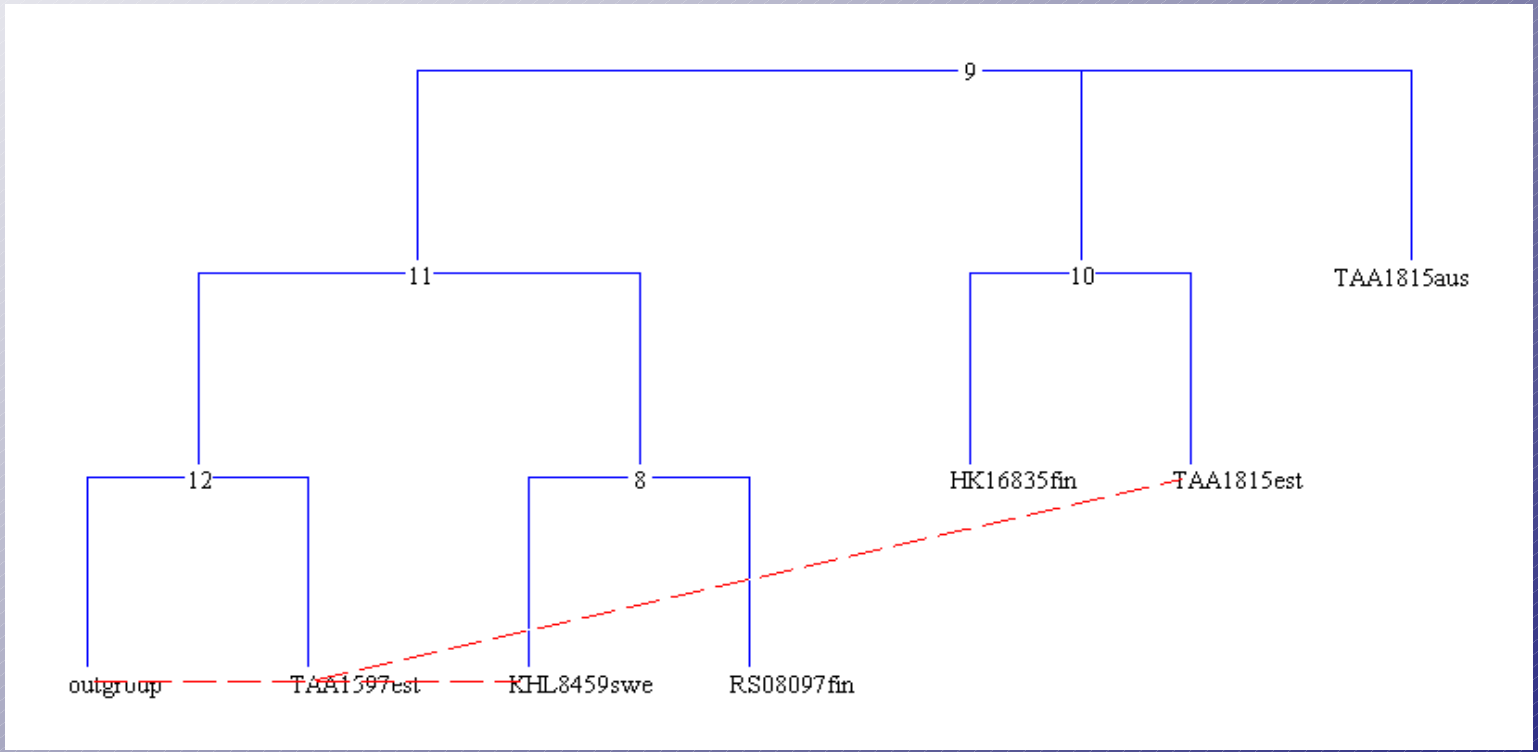
- data set is partitioned into set of sequences or 'splits'
- network is built by taking these splits and combining them successively
- when splits are incompatible, a loop is introduced to indicate that there are alternative splits
- fast, nucleotide or protein data, allows for the inclusion of models of nucleotide substitution or amino acid replacement, bootstrap evaluation
- SplitsTree web interface - http://bibiserv.techfak.uni-bielefeld.de/splits/

Statistical parsimony

- begins by estimating the maximum number of differences among haplotypes as a result of single substitutions with a 95% statistical confidence
- after that, haplotypes differing by one change are connected, then those that differing by two, by three and so on, until all the haplotypes are included in a single network or the parsimony connection limit is reached
- the statistical parsimony method emphasizes what is shared among haplotypes that differ minimally rather than differences among the haplotypes, and provides an empirical assessment of deviations from parsimony
- TCS - http://darwin.uvigo.es/software/tcs.html

Reticulogram

- addition of reticulations to a bifurcating tree
- the minimum number of reticulations required to maximize the fit of the network to the data is calculated
- T-Rex - http://www.info.uqam.ca/~makarenv/trex.html



Other methods

- Median networks (no program)
- Median-joining networks (not suitable at the population level, requires the absence of recombination)
- Molecular-variance parsimony (uses sampled haplotype frequencies and geographic subdivisions to present the solution in the form of a set of (near) optimal networks, http://lgb.unige.ch/arlequin/)
- Netting (no program)
- Likelihood network (open-source Java library, > 120 modules, http://www.pal-project.org/)

Instraspecific gene genealogies: trees grafting into networks, David Posada and Keith A. Crandall, TRENDS in Ecology & Evolution Vol. 16 No. 1 January 2001