Phylogenetic inference from raw word lists

Gerhard Jäger

Institute of Linguistics, Tübingen University

Language Diversity Congress Groningen

July 20, 2013

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Overview

- most work in computational phylogeny of languages uses expert cognate judgments
- such data are only available for few language families
- alternative:
  - phylogenetic inference directly from (uniform phonetic transcriptions of) raw Swadesh lists
  - distance-based instead of character-based phylogenetic inference
- assessing the quality of derived phylogenies poses a considerable challenge

The Automated Similarity Judgment Program

- Project at MPI EVA in Leipzig around Sören Wichmann
- covers more than 5,000 languages
- basic vocabulary of 40 words for each language, in uniform phonetic transcription
- freely available

used concepts: *I, you, we, one, two, person, fish, dog, louse, tree, leaf, skin, blood, bone, horn, ear, eye, nose, tooth, tongue, knee, hand, breast, liver, drink, see, hear, die, come, sun, star, water, stone, fire, path, mountain, night, full, new, name*

Automated Similarity Judgment Project

concept	Dutch	English
I	ik	Ei
you	yEi, y3	yu
we	vEi, v3	wi
one	en	8is
two	tve	8Et
person	%pErson, mEns	pers3n
fish	vis	fiS
dog	hont	dag
louse	l3is	laus
tree	bom	tri
leaf	blat	lif
skin	h3id, vEl	skin
blood	blut	bl3d
bone	ben	bon
horn	horn	horn
ear	or	ir
eye	oX	Ei

concept	Dutch	English
nose	nes	nos
tooth	tant	tu8
tongue	toN	t3N
knee	kni	ni
hand	hant	hEnd
breast	borst	brest
liver	lev3r	liv3r
drink	driNk3n	drink
see	zin	si
hear	hor3n	hir
die	stErv3n	dEi
come	kom3n	k3m
sun	zon	s3n
star	ster	star
water	vat3r	wat3r
stone	sten	ston
fire	vir	fEir

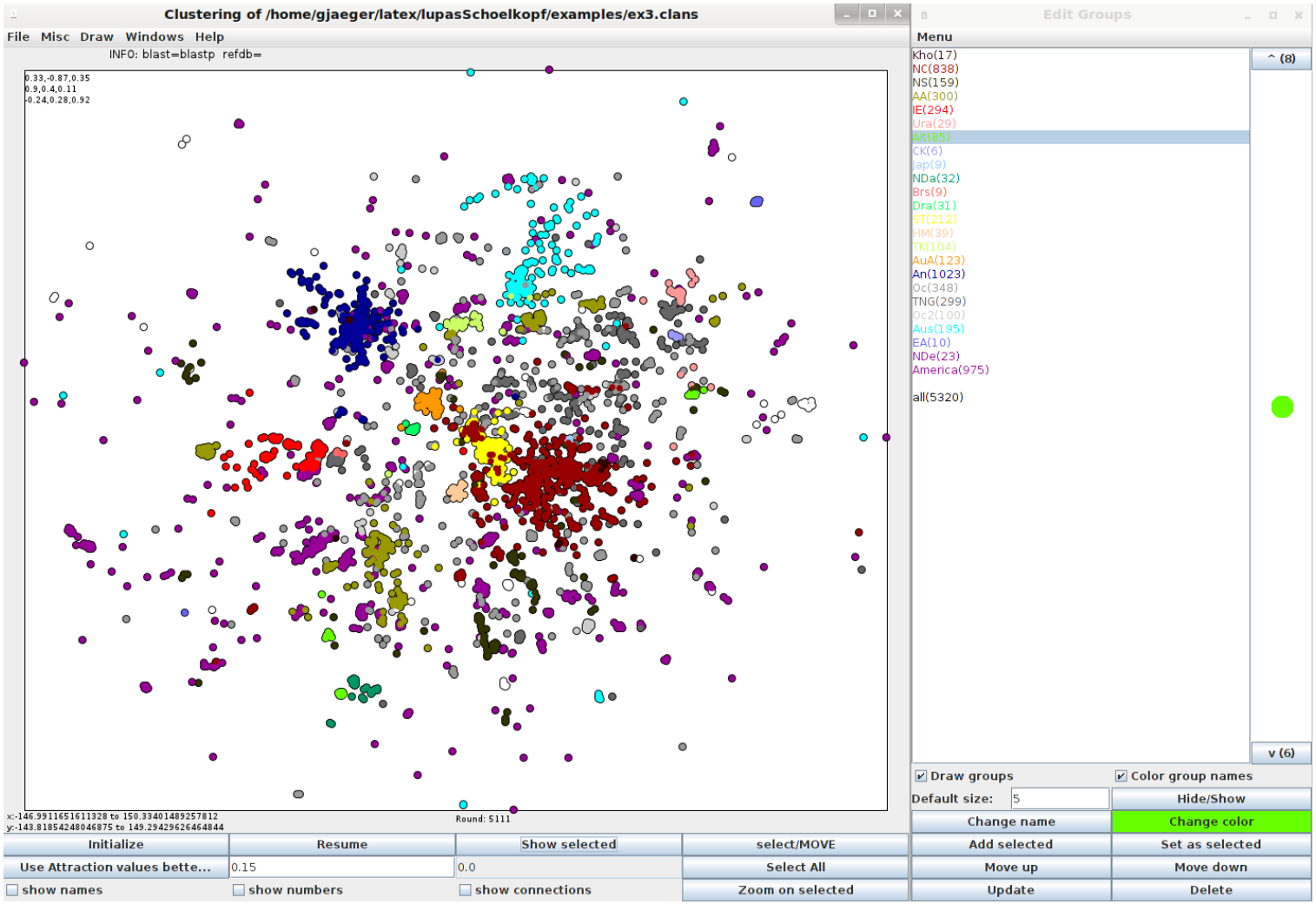
Benchmark: LDND measure

English / Dutch

	Ei	yu	wi	w3n	tu	fiS	...
ik	1	1	1	1	1	$^2/_3$	
y3	1	$^1/_2$	1	$^2/_3$	1	1	
v3	1	1	1	$^2/_3$	1	1	
en	1	1	1	$^2/_3$	1	1	
tve	1	1	1	1	$^2/_3$	1	
vis	$^2/_3$	1	$^2/_3$	1	1	$^2/_3$	
⋮							

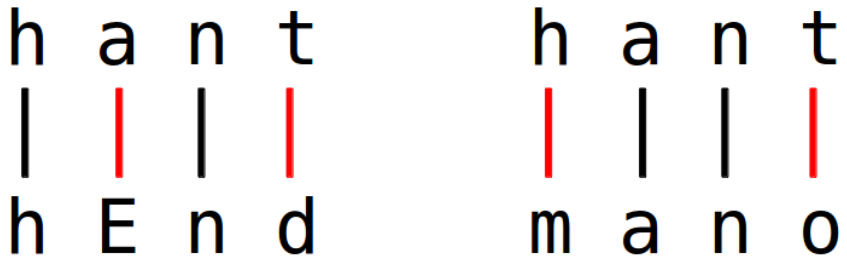
- average normalized Levenshtein distance along diagonal: 0.56
- average normalized Levenshtein distance off diagonal: 0.89
- LDND: $^{0.56}/_{0.91} = 0.62$

Benchmark: LDND measure



Weighted alignment

- Levenshtein distance is somewhat coarse grained



- simply normalized distance is 0.5 in both cases
- degrees of (dis-)similarity between sounds not taken into account

Alignment via Pointwise Mutual Information

- PMI (a.k.a. *log-odds*):

$$s(a,b) = \log \frac{p(a,b)}{q(a)q(b)}$$

  - $p(a,b)$: probability of sound a being etymologically related to sound b in a pair of cognates
  - $q(a)$: relative frequency of sound a
- training of parameters requires corpus of (ideally: etymologically correctly) aligned cognate word pairs

Determining weights

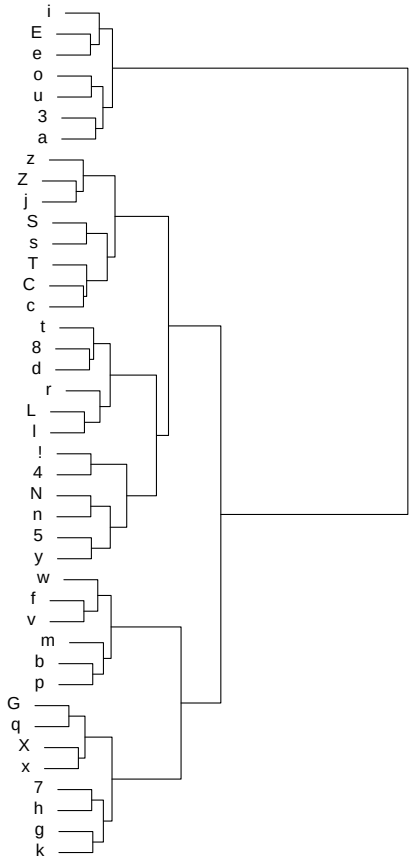
- quick and dirty method:
  1. estimate $q(a)$ as relative frequency of a in your entire training set
  2. define a collection of pairwise related language pairs
  3. do Levenshtein alignment of translation pairs
  4. estimate $p(a, b)$ as relative frequency of pairings of a with b
- starting with binary weights, steps 2–4 are repeated using the parameters obtained in the previous round; only word pairs with alignment score > 0 are used

Estimated PMI values

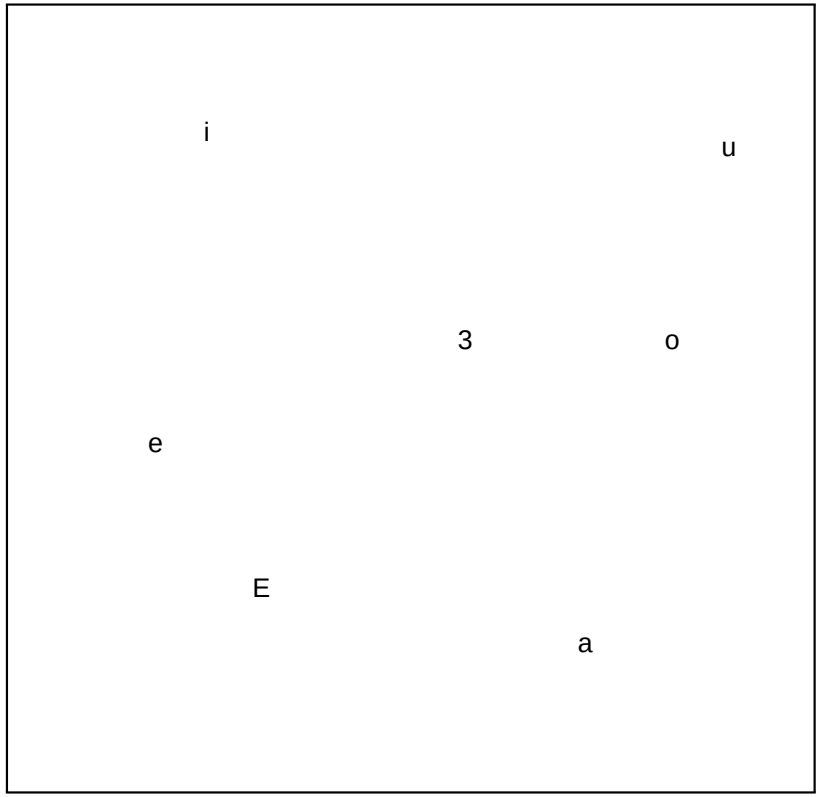
	a	e	i	o	u	b	p	d	t	g	k	h
a	1.93	−0.61	−2.04	−0.41	−1.57	−9.14	−8.94	−7.91	−8.14	−8.81	−6.72	−5.06
e	−0.61	2.34	−0.10	−0.64	−1.89	−7.99	−7.79	−5.56	−7.91	−6.57	−6.58	−6.46
i	−2.04	−0.10	2.19	−1.53	−1.03	−7.32	−8.50	−8.58	−6.54	−6.18	−6.67	−3.86
o	−0.41	−0.64	−1.53	2.21	0.70	−8.28	−8.08	−7.46	−8.89	−7.95	−8.26	−4.24
u	−1.57	−1.89	−1.03	0.70	2.70	−8.38	−6.80	−8.26	−9.00	−8.06	−8.37	−6.29
b	−9.14	−7.99	−7.32	−8.28	−8.38	3.76	0.46	−2.04	−1.79	−0.56	−2.43	−1.56
p	−8.94	−7.79	−8.50	−8.08	−6.80	0.46	3.60	−3.68	−1.65	−1.87	−1.67	−1.66
d	−7.91	−5.56	−8.58	−7.46	−8.26	−2.04	−3.68	3.81	0.38	−0.93	−1.91	−1.06
t	−8.14	−7.91	−6.54	−8.89	−9.00	−1.79	−1.65	0.38	3.40	−3.07	−0.97	−1.70
g	−8.81	−6.57	−6.18	−7.95	−8.06	−0.56	−1.87	−0.93	−3.07	3.22	0.45	−0.92
k	−6.72	−6.58	−6.67	−8.26	−8.37	−2.43	−1.67	−1.91	−0.97	0.45	2.83	−1.03
h	−5.06	−6.46	−3.86	−4.24	−6.29	−1.56	−1.66	−1.06	−1.70	−0.92	−1.03	3.03

Estimated PMI values

- hierarchical clustering



- MDS of vowels



Weighted alignment

left: Levenshtein alignment; right: weighted alignment

-iX	iX-
ego	ego
du	du
tu	tu
vir	vir
nos	nos
ains	ain-s
unus	-unus
cvai	cvai
-duo	duo-
---mEnS	mEnS---
persona	persona
---fiS	fiS---
piskis	piskis
hun-t	hun-t
kanis	kanis
-----laus	------laus
pedikulus	pedikul-us
-baum	--baum
arbor	arb-or

-blat	b-lat
folyu	folyu
haut--	haut--
-kutis	k-utis
---blut	---blut
saNgwis	saNgwis
knoX3n	knoX3n
--o--s	--os--
horn-	horn-
kornu	kornu
-au-g3	a-ug3-
okulus	okulus
na-z3	naz3-
nasus	nasus
chan	chan-
dens	d-ens
-chuN3	chuN--3
liNgwE	-liNgwE
-kni	k-ni
genu	genu

han-t	han-t
manus	manus
--brust	b--rust
pektus-	pektus-
leb3r	leb3r
yekur	yekur
triNk3n	triNk3n-
-bibere	-bi-bere
--ze3n	--ze3n
widere	widere
-her3n	--her3n
audire	audire-
Sterb3n	Sterb3n
-mor--i	-mor-i-
khom3n	khom3n---
wenire	w---enire
zon3	zon3
so-l	sol-
StErn	StErn
stela	stela

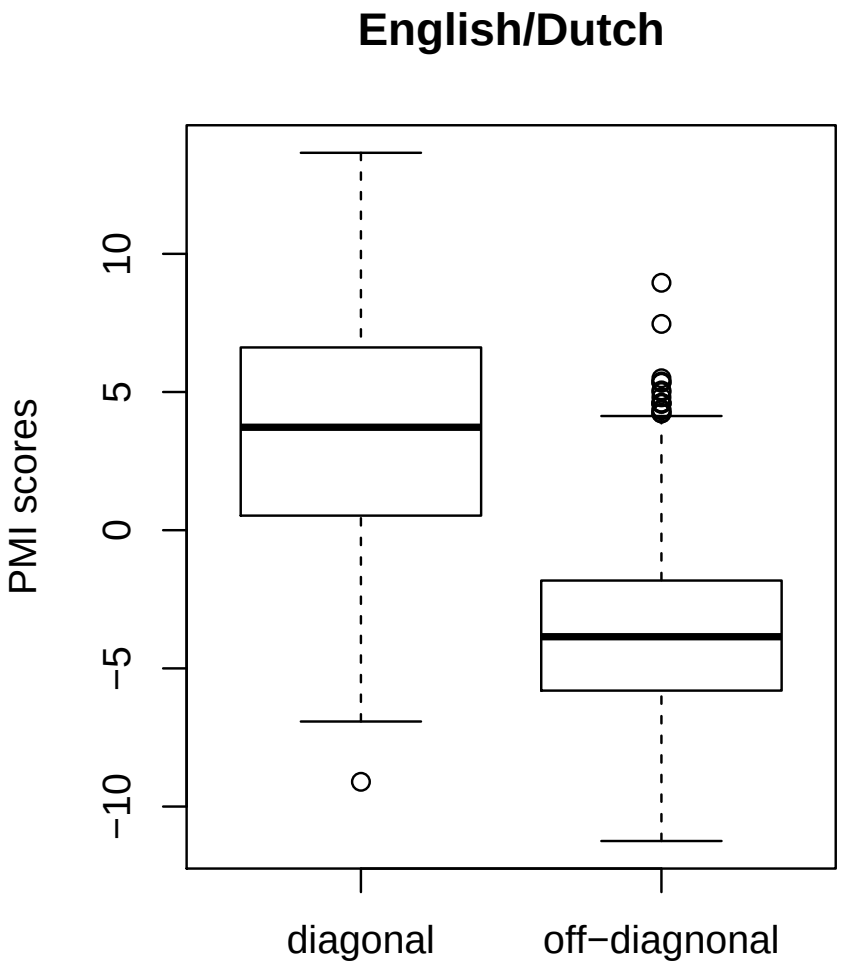
Aggregating PMI scores

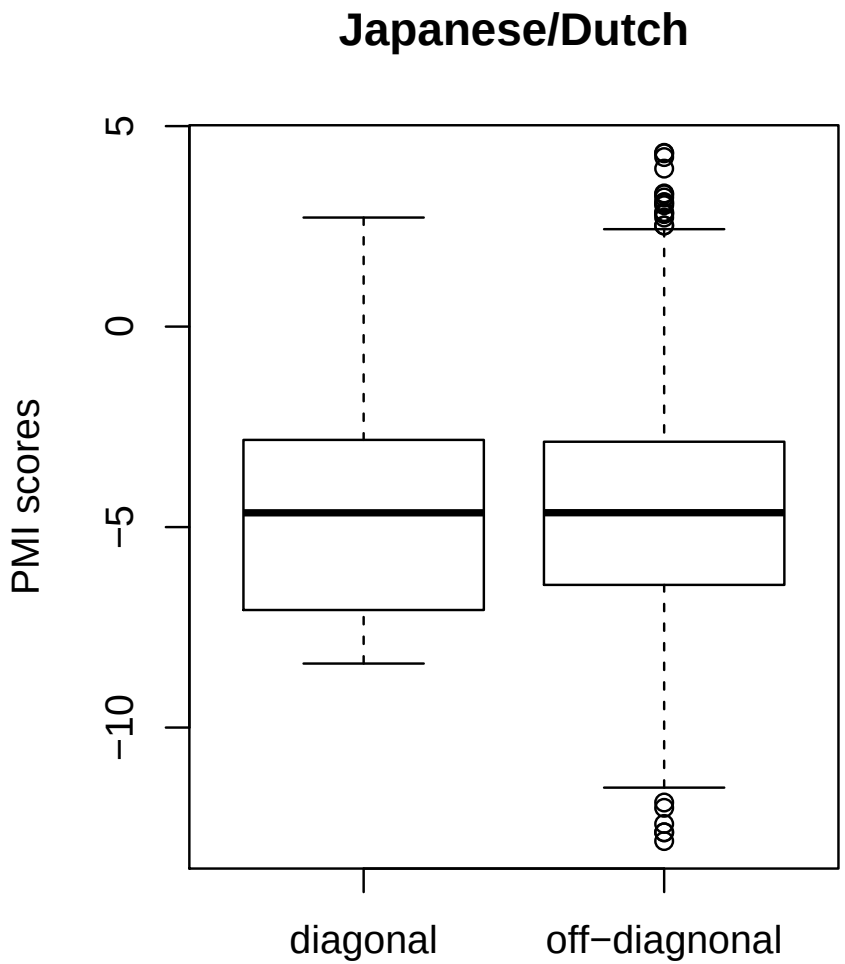
English / Dutch

	Ei	yu	wi	w3n	tu	fiS	...
ik	-1.41	-4.63	-1.41	-5.56	-4.57	-1.28	
y3	-3.79	2.92	-1.75	-0.52	-3.37	-6.09	
v3	-3.79	-3.58	0.87	2.1	-2.81	-0.11	
en	-3.7	-5.49	-3.7	0.43	-5.49	-6.07	
tve	-3.94	-6.94	-0.71	-2.77	-0.28	-1.68	
vis	-4.68	-6.3	1.58	-1.69	-5.48	5.77	
⋮							

Aggregating PMI scores

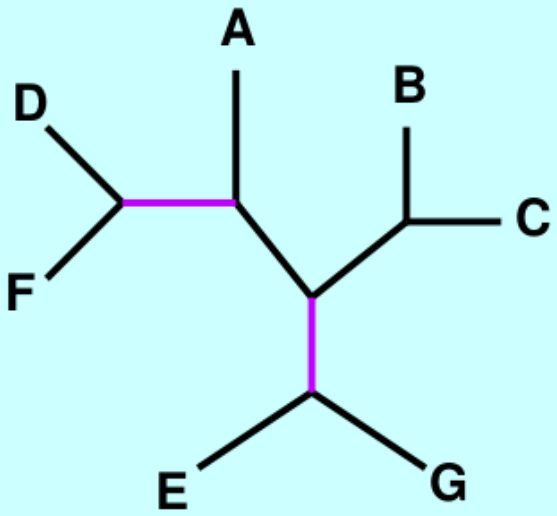
- distance between two languages is defined as distance between the diagonal distribution and the off-diagonal distribution

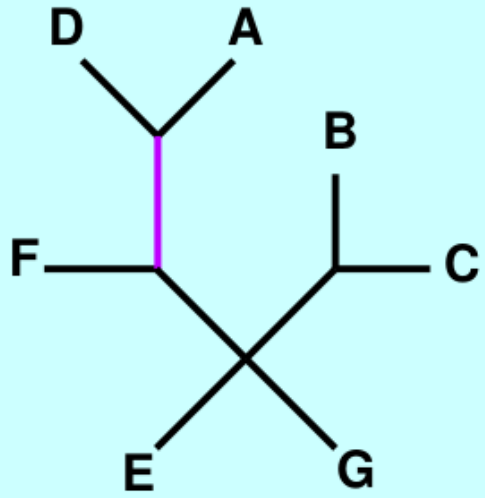




Tree distances: Robinson-Fould

The symmetric difference metric





Partitions

{ADF | BCEG}

{DF | ABCEG}

{BC | ADEFG}

{EG | ABCDF}

Partitions

{ADF | BCEG}

{AD | BCEFG}

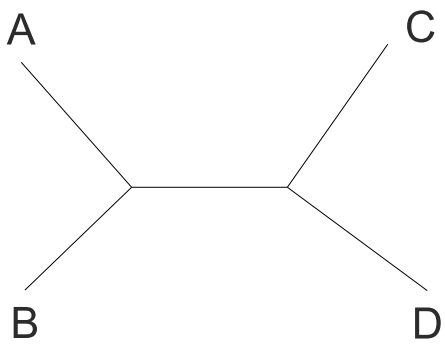
{BC | ADEFG}

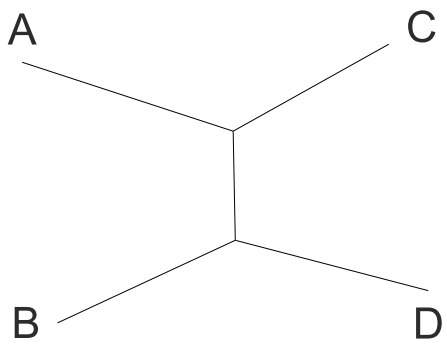
Tree distances: Robinson-Fould

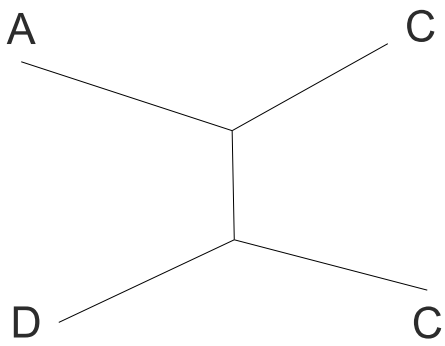
- normalized RF-distance: number of different partitions, divided by the total number of partitions in tree 1 + total number of partitions in tree 2
- in the example: $\frac{3}{4+3}$

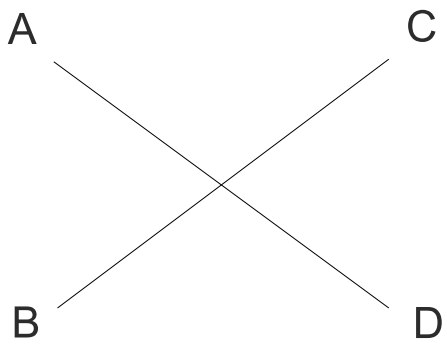
Tree distances: Quartet distance

- for a quartet of species, there are four possible tree topologies, 3 **butterflies** and 1 **star**





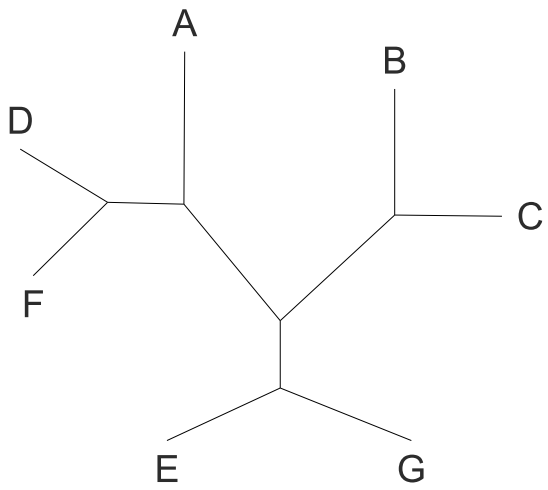


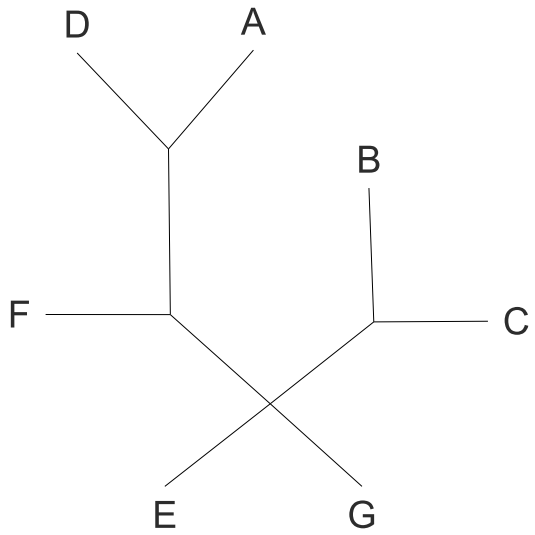


Tree distances: Quartet distance

- **quartet distance** between two unrooted trees is the number quartets that have a different topology in the two trees

Tree distances: Quartet distance

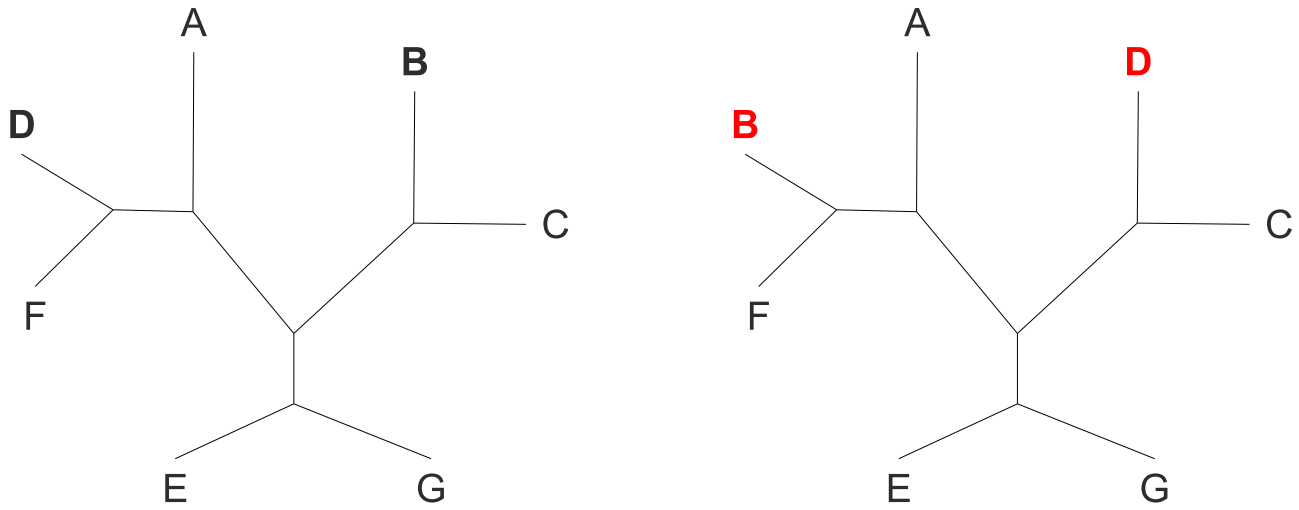




- $\binom{7}{4} = 35$ quartets in total
- 25 are shared, 10 are different
- normalized qdist: $\frac{10}{35}$

Tree distances

- Robinson-Fould distance is more intuitive, but quartet distance is more robust



- Robinson-Fould distance: 6; normalized $\frac{6}{8} = 0.75$
- quartet distance: 23; normalized $\frac{23}{35} \approx 0.66$

Expert trees

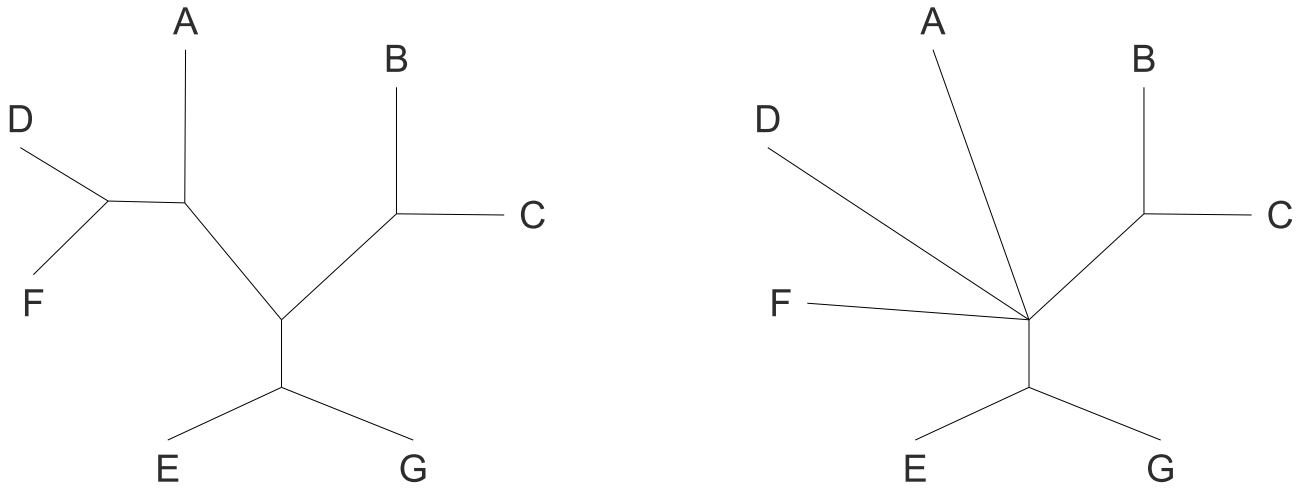
- quality of phylogenetic inference can be evaluated by comparison to **expert classifications**
- three commonly used classification systems:
  1. two-level taxonomy from WALS (World Atlas of Language Structure)
  2. multi-level taxonomy from Ethnologue
  3. more conservative multi-level taxonomy according to Harald Hammarström
- all three are part of the meta-data in ASJP

Expert trees and tree distances

- most nodes in the expert trees are multiple branching
- trees that are produced by phylogenetic software are always binary branching
- this leads to misleadingly high tree distances

Expert trees and tree distances

- suppose the left tree is extracted from the data and the right one is an expert tree



- as the left tree correctly captures all taxa in the right tree, this seems to be a perfect fit
- however:
  - normalized Robinson-Fould distance: 0.33
  - normalized quartet distance: 0.11

Expert trees and tree distances

- in practice
  - 5,644 languages in ASJP (excluding creoles etc.)
  - there 5,641 partitions in every inferred tree
  - Ethnologue: 1,803 partitions
  - WALS: 391 partitions
  - Hammarström: 1,735 partitions
- Robinson-Fould distance to WALS tree will be at least 0.68, no matter how well the algorithm performs
- minimum quartet distance: not easy to calculate, but also substantial

Measures of fit

- more realistic measures of goodness of fit:
  - **Robinson-Fould fit**:

$$\frac{\text{number of shared partitions}}{\text{total number of partitions in the expert tree}}$$

  - **quartet fit**:

$$\frac{\text{number of shared butterflies}}{\text{total number of butterflies in the expert tree}}$$

- these measures are always between 0 and 1
- 1 means that all groupings from the expert classification are correctly recovered

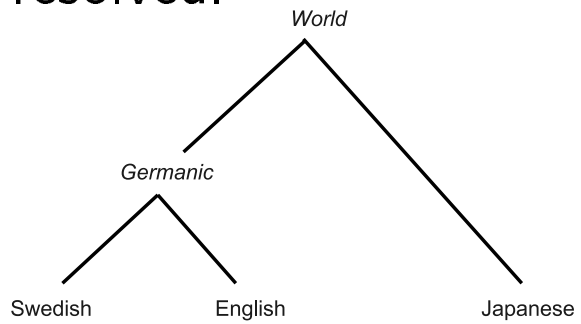
Triplet fit

- pick a triplet of languages A, B, C which has a resolved tree structure $((A, B), C)$ according to the expert tree
- determine predicted distances:

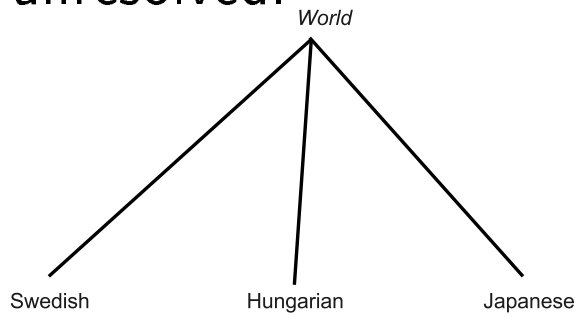
$$\begin{aligned} d(\text{Swedish}, \text{English}) &= 0.486 \\ d(\text{Swedish}, \text{Japanese}) &= 0.905 \\ d(\text{English}, \text{Japanese}) &= 0.897 \end{aligned}$$

- - $d(A, B) < \min(d(A, C), d(B, C)) \mapsto$ **correct**
  - otherwise $\mapsto$ **incorrect**
- triplet fit of a distance measure to an expert tree: proportion of resolved triplets that come out correct

resolved:

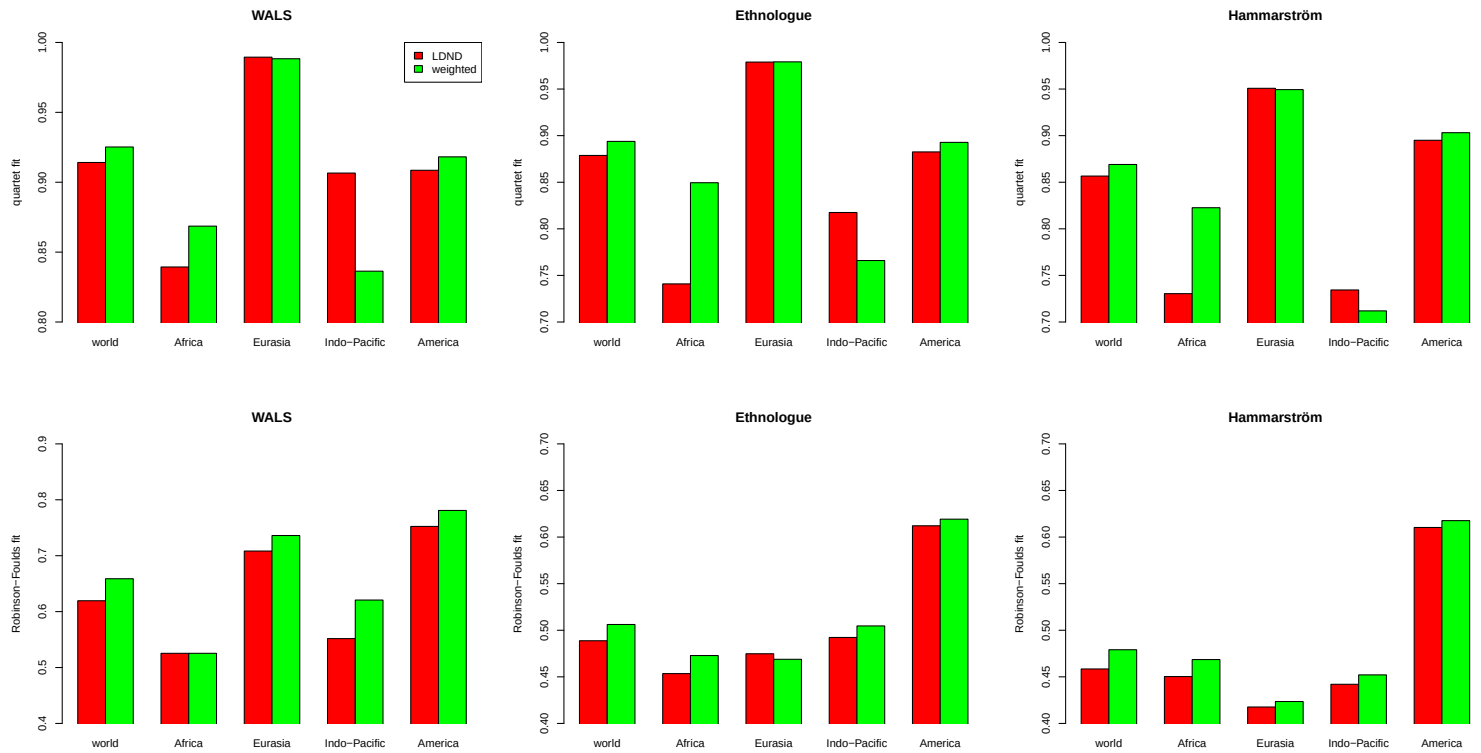


unresolved:



Evaluation

Neighbor-Joining algorith; split of data into four continental areas

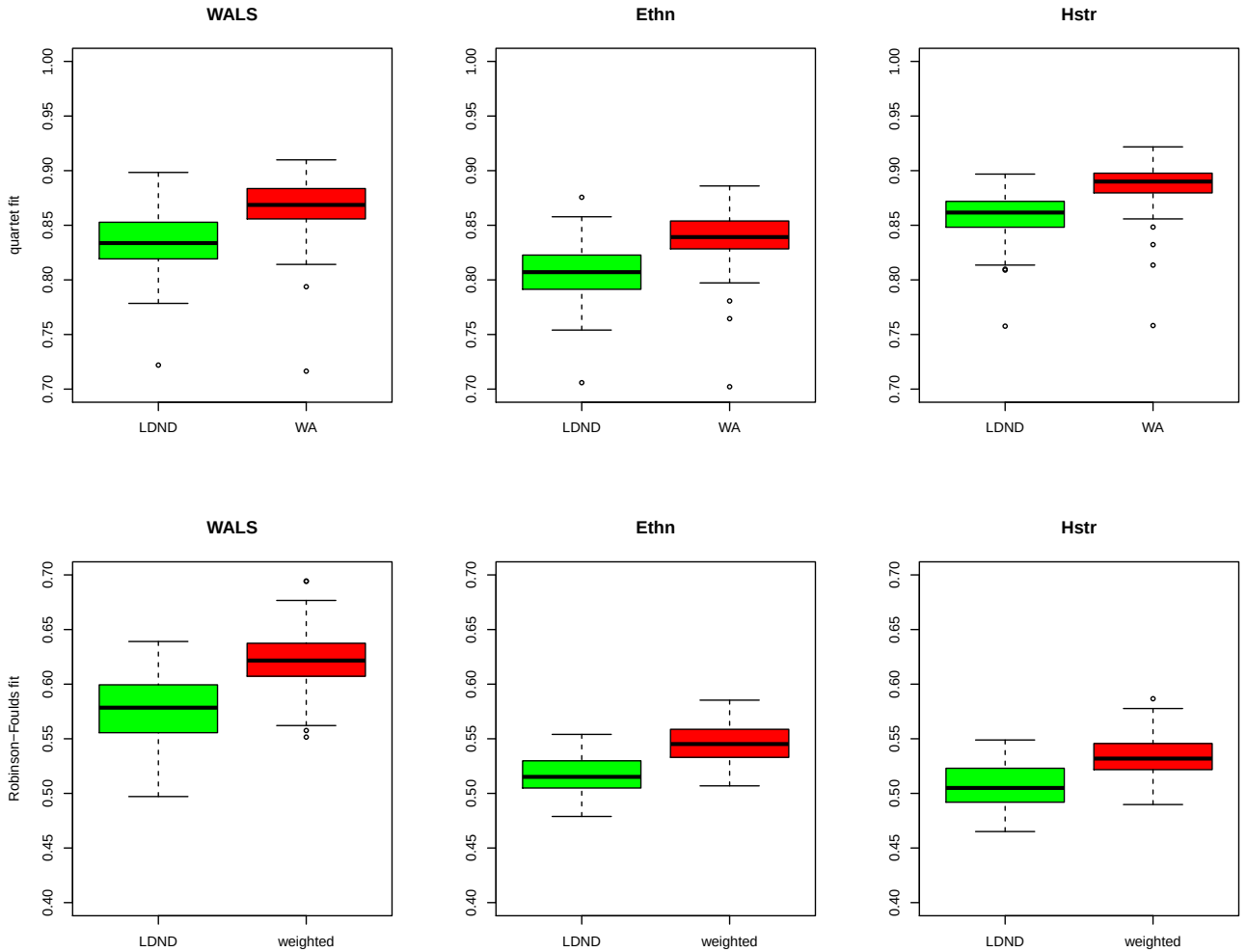


weighted alignment seems to perform slightly better, but results are somewhat inconclusive

Evaluation

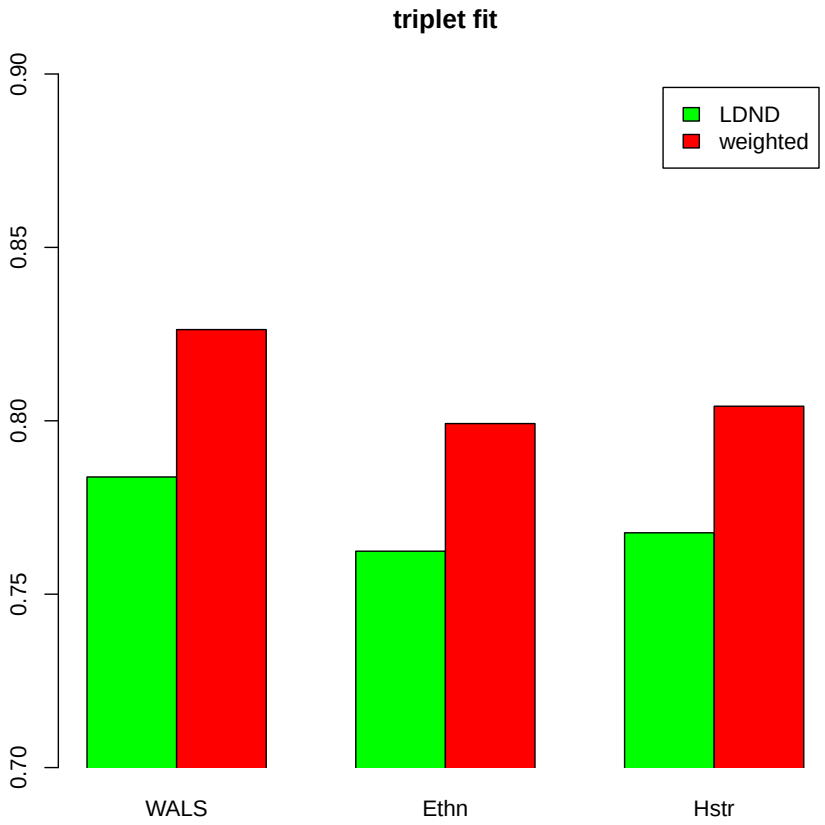
- alternative procedure:
  - draw 1,000 word lists at random from ASJP and compute quartet fit and Robinson-Foulds fit to the expert trees
  - repeat this procedure 100 times

Evaluation


WALS
Ethn
Hstr
quartet fit
Robinson−Foulds fit
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
LDND
WA
weighted

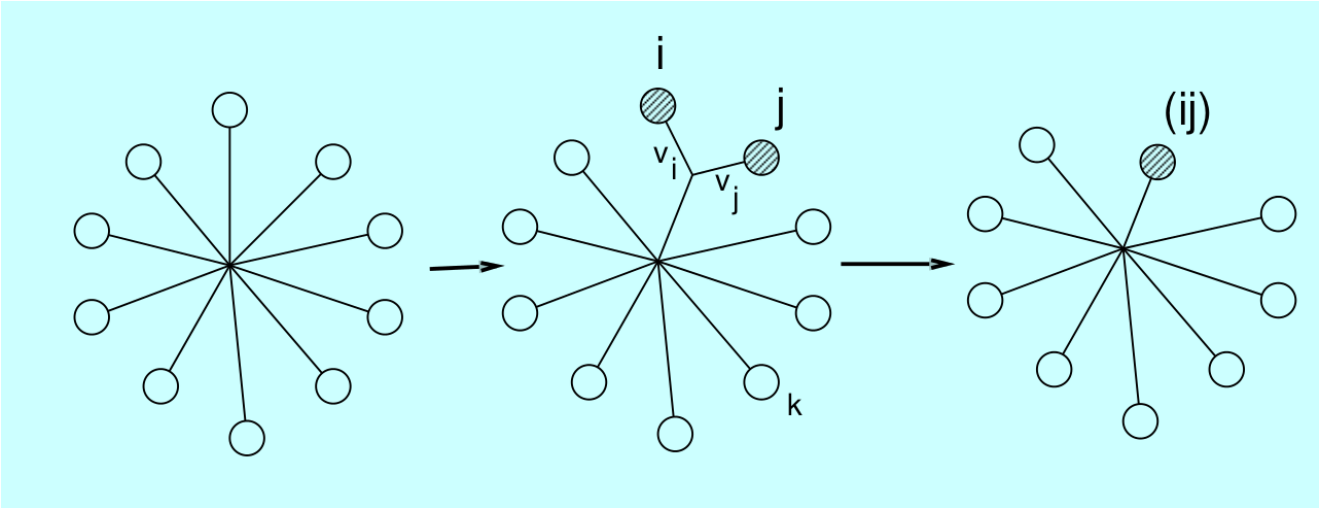
Evaluation

- triplet fit (based on a random sample of 10,000 resolved triplets)



Phylogenetic inference

- standard method for computing an (unrooted) tree from pairwise distances: **Neighbor Joining** (NJ)

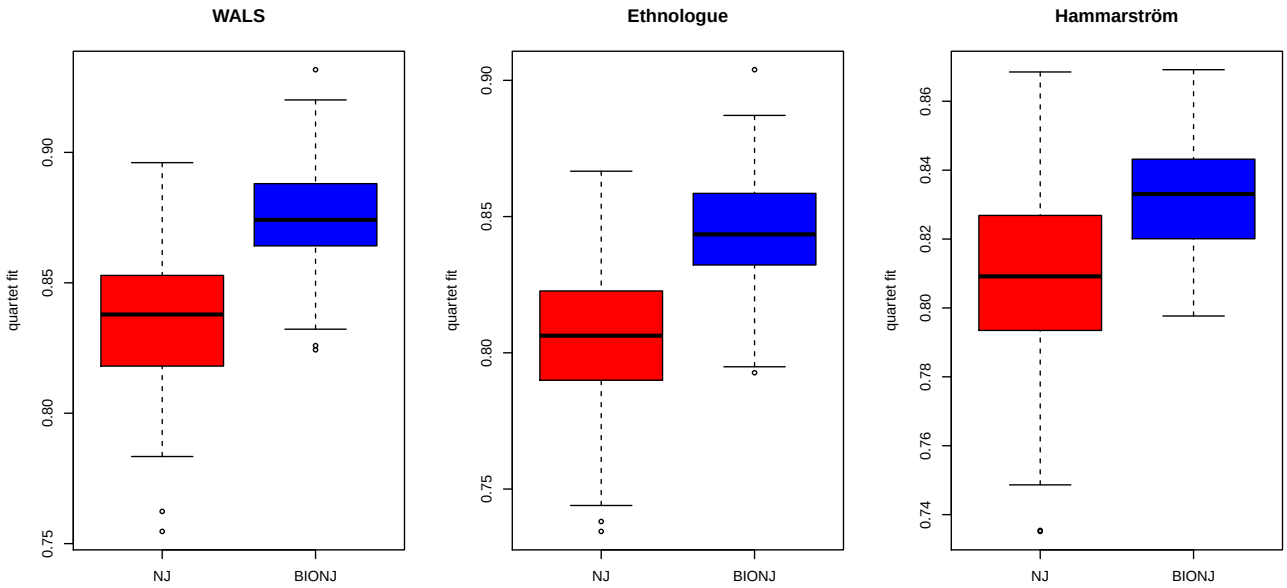


- criterion for choice of (i, j):

$$\min_{ij}[(N-2)d(i,j) - \sum_{k=1}^{N} d(i,k) - \sum_{k=1}^{N} d(j,k)]$$

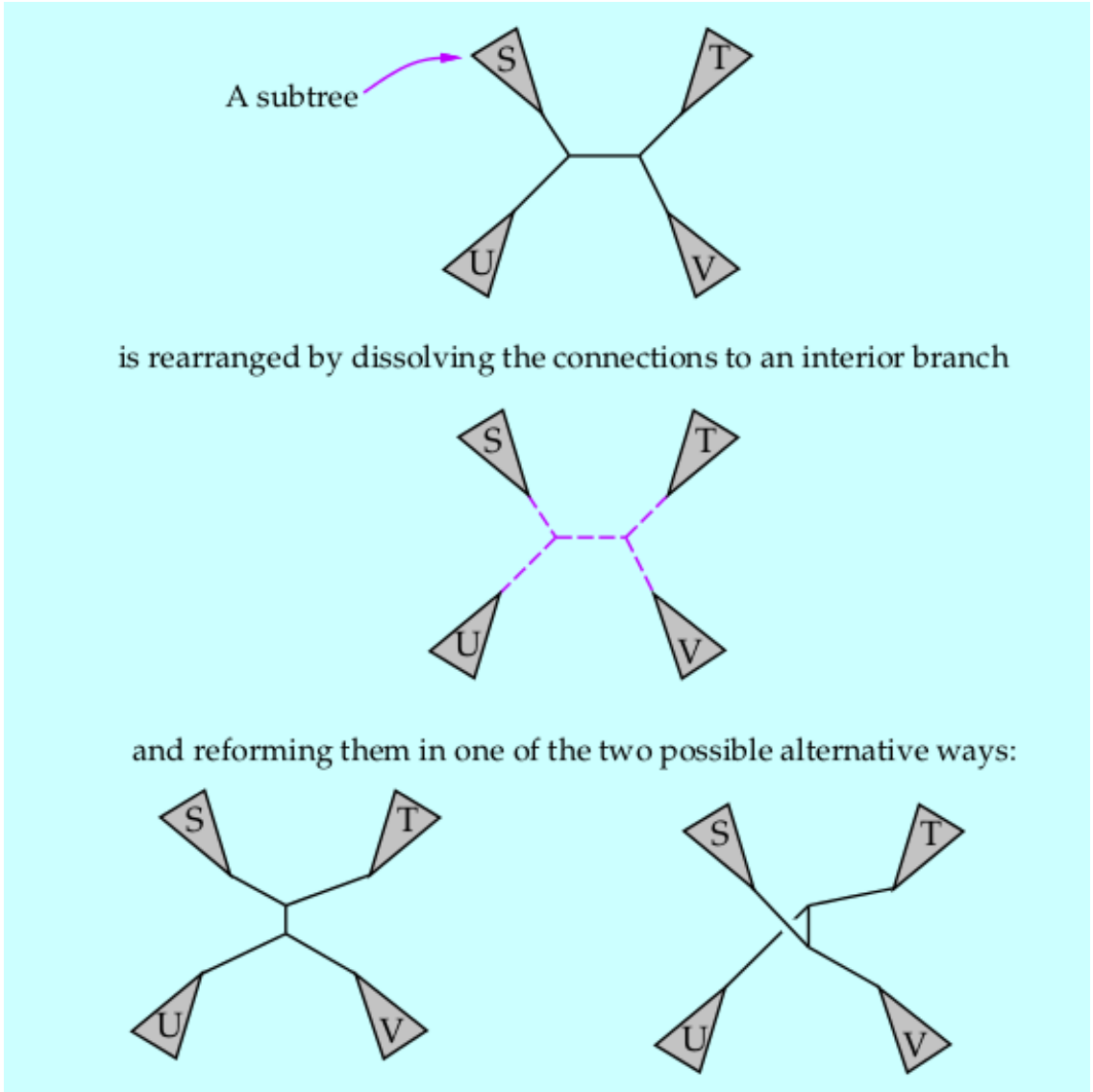
Phylogenetic inference

- alternative: BIONJ (Gascuel 1997)
- only difference to NJ: (i, j) are chosen so that the variance of the reduced distance matrix is minimized
- comparison NJ vs. BIONJ
  - evaluation over 1,000 random samples, each consisting of 1,000 word lists:

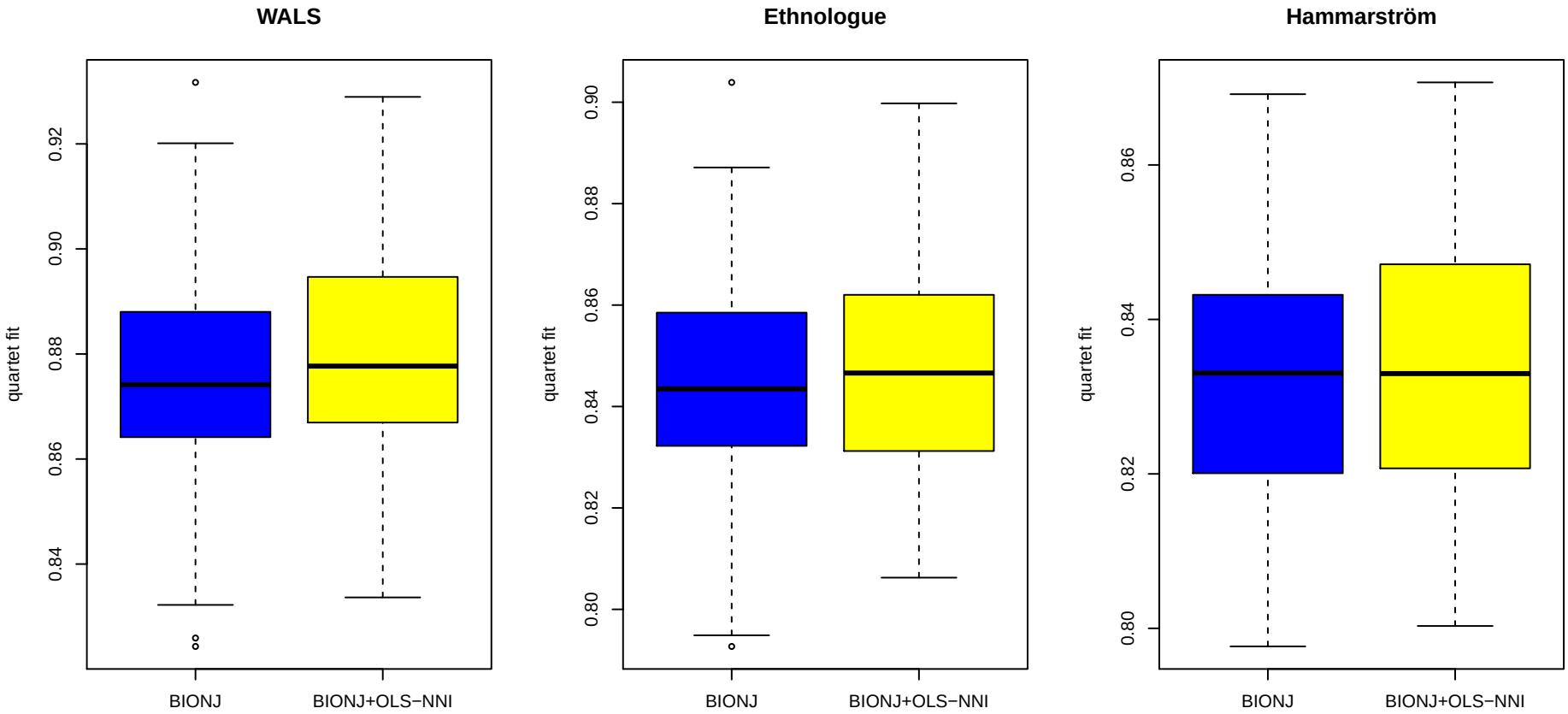


Optimization via Nearest Neighbor Interchange

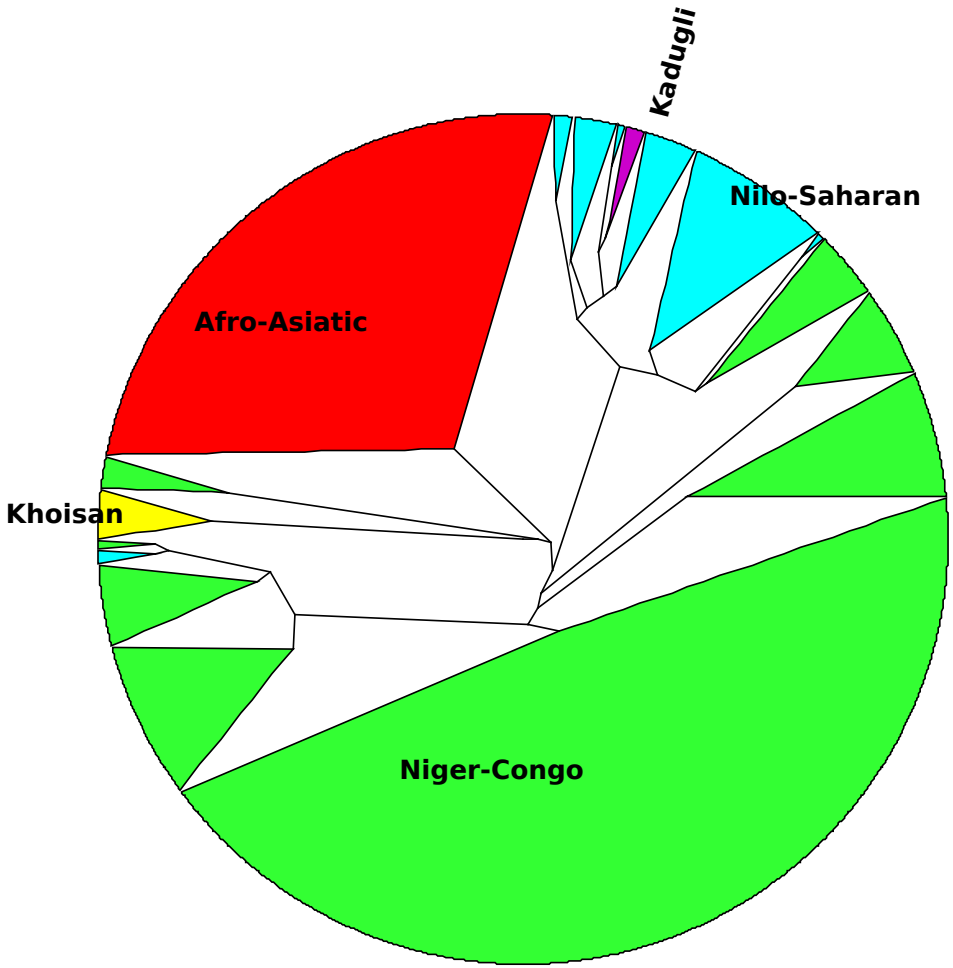
- criterion for optimization: *Ordinary Least Square*
- implemented in *FastME* software package



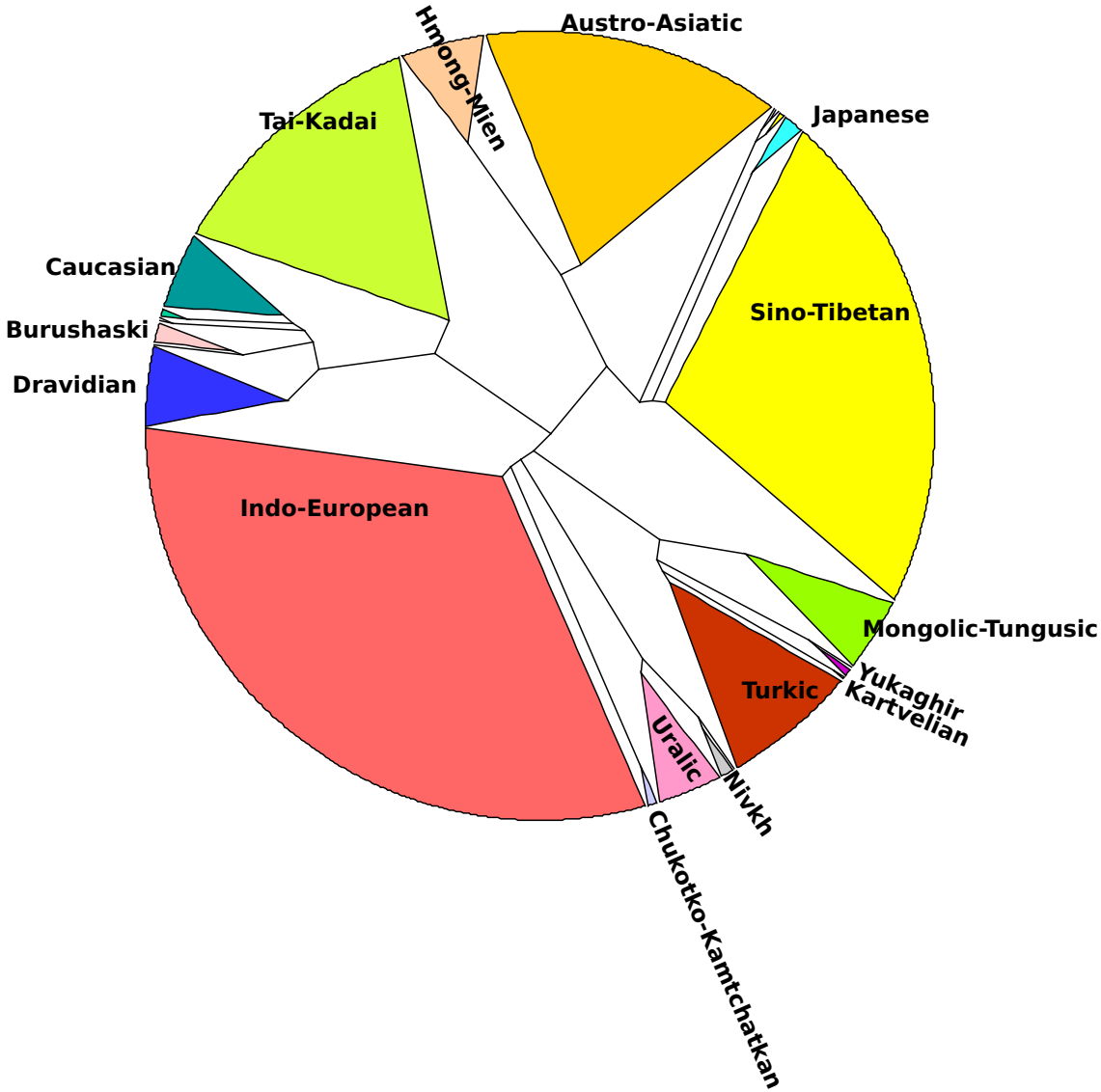
Optimization via Nearest Neighbor Interchange



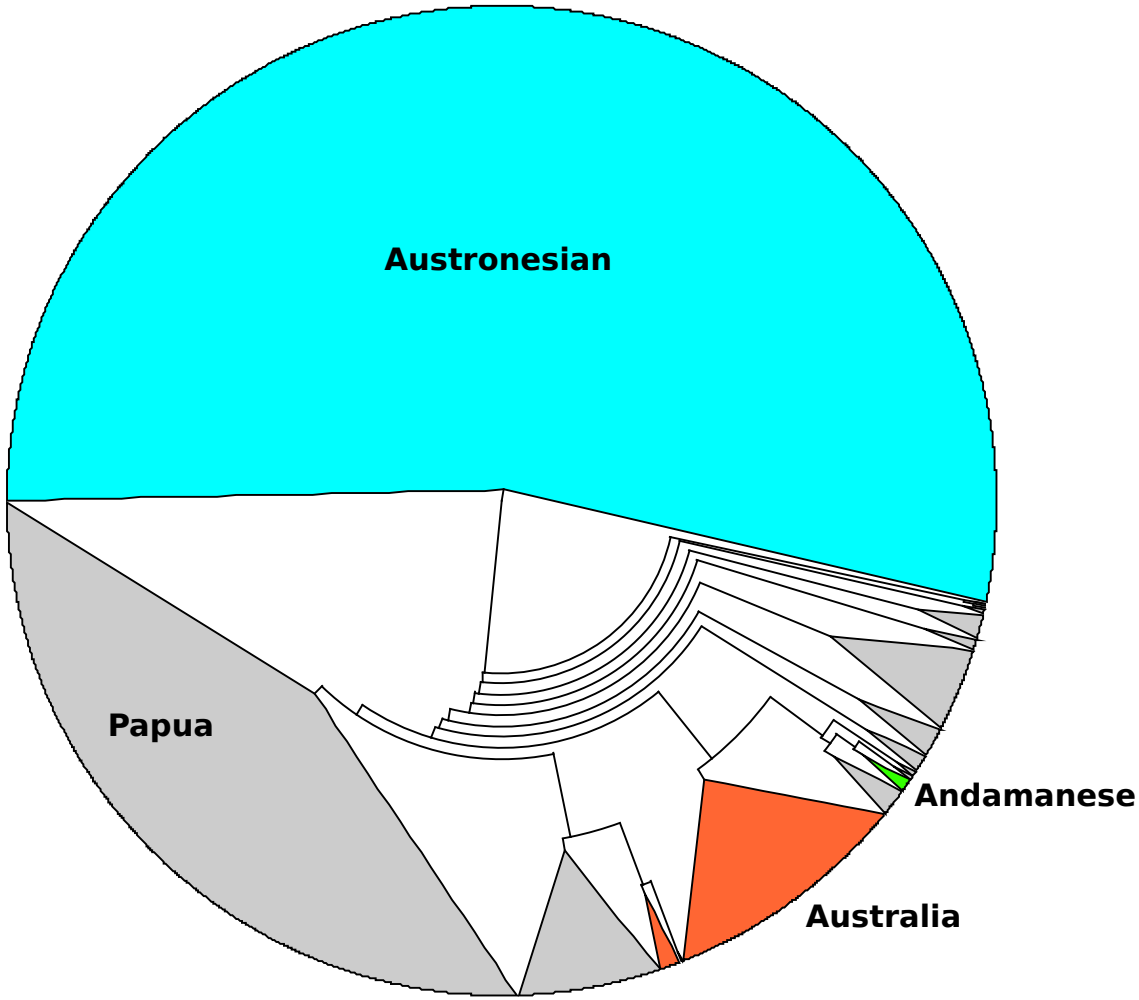
Results: Africa



Results: Eurasia



Results: Indo-Pacific



Results: America