Estimating and Visualizing Language Similarities Using ~~Weighted~~ Alignment and Force-Directed Graph Layout

Gerhard Jäger, Armin Buch, David Erschler and Andrei Lupas

University of Tübingen
MPI for Developmental Biology Tübingen

September 1, 2012, Stockholm

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

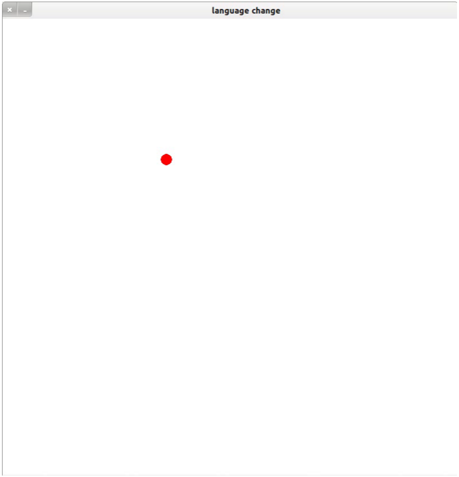
CLANS

- **Cl**uster **An**alysis of **S**equences (Frickey and Lupas 2004)
- Visualization of similarity matrices using *Force Directed Graph Layout*
- advantages in comparison to tree-based algorithms:
  - does not *a priori* assume a tree like signal (useful when lateral transfer plays a role)
  - fast (esp. in comparison to character based algorithms)
  - robust (noise in data items does not accumulate)
- general impression so far (Lupas, p.c.):
  - tree algorithms are more precise when evolutionary distances are small; CLANS is more sensitive to weak evolutionary signals

Is this true?

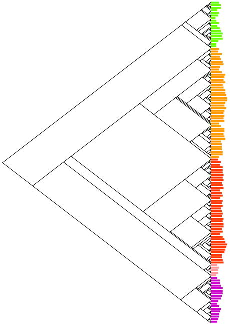
simLC

- simulation of language change:
  - "languages" are represented as vectors of identifiers (cognate classes, if you like)
  - languages are located on a two-dimensional surface
  - in each time step, each living language
    - moves a bit around in space
    - may replace words by some new, unrelated words
    - may borrow words from geographically neighboring languages
    - may split into two languages, and
    - may go extinct

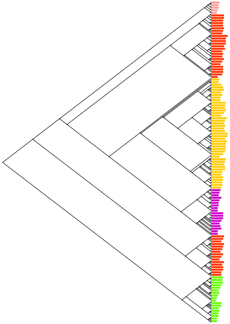

language change

Analyzing simulated data

True phylogenetic tree



Reconstructed tree (using neighbor joining)



Analyzing simulated data

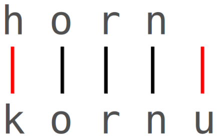
The Automated Similarity Judgment Program

- 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*

First shot: Levenshtein Distance

- first step: finde minmal edit distance between all translation pairs of the languages to be compared
- e.g. German ↔ Latin

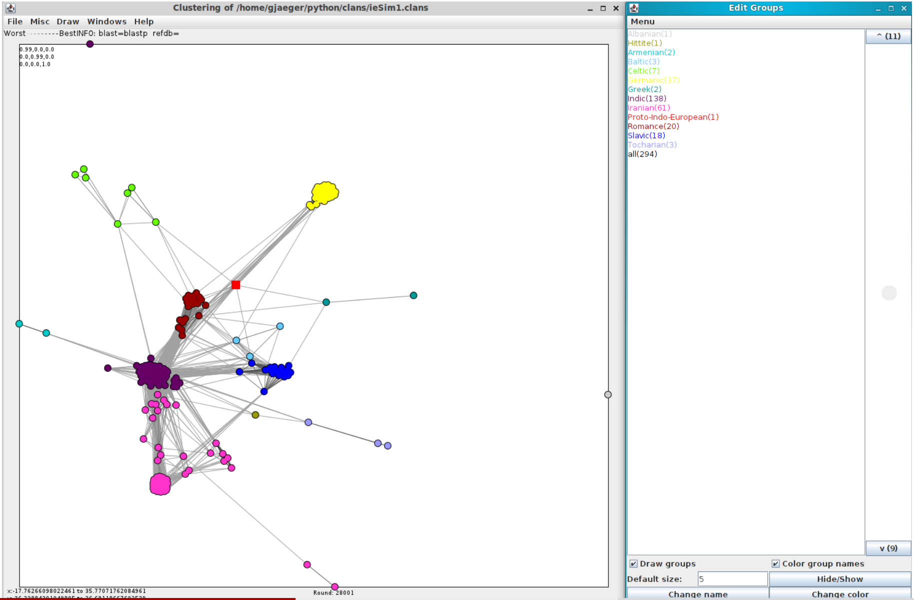


- edit distance = 2
- transformation into similarity measure

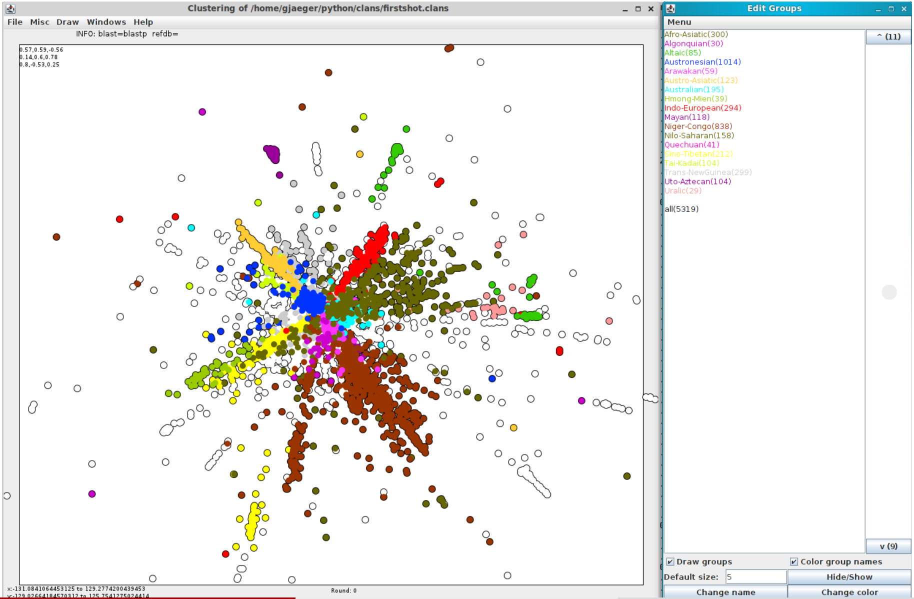
$$\mathrm{sim}(x, y) \doteq \frac{2(\max(l(x), l(y)) - d_{Lev}(x, y))}{l(x) + l(y)}$$

- similarity between L1 and L2: average similarity of translation pairs between L1 and L2

First shot: normalized Levenshtein Distance

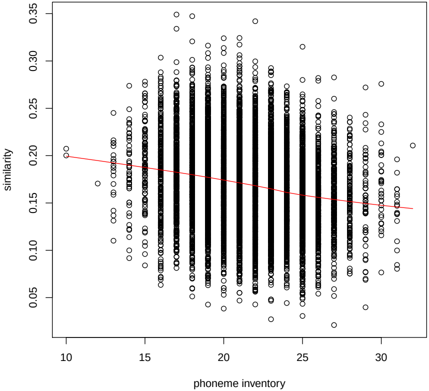


First shot: normalized Levenshtein Distance



First shot: normalized Levenshtein Distance

- basic problem here: the smaller the sound inventories of the languages compared, the higher is the probability of false positives



Benchmark: LDND measure

- Wichmann et al.: doubly normalized Levenshtein distance (**L**evenshtein **D**istance **N**ormalized and **D**ivided)
- normalization for word length

$$\mathrm{ldn}(x,y) \doteq \frac{d_{Lev}(x,y)}{\max(l(x),l(y))} \tag{1}$$

- normalization for language specific patterns (including sound inventory size):
  - normalization factor $1/\mu$
  - μ_{L_1,L_2}: mean of $\{ldn(x,y) | x \in L_1, y \in L_1, \|x\| \neq \|y\|\}$

$$\mathrm{ldnd}(x,y,L_1,L_2) \doteq \frac{ldn(x,y)}{\mu_{L_1,L_2}}$$

$$\mathrm{ldnd}(L_1,L_1) \doteq \frac{\sum_{x\in L_1, y\in L_2}\{\mathrm{ldnd}(x,y,L_1,L_2) : \|x\| = \|y\|\}}{\#\{x,y : \|x\| = \|y\|\}}$$

Benchmark: LDND measure

English / Swedish

	Ei	yu	wi	w3n	tu	fiS	...
yog	1	$^2/_3$	1	1	1	1	
du	1	$^1/_2$	1	1	$^1/_2$	1	
vi	$^1/_2$	1	$^1/_2$	1	1	$^2/_3$	
et	1	1	1	1	1	1	
tvo	1	1	1	1	$^2/_3$	1	
fisk	$^3/_4$	1	$^3/_4$	1	1	$^1/_2$	
⋮							

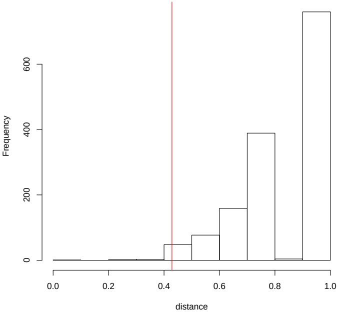
- average LDN along diagonal: 0.56
- average LDN off diagonal: 0.91
- LDND: $^{0.56}/_{0.91} = 0.61$

A bit of information theory

Swedish *fisk* = English *fish*?
Turkish *dört* = English *dirt*?

- first guess is good because the words sound similars **and the languages are closely related**
- second guess is bad (and wrong) even though the words sound similar **because the languages are not related**
- If two languages are related, knowing a word from one language reduces the uncertainty about its form in the other language
- *Hypothesis: degree of similarity between two languages* ≈ *average amount of information that the form of a word in one language carries about the form of its translation into the other language*

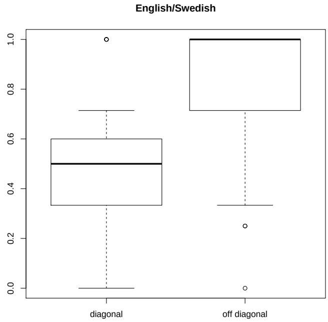
English and Swedish again



- Histogramm: off-diagonal distances
- red line: distance *fiS* $\sim$ *fisk* $(= 4.3)$
- relative frequency of off-diagonal entries ≤ 4.3: 0.004
- can be interpreted as p-value for the null hypothesis that the two words are not cognates
- $-log_2(0.004) = 7.9$ bit: amout of information that [fisk] carries about [fiS], given the general pattern of phonotactic similarities between unrelated English and Swedish words

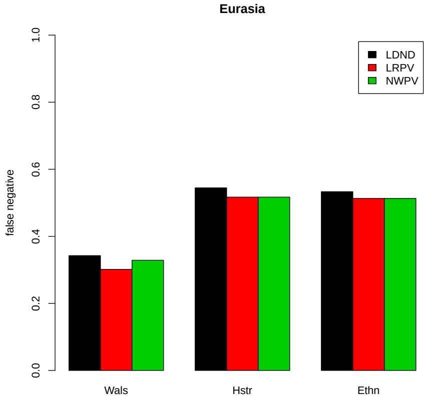
Information theoretic estimate of language similarity

- similarity between two languages: average amount of information that a word from one language carries about its translation
- formally: average binary logarithm of the p-values for all Swadesh items in the date base



Benchmarking

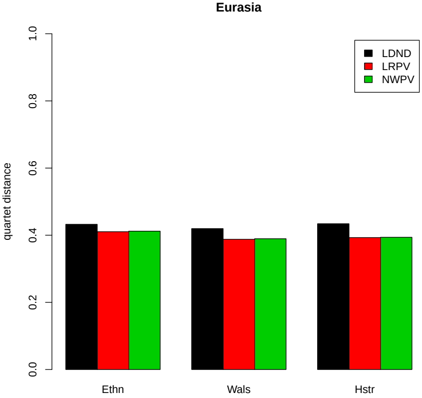
- comparison of Neighbor-Joining tree with three expert classifications:
  - WALS
  - Ethnologue
  - Hammarström 2010
- measure: proportion of false negatives, i.e. clades in the expert tree that are not recognized in the automatically obtained tree



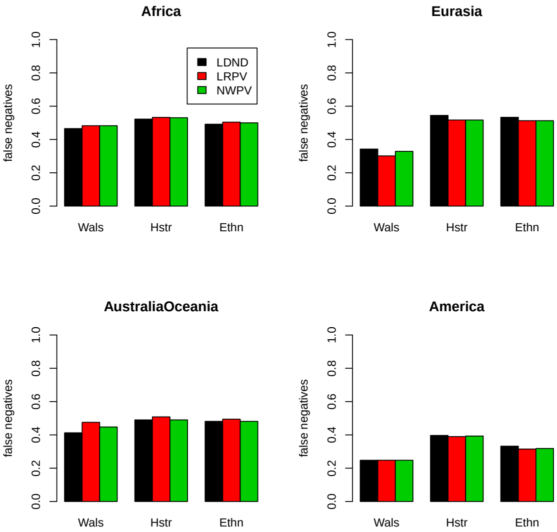
Benchmarking

- same, but using **quartet distance**:
  - all quadruples of languages are considered
  - there are three ways how a quadruple can be organized into an unrooted binary tree
  - *quartet distance* counts the proportion of quadruples where the expert tree assumes another structure than the induced tree

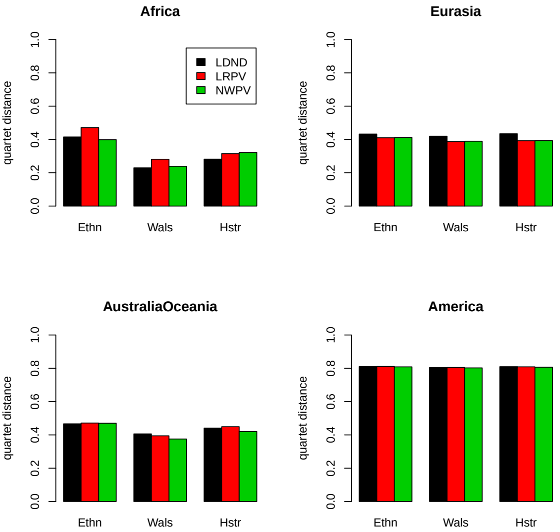
LRPV seems to fare pretty well, but...



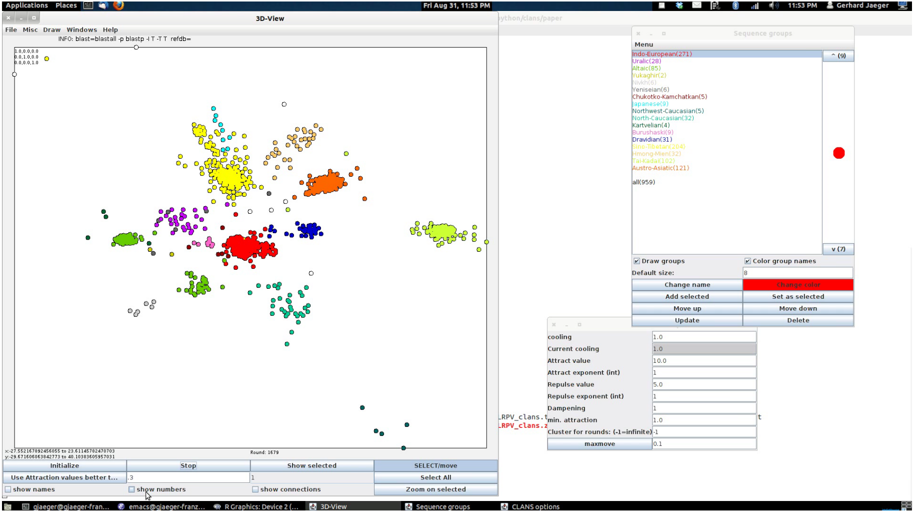
Benchmarking



Benchmarking



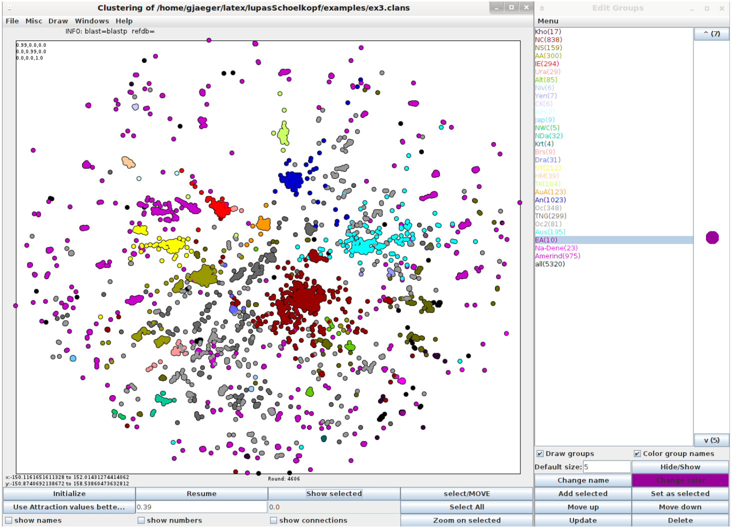
Visualization with CLANS



Qualitative results

- several stable recurring patterns:
  - primary division of Eurasia into
    - Sino-Tibetan/Austro-Asiatic/Japanese/Hmong-Mien (/Tai-Kadai)
    - Indo-European/Altaic/Uralic/Caucasian languages/Siberian languages/Chukotko-Kamchatkan
    - Dravidian languages always end up in the center, right between the two major groups
  - Sino-Tibetan/Japanese
  - Indo-European/Burushaski
  - Turkic/Uralic
  - Tai-Kadai/Austronesian
  - Nilo-Saharan/Niger-Congo

The world is not enough

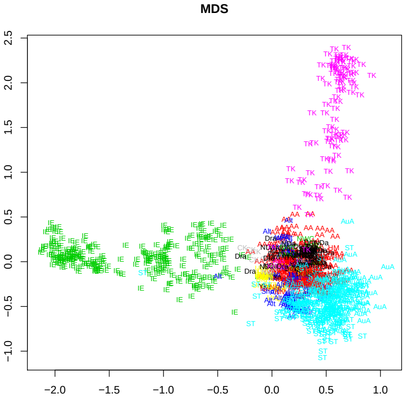


CLANS and dimensionality reduction

- CLANS performs a kind of (non-deterministic) dimensionality reduction
- How does this relate to more established methods?

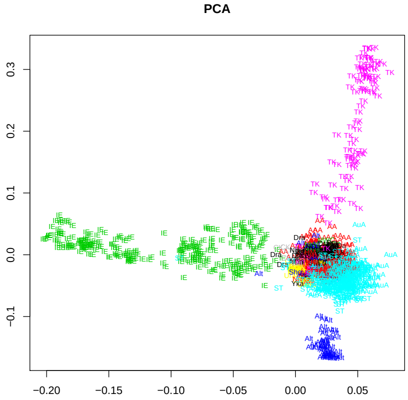
CLANS vs. Multi-Dimensional Scaling

- MDS applied to NWPV-matrix of the Eurasian languages



CLANS vs. Principal Component Analysis

- PCA applied to NWPV-matrix of the Eurasian languages



CLANS and dimensionality reduction

- language families massively vary in size
- MDS and PCA only provide information about the largest families
- CLANS is sensitive to local patterns

Frickey, T. and A. N. Lupas (2004). Clans: a java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**(18):3702–3704.