Can we build an accurate language phylogenetic tree with (*just*) 10 words?

Lucas Lacasa[1], Niko Komin[1], Andrew Berdahl[2]

[1] IFISC

[2] Dept. Ecology and Evol. Biology, Princeton Univ.

*IFISC

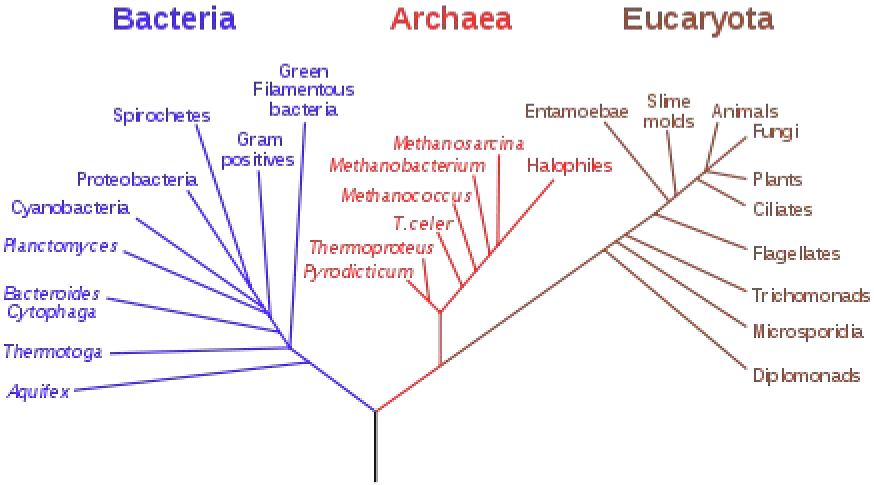
What's a phylogenetic tree ?

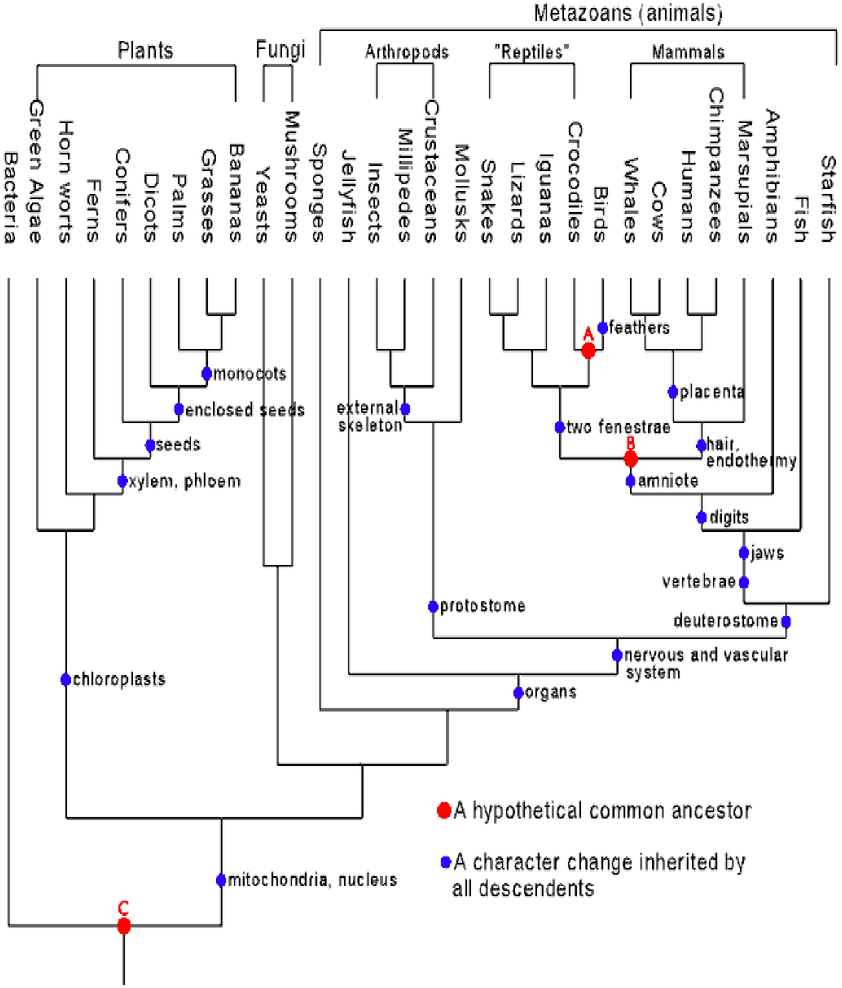
Tree showing the evolutionary relationships among entities that are shown to have a common ancestor.

Examples:

- Biological species (molecular data, DNA)
- Languages (texts)

Phylogenetic Tree of Life





Build a phylogenetic tree in 4 steps

1. Defining the ENTITY

Biological species, DNA, RNA, human languages, programming languages??

2. Defining the DATA: essentially two possibilities

- TEXT in several languages (e.g. Human rights declaration)
- list of Keywords (Swadesh list >100 words) in several languages

3. Defining the DISTANCE between data... many ones!

- Pairwise: Levenshtein distance, Sequence allignment methods
- Global: Kolmogorov Complexity (Zip) *ex: Benedetto et al, PRL 88 (2002)*

4. Define the ALGORITHM

- Unweighted Pair Group Method Average (UPGMA) (absence of implicit evolution model)
- Maximal Parsimony (with implicit evolution model)
- ...

SO, our motivation is

Which is the smallest amount of information such that the tree is OK?

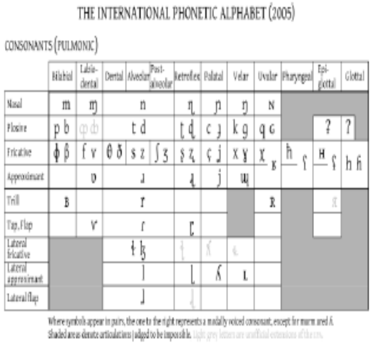
And we think that

Numbers are fundamental elements within a language.

ENTITY: languages
DATA: numbers from one to ten in several languages
DISTANCE: Alphabet-Codon mapping + Global sequence allignment
ALGORITHM: Matlab...

{uno,dos,tres,cuatro,cinco,seis,siete,ocho,nueve,diez}
{one,two,three,four,five,six,seven,eight,nine,ten}
{un,deux,trois,quatre,cinq,six,sept,huit,neuf,dix}
...

Alphabet mapping: criteria



THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ	n			ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ	t d			ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ	ɹ			ɻ	j	ɰ				
Trill	ʙ		r						ʀ		ᴙ	
Tap, Flap		ⱱ	ɾ			ɽ						
Lateral fricative			ɬ ɮ			ꞎ	ʎ̝	ʟ̝				
Lateral approximant			l			ɭ	ʎ	ʟ				
Lateral flap			ɺ			ɭ̆						

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ɦ. Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.

- Each letter maps into a 3-nucleotide string from {A,T,C,G}
- Phonetic and feature-based properties are encoded in the mapping
- We finally have a new alphabet: each of the 26 letters is a 3-nucleotide string
- We concatenate the numbers in a single string
- We make global sequence allignment

The mapping recipe

A	AAA
E	ACA
I	AGA
O	ATA
U	ATC
Y	AGG

B	CGT
P	CGA
V	CGG
F	CAG
W	AGG

C	GTA
D	GAA
T	GTT
Z	GTG

S	ATG
X	TTG
H	TTT
L	CCC
M	CTA
N	CTC
R	AGC

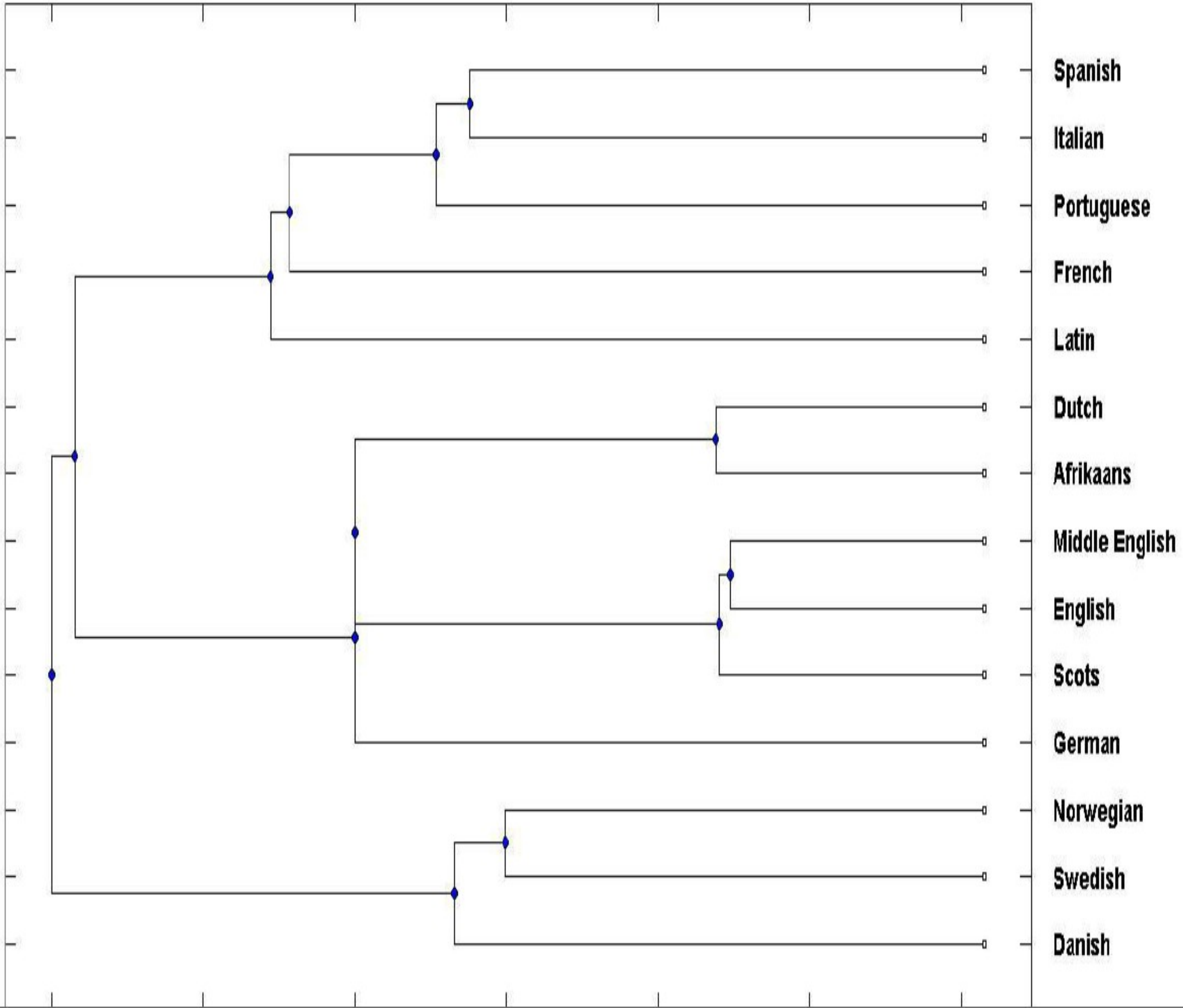
Example of sequence alignment

6 (B)

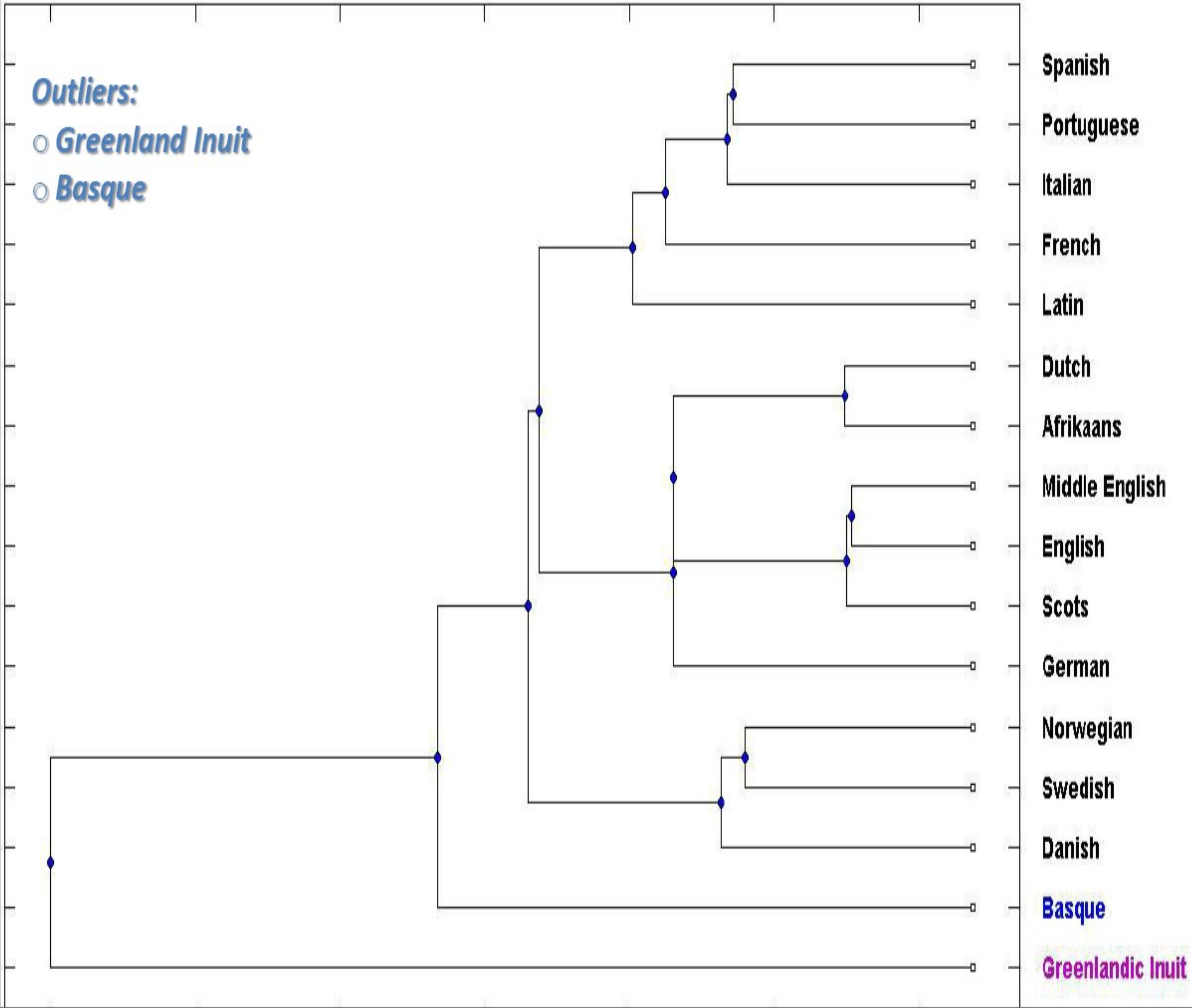
```
GAATTCAG      GAATTCAG
| | || |      | || | |
GGA-TC-G      GCAT-C-G

GAATTC-A      GAATTC-A
| | || |      | || | |
GGA-TCGA      GCAT-CGA
```

Results (I): subset of 'familiar' languages: *WORKS!*



Results (II): detecting the outliers: ***WORKS!***



Outliers:

- ***Greenland Inuit***
- ***Basque***