UPGMA and CMM

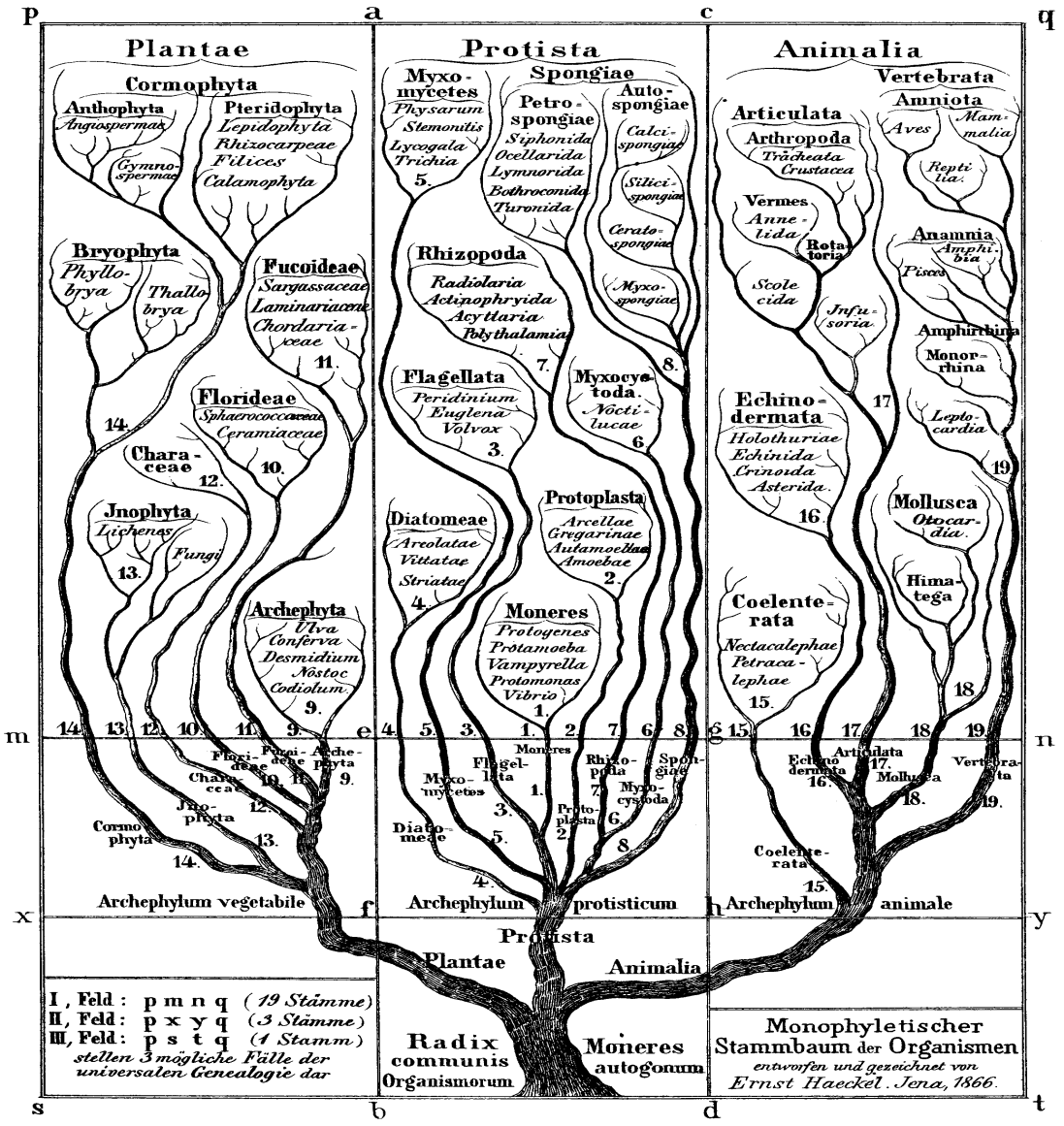
Peter Revesz

CSCE 915
Computer Science and Engineering
University of Nebraska-Lincoln

Evolutionary trees

The earliest evolutionary trees were constructed manually by biologists based on morphological similarities. This works reasonably for a few species, but is not good for a large number of species or for individuals within the same or closely related species.

Today computer algorithms can construct evolutionary trees based on genetic similarities. Thousands of species or individuals can be placed into an evolutionary tree.



(Ernst Haeckel 1866)

From DNA Strings to Distance Matrices

DNA String			
S_1	AGCTA	CTAGT	AATCA
S_2	AGCTA	CGAGT	AATCA
S_3	ATCCA	CTAGT	ACACT
S_4	ATCCA	CTAGT	ATACT
S_5	CGGTA	TTTGT	AAGCT
S_6	CGGTT	CATCA	AATGC
S_7	AGGTA	CTTGA	AATCC

The Hamming distance δ between two DNA strings is the number of different nucleotides.

$$\delta(S1, S2) = 1$$

because $S_1[7]$ = T while $S_2[7]$ = G.

Distance matrix based on Hamming distance:

	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1	0	1	5	5	6	9	4
S_2	1	0	6	6	7	9	5
S_3	5	6	0	1	8	13	8
S_4	5	6	1	0	8	13	8
S_5	6	7	8	8	0	8	5
S_6	9	9	13	13	8	0	5
S_7	4	5	8	8	5	5	0

The UPGMA algorithm

- Unweighted Pair Group Method with Arithmetic mean
- Always combine the closest pairs, which is S1 and S2.
- After combination the distance from S12 to others with be the average of the distances from S1 and S2.

	S_{12}	S_3	S_4	S_5	S_6	S_7
S_{12}	0	5.5	5.5	6.5	9	4.5
S_3	5.5	0	1	8	13	8
S_4	5.5	1	0	8	13	8
S_5	6.5	8	8	0	8	5
S_6	9	13	13	8	0	5
S_7	4.5	8	8	5	5	0

After combining S3 and S4:

	S_{12}	S_{34}	S_5	S_6	S_7
S_{12}	0	5.5	6.5	9	4.5
S_{34}	5.5	0	8	13	8
S_5	6.5	8	0	8	5
S_6	9	13	8	0	5
S_7	4.5	8	5	5	0

After combining S12 and S7:

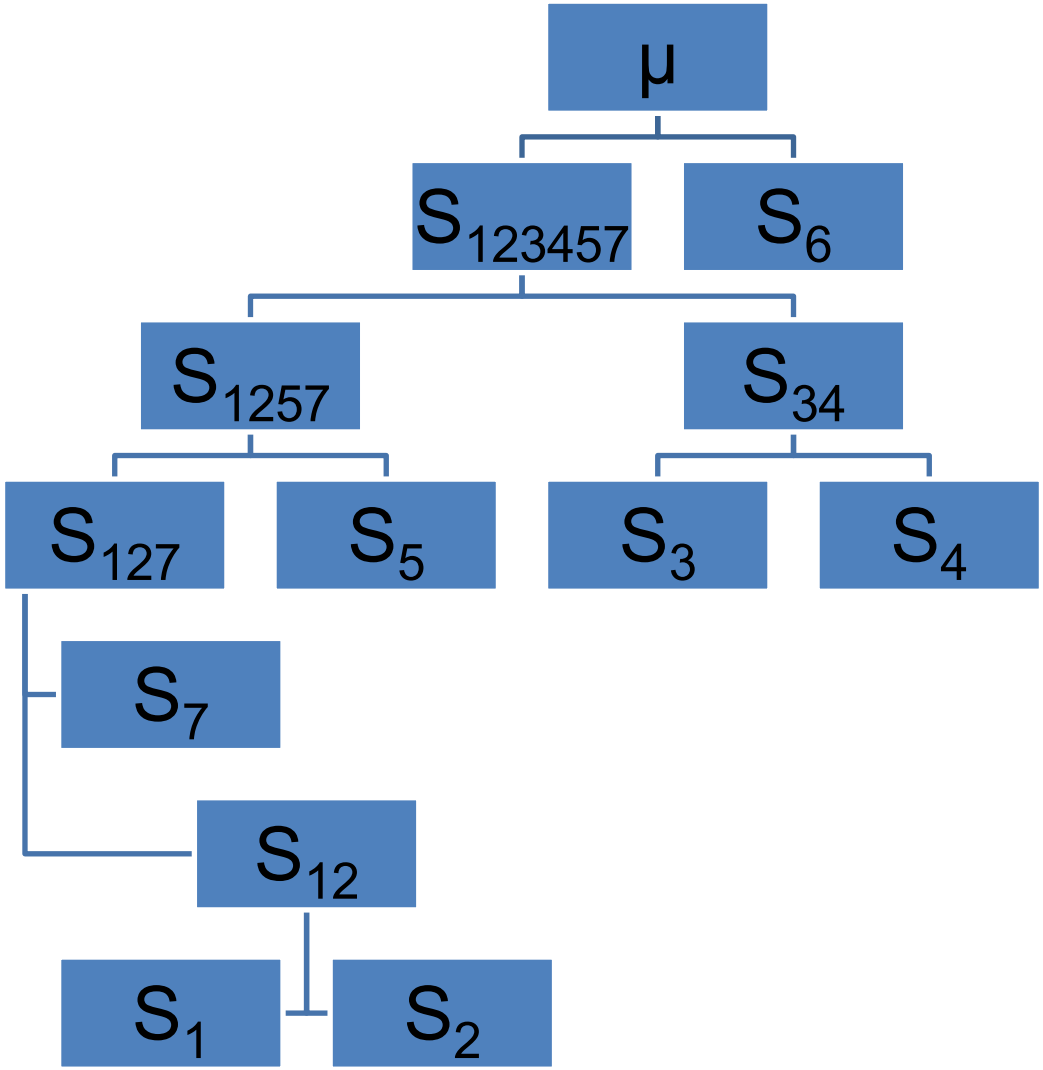
	S_{127}	S_{34}	S_5	S_6
S_{127}	0	6.75	5.75	7
S_{34}	6.75	0	8	13
S_5	5.75	8	0	8
S_6	7	13	8	0

After combining S127 and S5:

	S_{1257}	S_{34}	S_6
S_{1257}	0	7.375	7.5
S_{34}	7.375	0	13
S_6	7.5	13	0

After combining S1257 and S34:

	S_{123457}	S_6
S_{123457}	0	10.25
S_6	10.25	0



The evolutionary tree
Generated by UPGMA.

A new algorithm based on Common Mutation Matrixes.

- Common mutations suggest evolutionary similarities.
- In some cases common mutations may be more reliable evidence of an evolutionary relationship than similarities.

DNA String			
S_1	AGCTA	CTAGT	AATCA
S_2	AGCTA	CGAGT	AATCA
S_3	ATCCA	CTAGT	ACACT
S_4	ATCCA	CTAGT	ATACT
S_5	CGGTA	TTTGT	AAGCT
S_6	CGGTT	CATCA	AATGC
S_7	AGGTA	CTTGA	AATCC
μ	AGCTA	CTAGT	AATCT

Consensus sequence.

Sequence	Set of Mutations Δ
S_1	15A
S_2	7G, 15A
S_3	2T, 4C, 12C, 13A
S_4	2T, 4C, 12T, 13A
S_5	1C, 3G, 6T, 8T, 13G
S_6	1C, 3G, 5T, 7A, 8T, 9C, 10A, 14G, 15C
S_7	3G, 8T, 10A, 15C

Set of mutations.

Common mutation matrix.

	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1	15A	15A	Ø	Ø	Ø	Ø	Ø
S_2		7G, 15A	Ø	Ø	Ø	Ø	Ø
S_3			2T, 4C, 12C, 13A	2T, 4C, 13A	Ø	Ø	Ø
S_4				2T, 4C, 12T, 13A	Ø	Ø	Ø
S_5					1C, 3G, 6T, 8T, 13G	1C, 3G, 8T	3G, 8T
S_6						1C, 3G, 5T, 7A, 8T, 9C, 10A, 14G, 15C	3G, 8T, 10A, 15C
S_7							3G, 8T, 10A, 15C

	S_1	S_2	S_3	S_4	S_5	S_{67}
S_1	15A	15A	Ø	Ø	Ø	Ø
S_2		7G, 15A	Ø	Ø	Ø	Ø
S_3			2T, 4C, 12C, 13A	2T, 4C, 13A	Ø	Ø
S_4				2T, 4C, 12T, 13A	Ø	Ø
S_5					1C, 3G, 6T, 8T, 13G	3G, 8T
S_{67}						3G, 8T, 10A, 15C

Combine S6 and S7.

	S_1	S_2	S_{34}	S_5	S_{67}
S_1	15A	15A	Ø	Ø	Ø
S_2		7G, 15A	Ø	Ø	Ø
S_{34}			2T, 4C, 13A	Ø	Ø
S_5				1C, 3G, 6T, 8T, 13G	3G, 8T
S_{67}					3G, 8T, 10A, 15C

Combine S3 and S4.

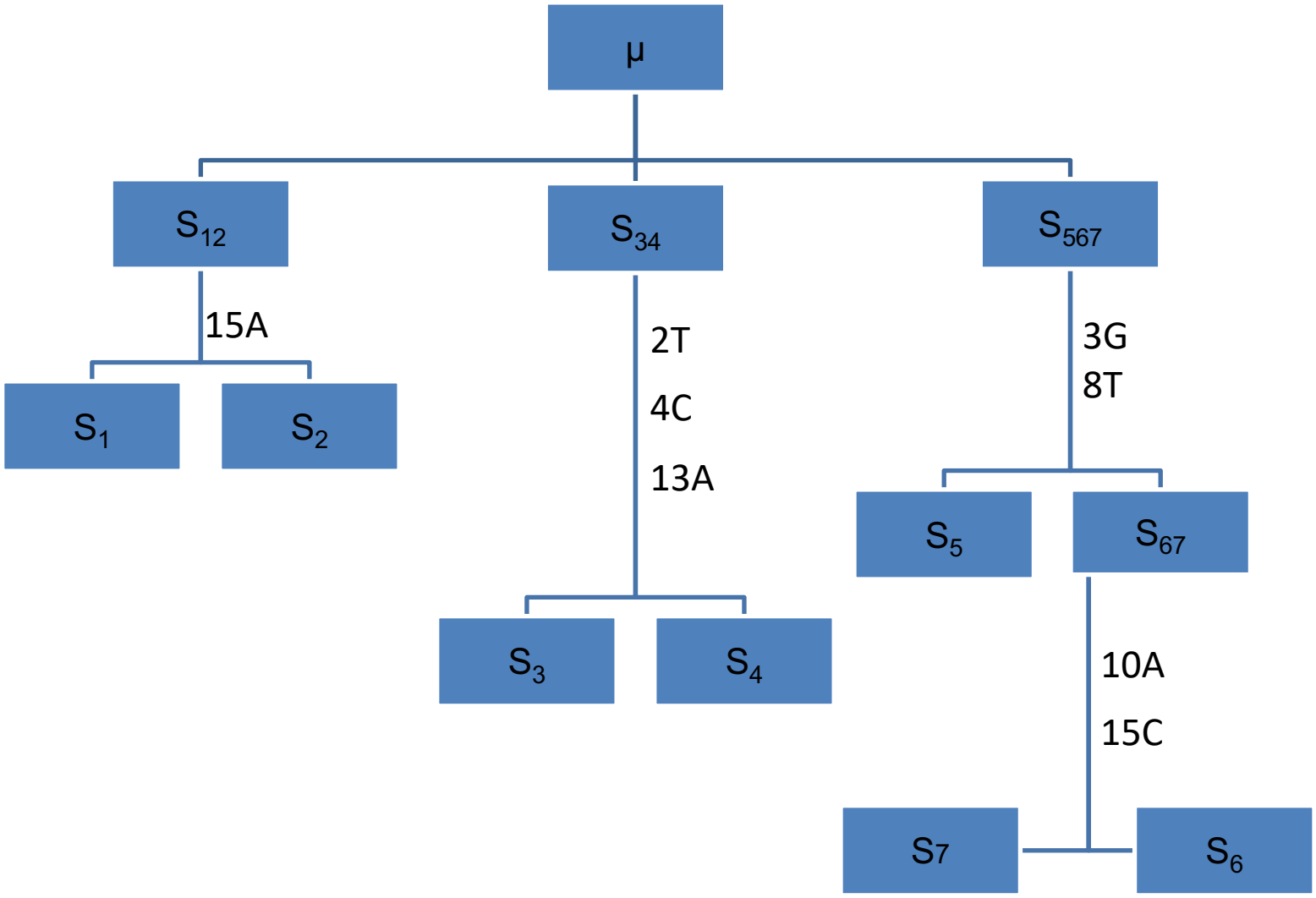
	S_1	S_2	S_{34}	S_{567}
S_1	15A	15A	Ø	Ø
S_2		7G, 15A	Ø	Ø
S_{34}			2T, 4C, 13A	Ø
S_{567}				3G, 8T

Combine S_5 and S_{67}.

	S_{12}	S_{34}	S_{567}
S_{12}	15A	Ø	Ø
S_{34}		2T, 4C, 13A	Ø
S_{567}			3G, 8T

Combine S_1 and S_2.

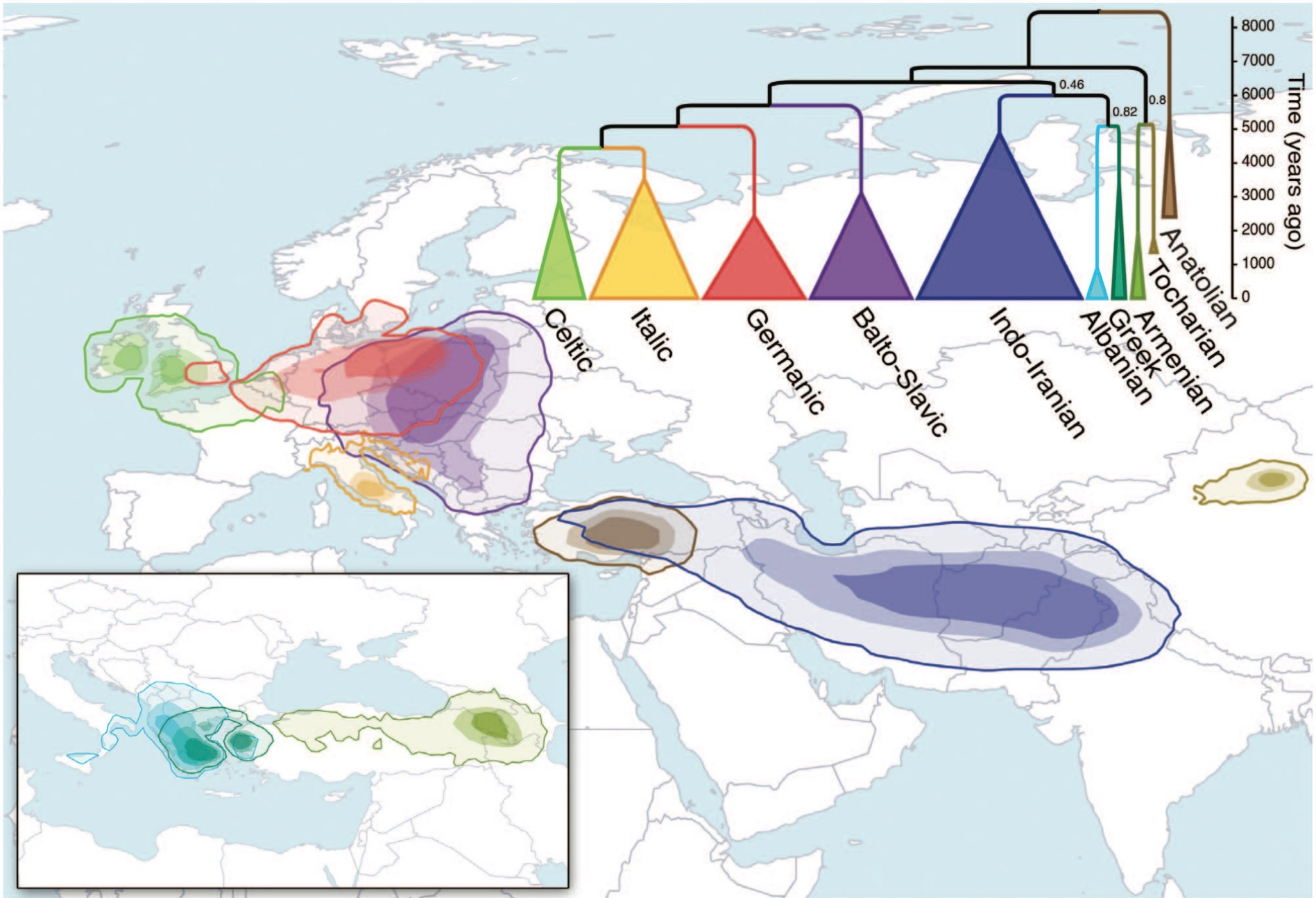
The evolutionary tree generated by the CMM algorithm



Comparison of the UPGMA and the CMM algorithms

- The two evolutionary trees are different.
- CMM is better in dealing with long distance branches.
- Two long distance branch may look more similar to a small branch which does not contain many mutations than to each other even though these have common mutations.
- Hence CMM is more robust in these cases.
- CMM complexity is $O(n^2 m)$, where n is number of input sequences and m is the length of the sequences.

Can UPGMA and CMM be applied to language families?



Swadesh list for Germanic languages (detail)

English	Scots	West Frisian	Dutch	Afrikaans	Low Saxon	Limburgish	Central Franconian	Luxembourgish	German
I	A	ik	ik	ek	ik	ich	ich, eich, ech	ech	ich
you (singular) thou (dialectal, literary, or archaic)	thoo, you, ye	do (dû) (informal), jo (formal)	jij, je (informal), u (formal)	jy (informal), u (formal)	du	doe (informal), geer (formal)	du, de	du, de	du
he	he	hy, er	hij	hy	he	hae	hä, hän, er, e	hien, en	er
we	we	wy	wij, we	ons	wi	weer	mir, mer	mir, mer	wir
you (plural)	you(se), ye(se)	jimme, me	jullie (informal), u (formal)	julle (informal), u (formal)	ji	geer	ühr, ihr, er, dir, der	dir, der	ihr
they	thay	sy, hja	zij, ze	hulle	se	die	sei, sie, se	si, se	sie
this	this	dit, dizze	deze, dit	dié, hierdie	düsse, düt	dit	disse, diss, diss/ditt	dësen, dës, dëst	dieser, -e, -es etc.

Algorithm to generate a language family tree

- Take a Swadesh list for a group of related languages.
  Suppose the list has N rows where groups of cognate words are marked with different colors. The cognate words are assumed to have a common origin.

- Find for each pair of languages Li, Lj how many cognates they have.
  If they have k cognates, then their Hamming distance is:

  $$\delta(Li, Lj) = N-k.$$

- Build a distance matrix for the group of languages.
  Apply the UPGMA algorithm to generate a tree.

- Build a common novel cognates matrix.
  Apply the CMM algorithm to generate a tree.