

# ***Sequence Analysis***

*Winter 2023/2024*

*Arne Dür*

- **Scope:**

**Sequence analysis** is an interdisciplinary field involving

- discrete mathematics
- bioinformatics
- molecular biology

to analyze biological sequence data (or, more generally, arbitrary text strings).

- Aims:

The purpose of this course is to explain how

data structures and algorithms

from discrete mathematics and bioinformatics can be used

- to find similar sequences in databases,
- to construct a phylogeny from homologous sequences, or
- to visualize the differences between sequences from a geographic region.

- Example 1:

How to *align* sample from lab to revised Cambridge Reference Sequence ?

	1	1	1	1	1	1	1	1	1	1	1	1	1
	6	6	6	6	6	6	6	6	6	6	6	6	6
	1	1	1	1	1	1	1	1	1	1	1	1	1
	8	8	8	8	8	8	8	9	9	9	9	9	9
	3	4	5	6	7	8	9	0	1	2	3	3	4
												.	
												1	
rCRS	A	C	C	C	C	C	T	C	C	C	C	–	A
SWGDM	A	C	C	C	C	–	T	C	C	C	C	C	A
transcript	M	M	M	M	M	D	M	M	M	M	M	I	M
EMPOP	A	C	C	C	C	T	C	C	C	C	C	–	A
transcript	M	M	M	M	M	R	R	M	M	M	M	–	M

- Example 2:

How to find the *most parsimonious tree* for given sequences ?

S1 = AAAAAAA

S2 = GAAAGAA

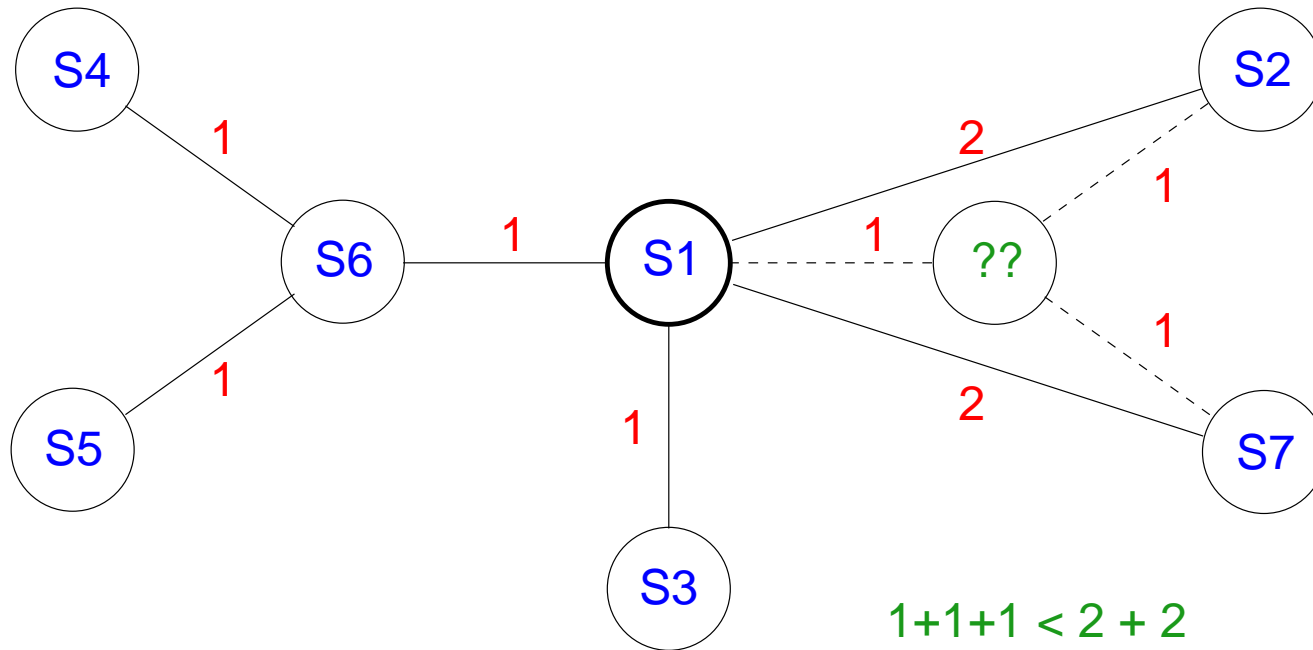
S3 = AAGAAAA

S4 = AGAAAGA

S5 = AGAAAAG

S6 = AGAAAAA

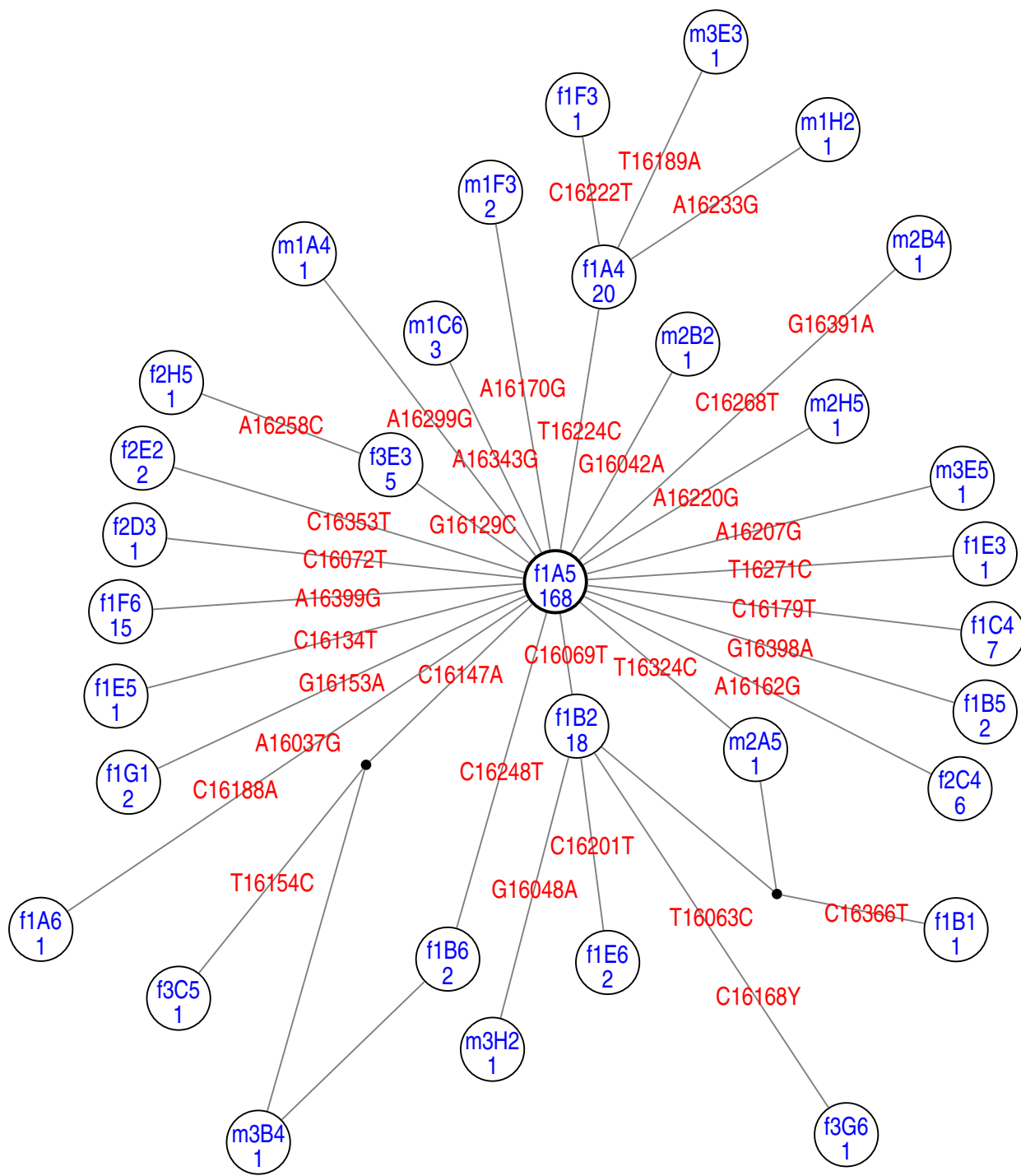
S7 = GAAGAAA

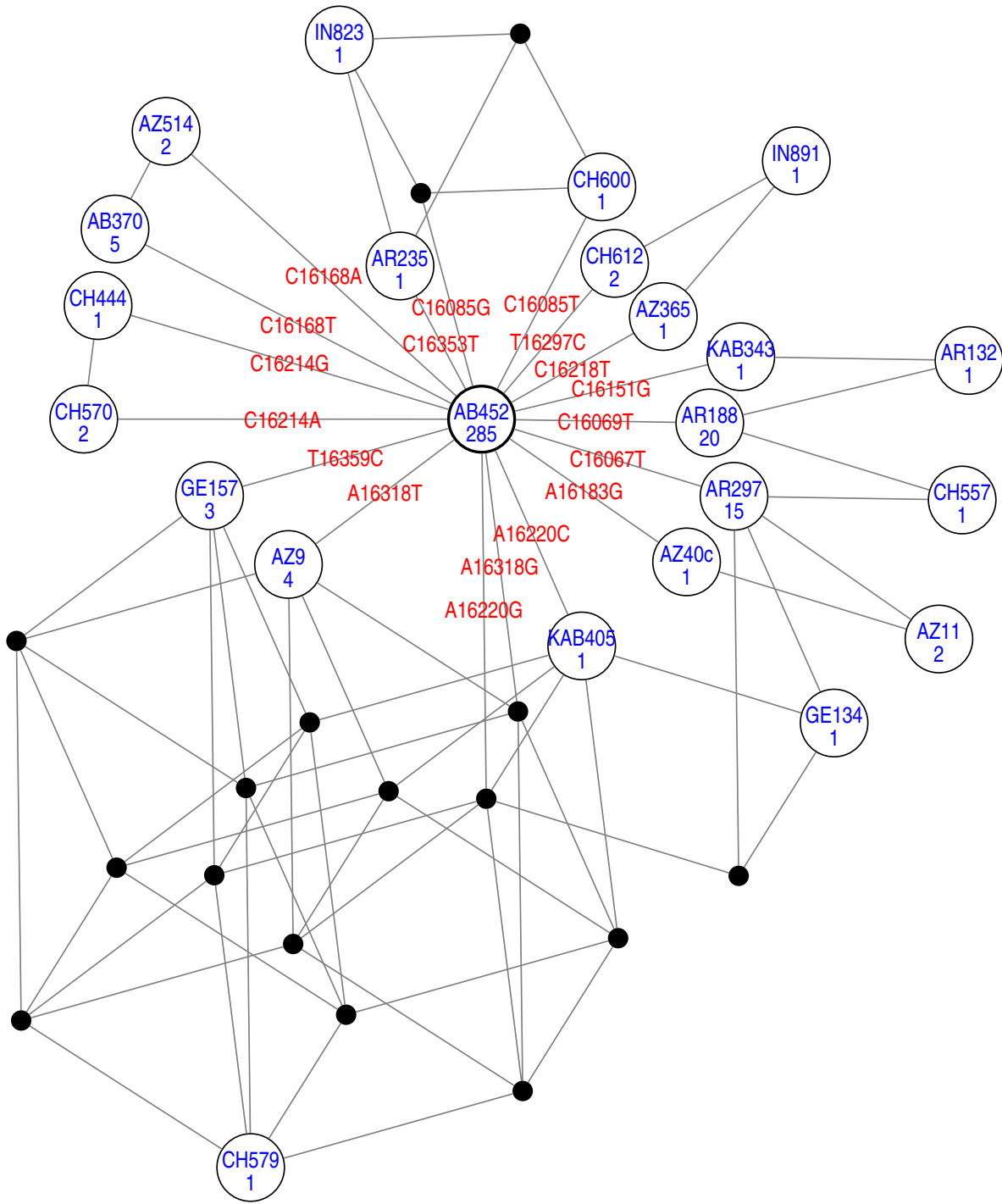


- Example 3:

How to *visualize* sequences from a geographic region ?

- Quasi-median network of 273 samples from **Austria** 2005
- Quasi-median network of 353 samples from **Caucasus** 2001





- Contents:

- String Matching

- \* Finite automaton (Knuth-Morris-Pratt algorithm)
    - \* Suffix tree (Ukkonen's algorithm)

- Sequence Alignment

- \* Global alignment, edit transcript, and edit distance
    - \* Dynamic programming, distance table, and edit graph
    - \* Similarity of strings
    - \* Local alignment



- Contents (continued):
  - Strings and Evolutionary Trees
    - \* The Perfect Phylogeny Problem
    - \* Compatibility and strong compatibility of characters
    - \* The Maximum Parsimony Problem and Steiner trees
  - Quasimedian Networks
    - \* Quasimedian algebra
    - \* Construction of quasimedian networks
    - \* Visualization

- Scheduled Time and Place:

- Lecture: VO2 Thursday 10:15-12:00, HSB 7 (starting on 5.10.)  
On-line registration necessary to get OLAT access !

- Proseminar: PS1 Thursday 12:15-13:00, HSB 7 (starting on 5.10.)  
On-line registration necessary to participate !

- Materials:

Lecture notes, problem sheets for weekly exercises, and supplementary materials can be downloaded from OLAT.