Bioinformatics and Advanced Programming

Jan T. Kim

THE Pirbright INSTITUTE

BCS Advanced Programming SG, 11 Dec 2014

Abstract

Bioinformatics can be defined strictly as the science of information in biological systems, or more broadly as developing and applying computational tools for analysing biological data. The biological processes that generate this information, particularly evolution, are highly complex, and therefore analysis of biological information is often computationally challenging. I will present the following selected topics and highlight the advanced computing challenges they involve, and also outline advances in the biosciences that have been enabled by tackling these challenges.
Many bioinformatics analyses are based on DNA sequences which today can be determined at very high volume through "Next Generation Sequencing" (NGS) techniques. As a result, the volume of publicly available sequence data has reached the range of petabytes. Searching this body of data requires highly optimised computational approaches, such as BLAST ("Basic Local Alignment Search Tool").
More recently, NGS methods that generate very large numbers of "short reads", i.e. strings of sequence . Central computational challenges resulting from these new technologies are "de novo assembly" of the original long sequence(s) from short reads, and mapping very large numbers of short reads to a known reference sequence.
Phylogeny analysis, i.e. reconstruction of ancestry relationships among species, is a classical field of bioinformatics which typically involves two steps, first a multiple alignment of the sequences is computed which subsequently is used to compute a tree. Computing multiple alignments is an optimisation problem that can only be approximately solved.

The Pirbright Institute

THE 100 YEARS Pirbright INSTITUTE

Preventing and controlling viral diseases

Core funding:

BBSRC bioscience for the future

Project funding by BBSRC and many others.

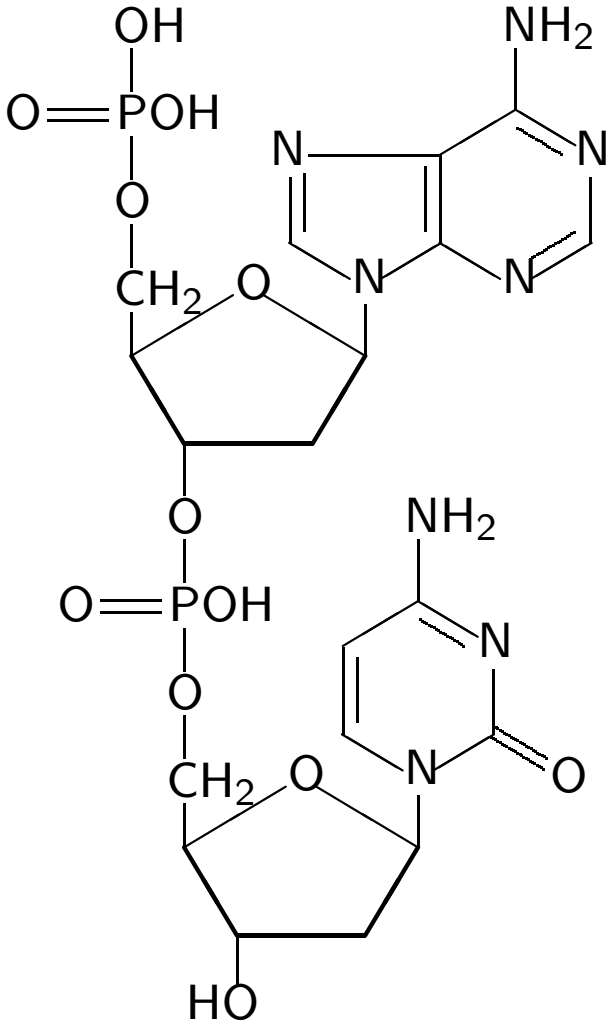
Outline

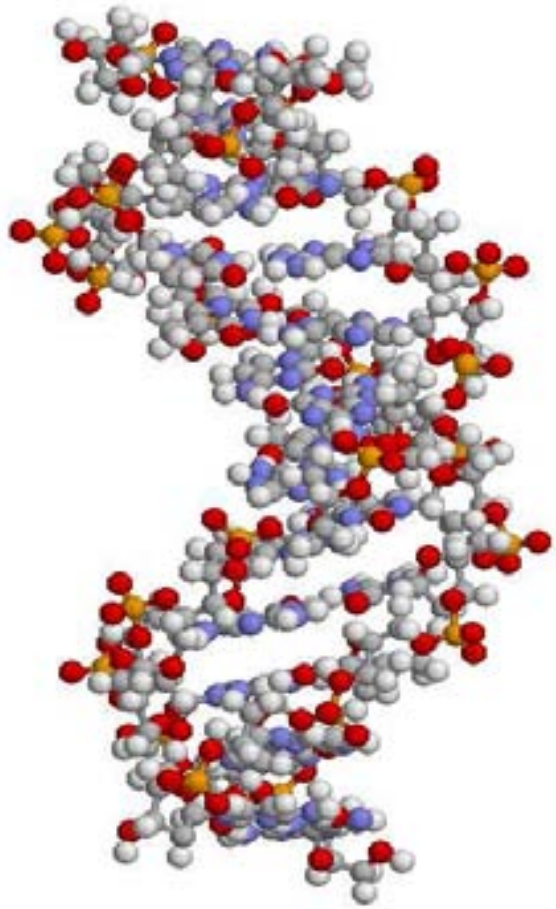
Bioinformatics: Definition(s)

- Scientific inquiry into **information in biological systems**.
- Computational analysis of **biological data**.
- Computer assisted mining of **biological literature**.

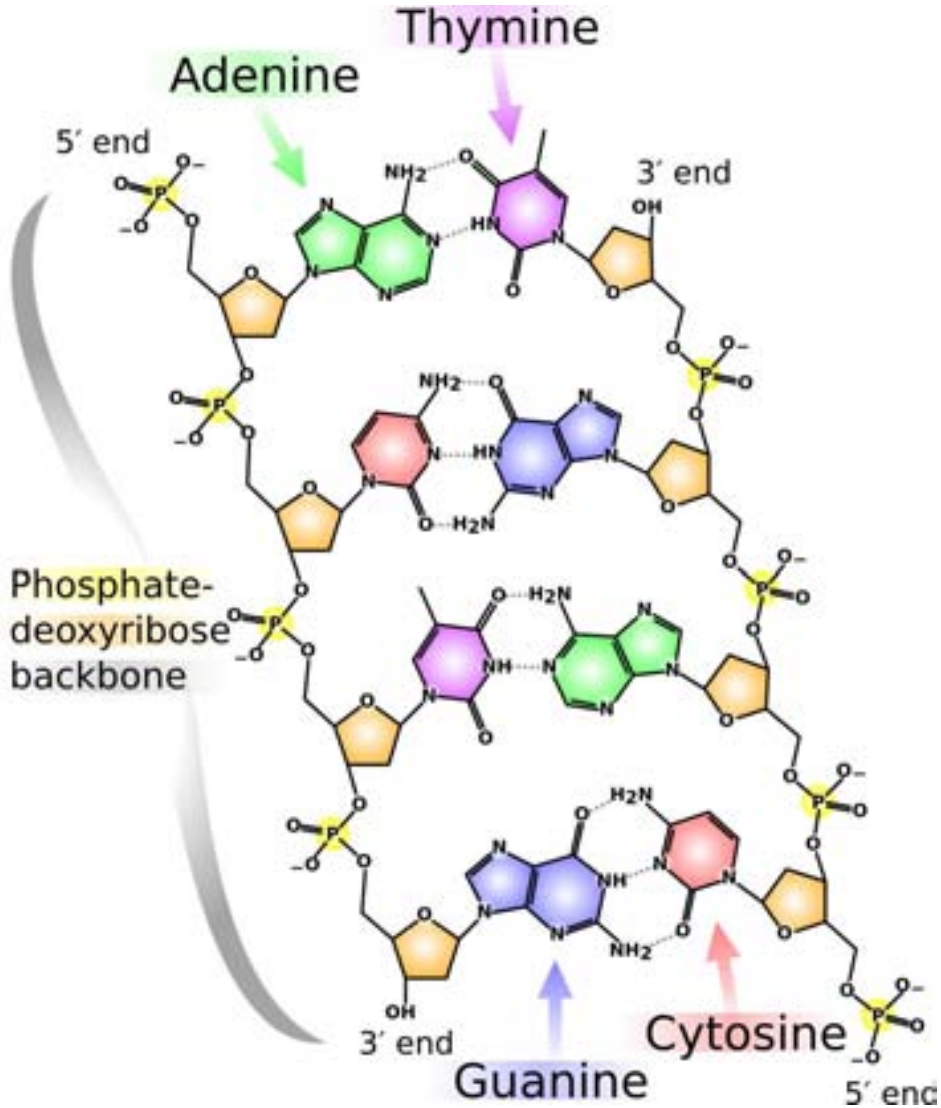
Outline

DNA: Structure



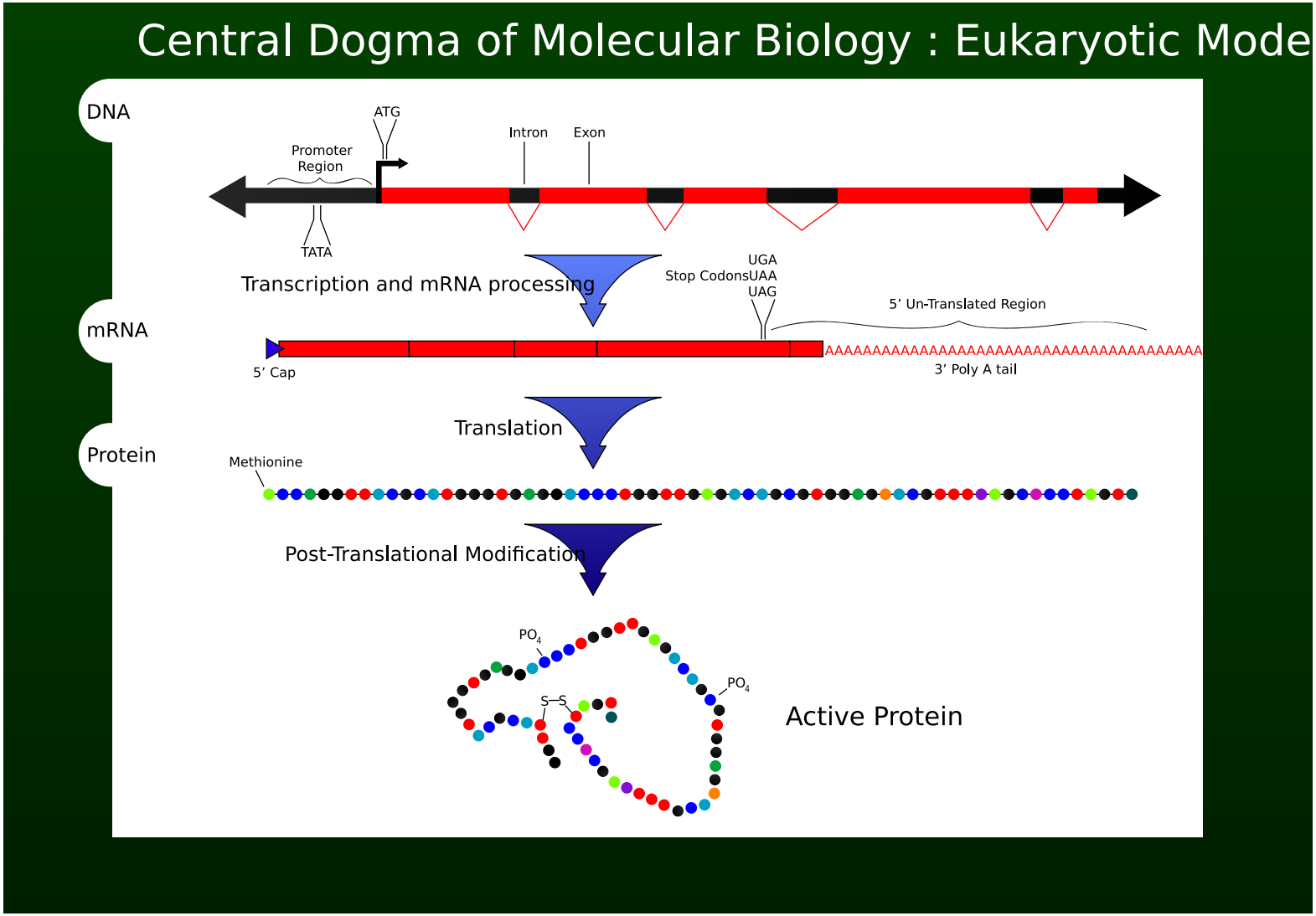


Base Complementarity



http://commons.wikimedia.org/wiki/File:DNA_chemical_structure.svg

The "Central Dogma"



http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology
http://en.wikipedia.org/wiki/File:Cdmb.svg

The Success of Bioinformatics

- **The Object: Information in biological systems**:

  In living systems, a dynamics of information has gained control over the dynamics of energy, which determines the behavior of most non-living systems.

  [Langton, 1992]

- Genetic information is **digital**.

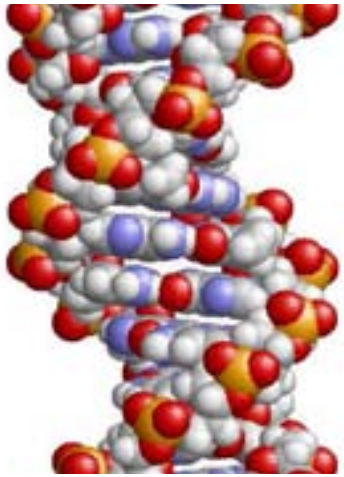
The Success of Bioinformatics

- **The Object: Information in biological systems**:

  In living systems, a dynamics of information has gained control over the dynamics of energy, which determines the behavior of most non-living systems.

  [Langton, 1992]

- Genetic information is **digital**.



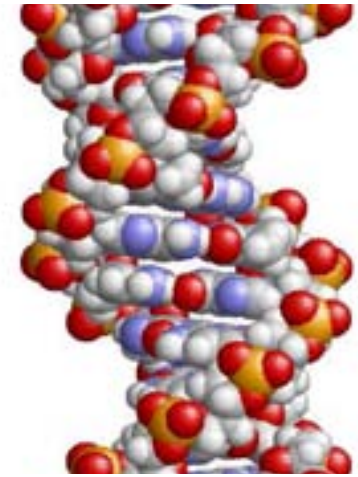
The Success of Bioinformatics

- **The Object: Information in biological systems**:

  In living systems, a dynamics of information has gained control over the dynamics of energy, which determines the behavior of most non-living systems.

  [Langton, 1992]

- Genetic information is **digital**.



```
TACCGTCAC
CTACACCAT
ACCTACATG
TTCACATTA
```

Sequence Data Is Big Data

- NCBI-GenBank Flat File Release 204.0 (15 Oct 2014):

  `ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt`

  - 178,322,253 loci,
  - 181,563,676,918 bases.

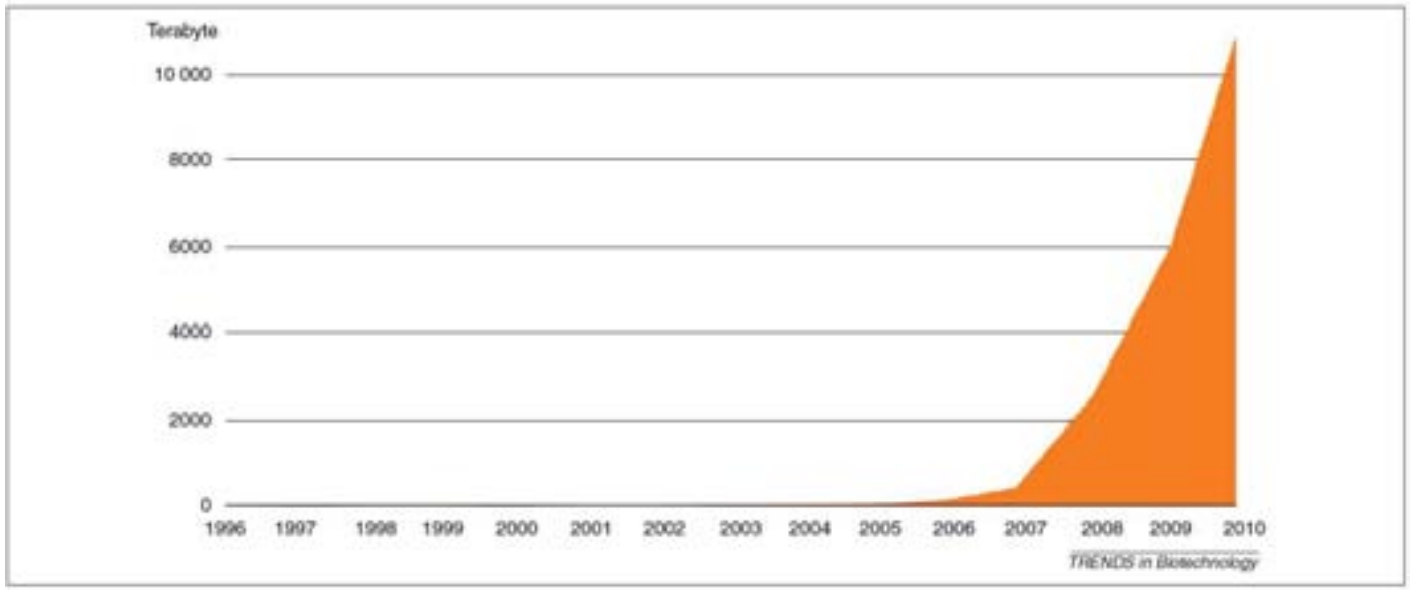


Figure 1. The scale of data growth. The chart shows the trend in storage capacity needed to store biological data at EMBL-EBI (a terabyte is a million million bytes).

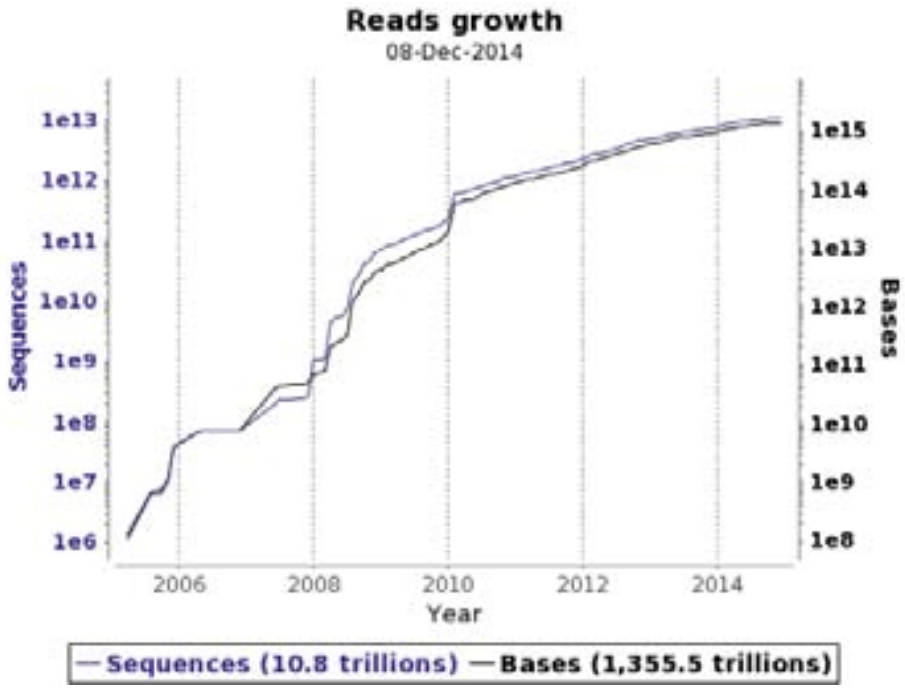
[Crosswell and Thornton, 2012]

Sequence Data Is Big Data

- NCBI-GenBank Flat File Release 204.0 (15 Oct 2014):

  `ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt`

  - 178,322,253 loci,
  - 181,563,676,918 bases.



`http://www.ebi.ac.uk/ena/about/statistics`

Outline

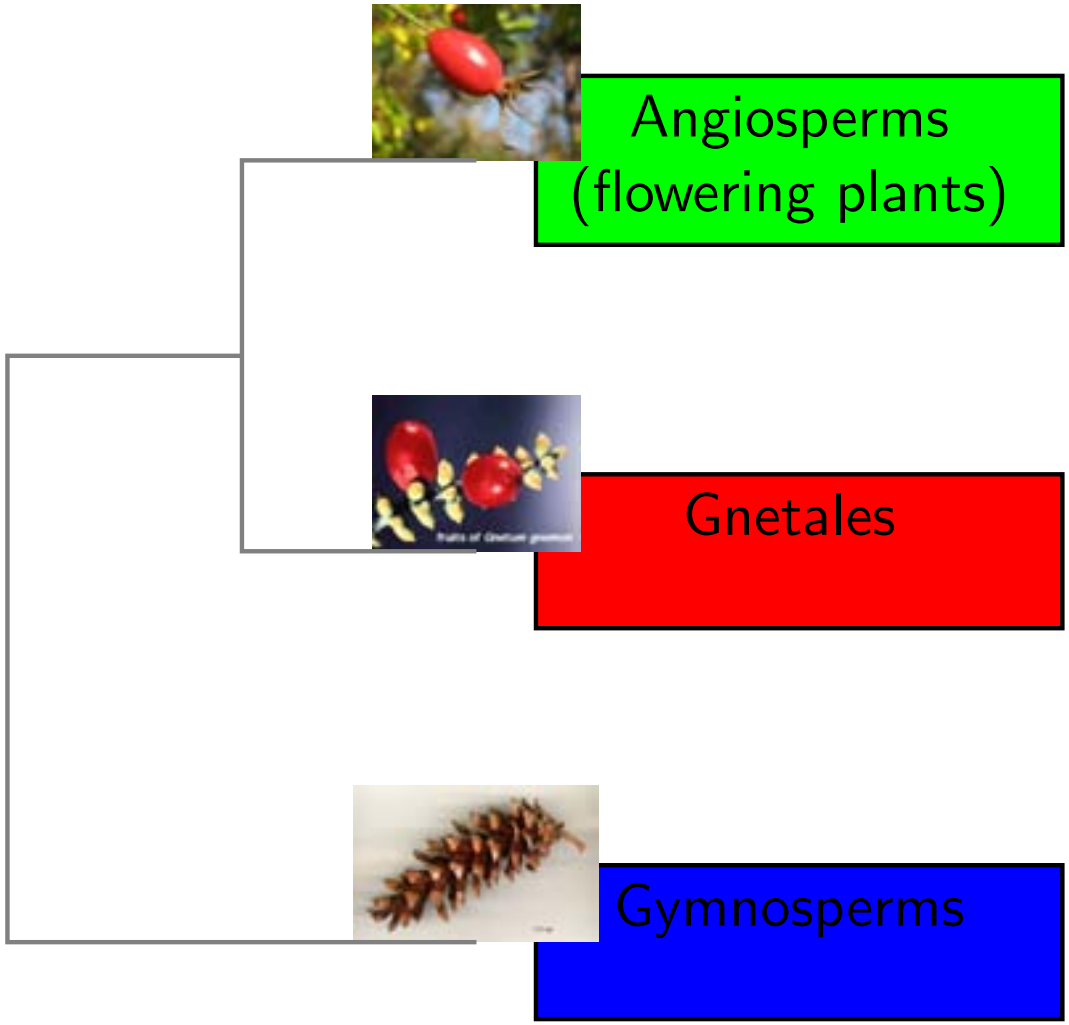
Land Plant Phylogeny

Angiosperms (flowering plants)

Gnetales

Gymnosperms

Land Plant Phylogeny



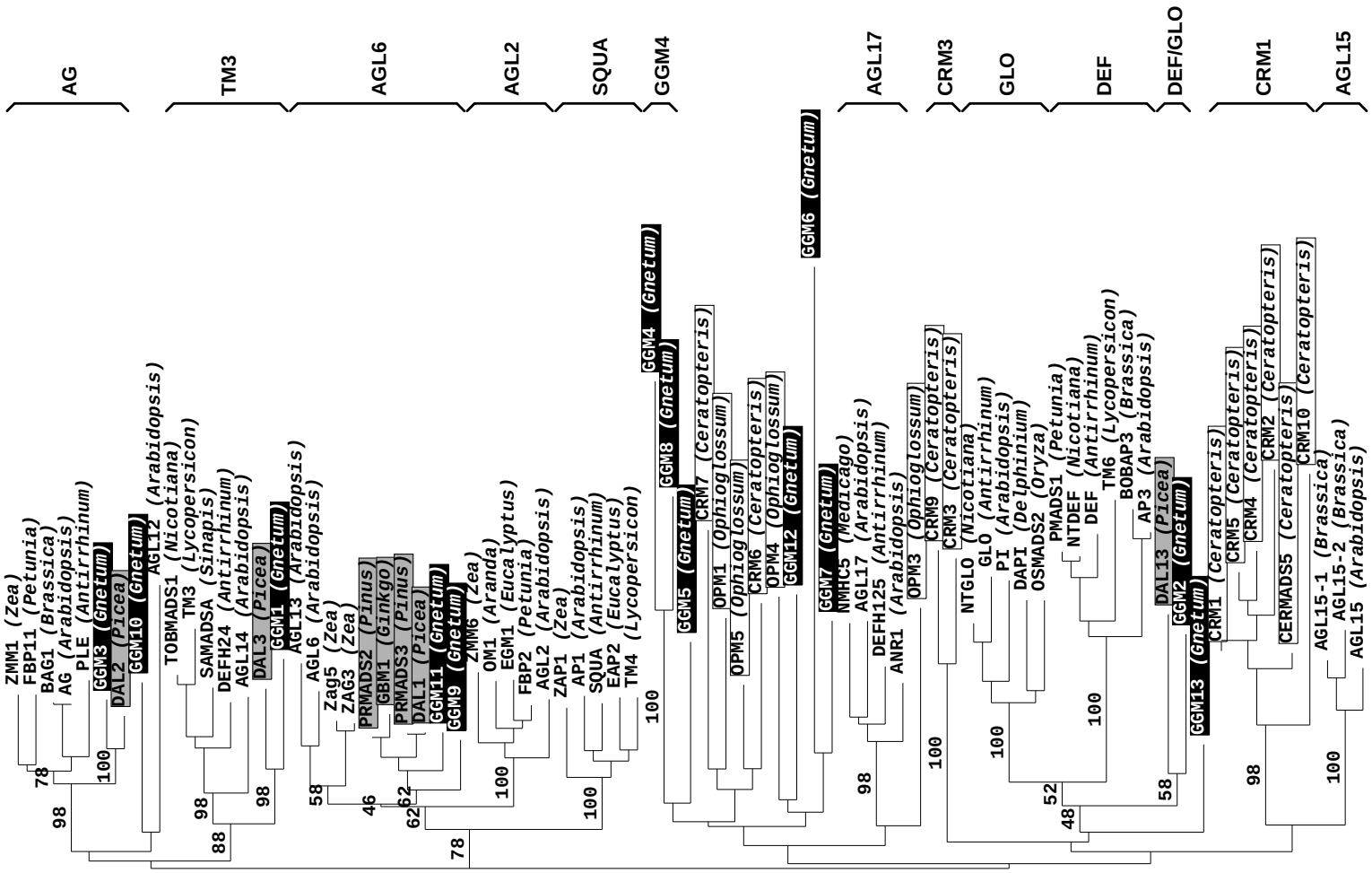
Land Plant Phylogeny

Angiosperms (flowering plants)

Gnetales

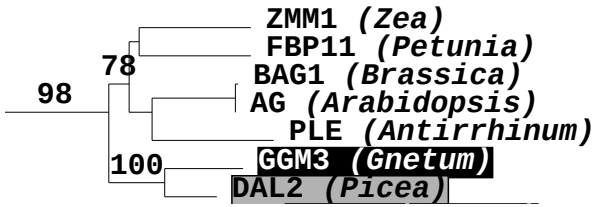
Gymnosperms

Phylogeny of MADS Proteins



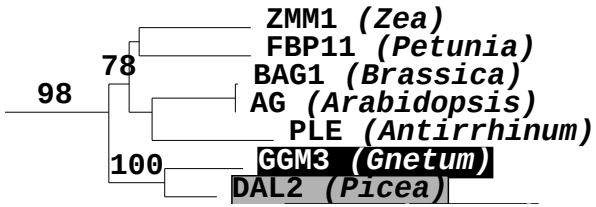
Angiosperms | Gnetales | Gymnosperms

The AG and AGL6 Subfamilies

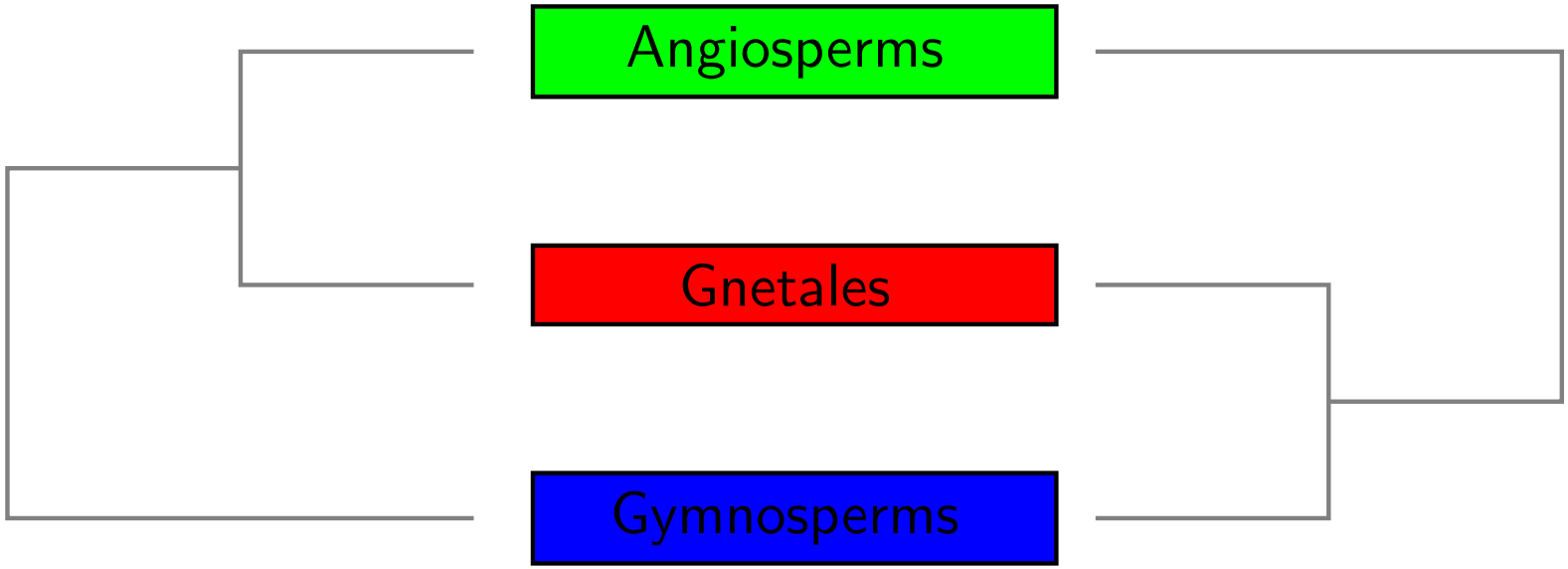


AG subfamily

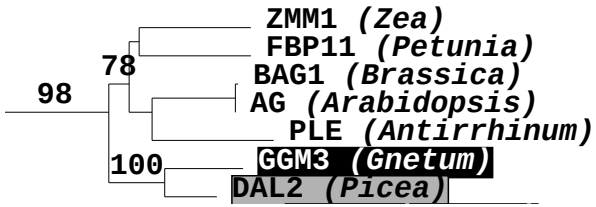
The AG and AGL6 Subfamilies



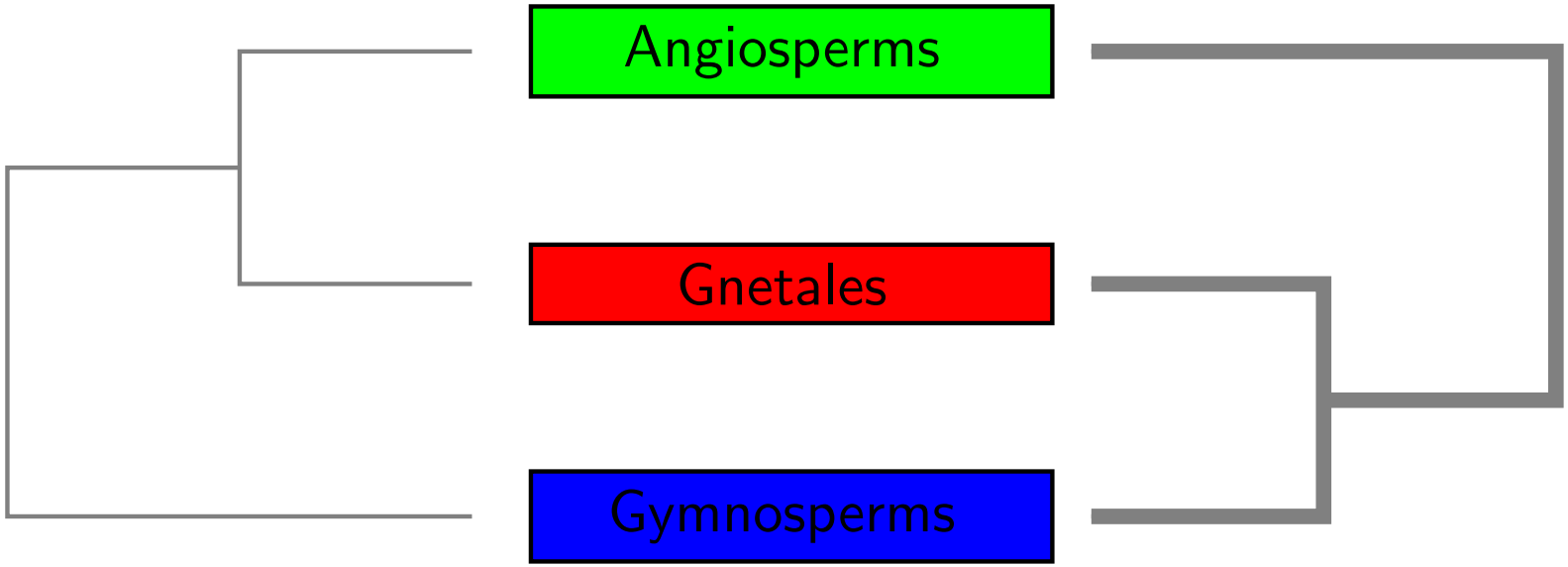
AG subfamily



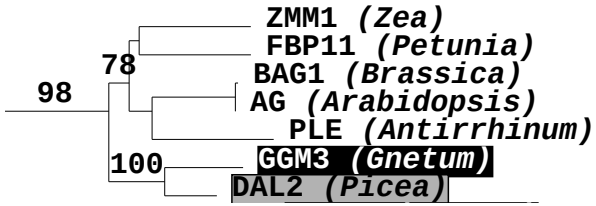
The AG and AGL6 Subfamilies

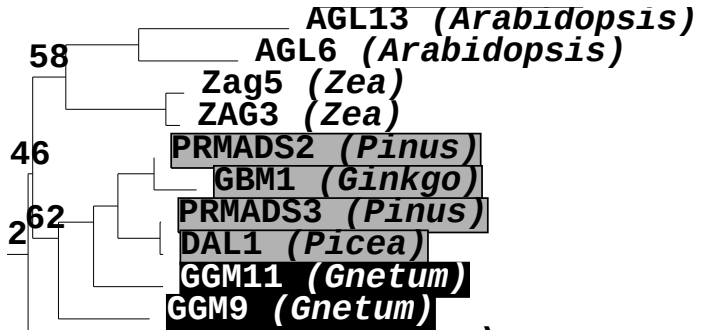


AG subfamily



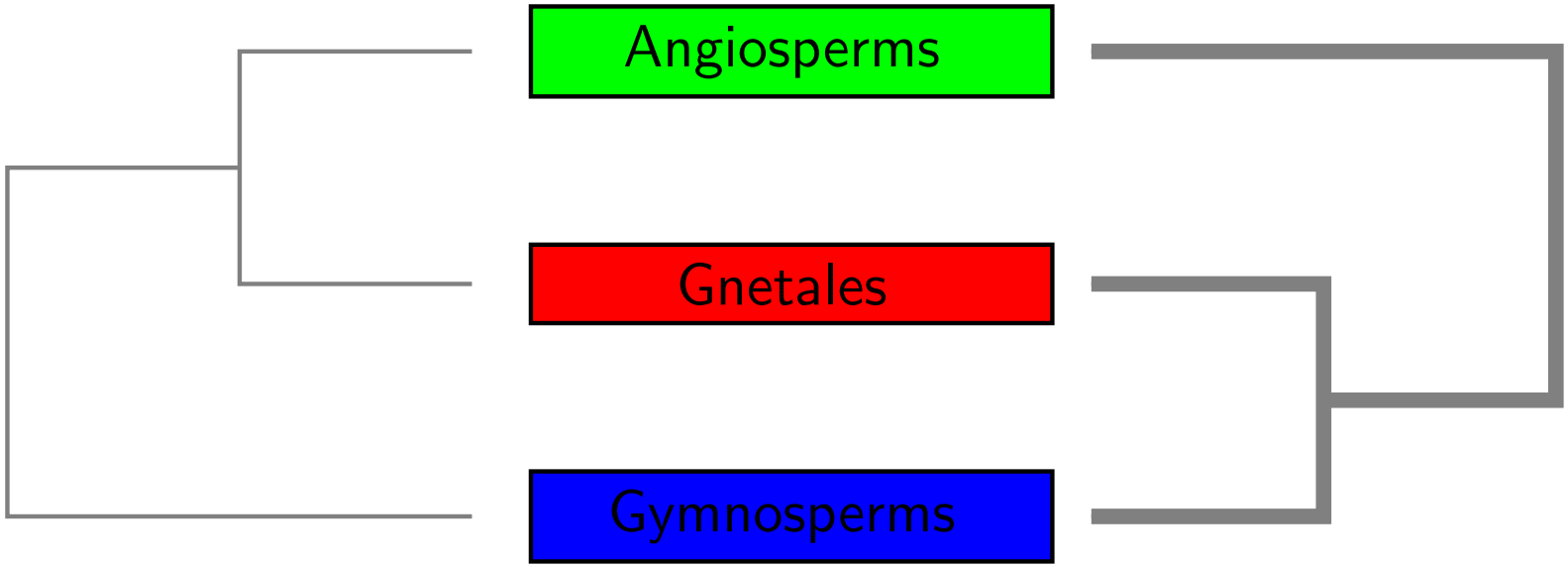
The AG and AGL6 Subfamilies



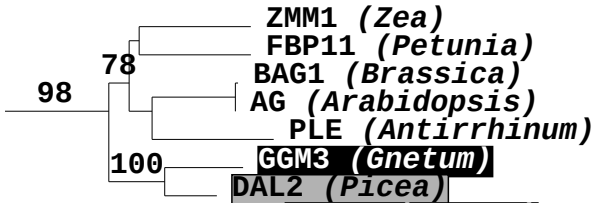


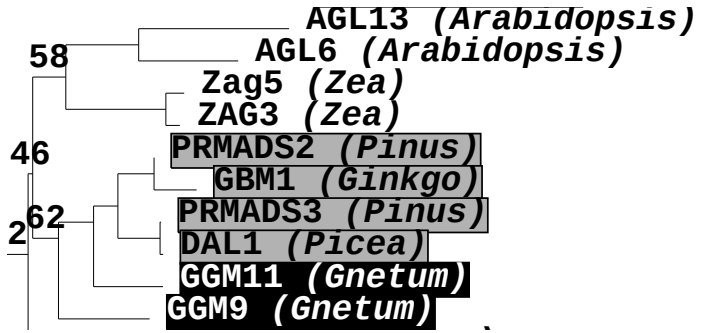
AG subfamily

AGL6 subfamily



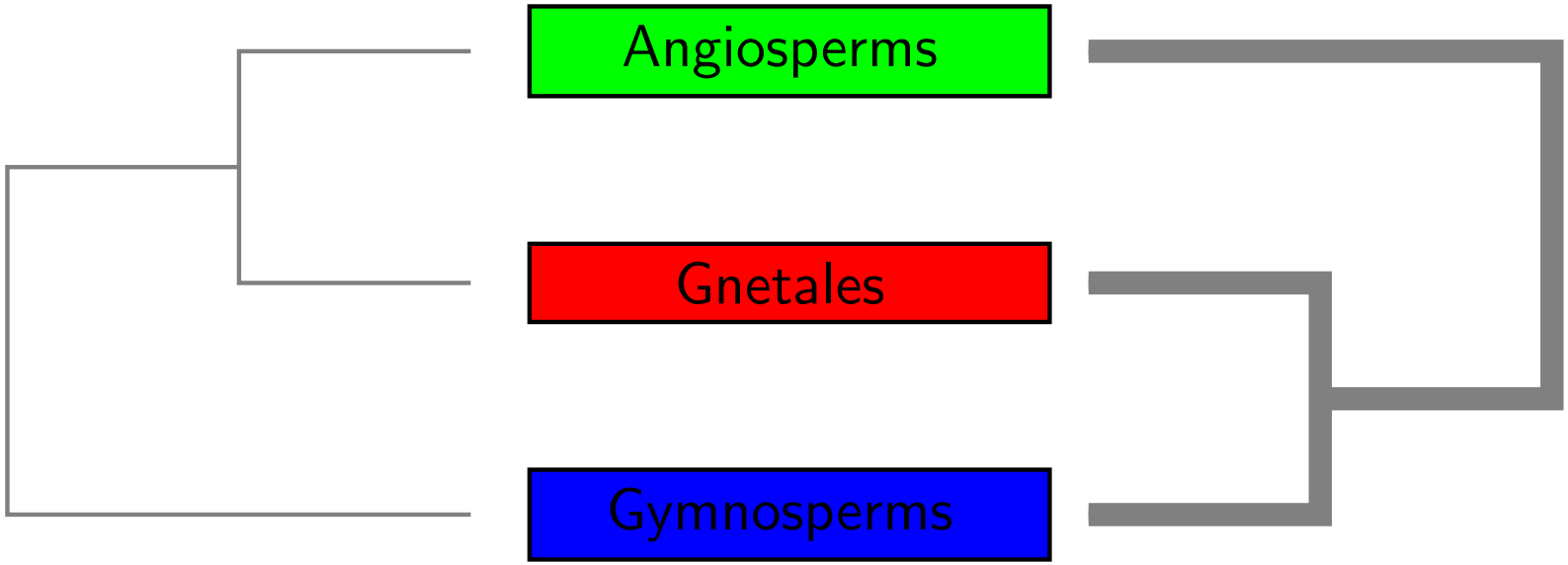
The AG and AGL6 Subfamilies



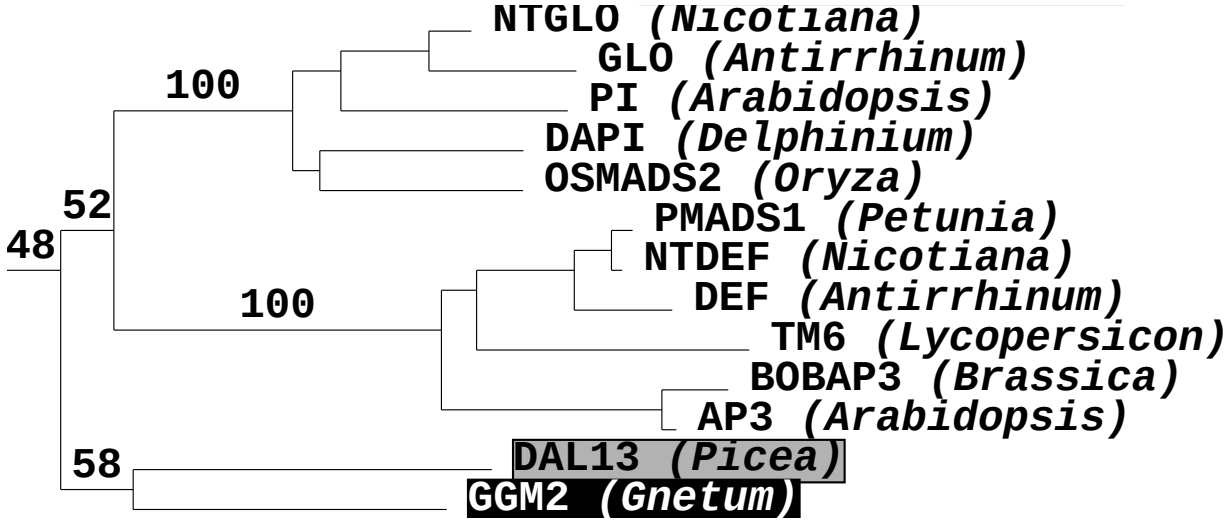


AG subfamily

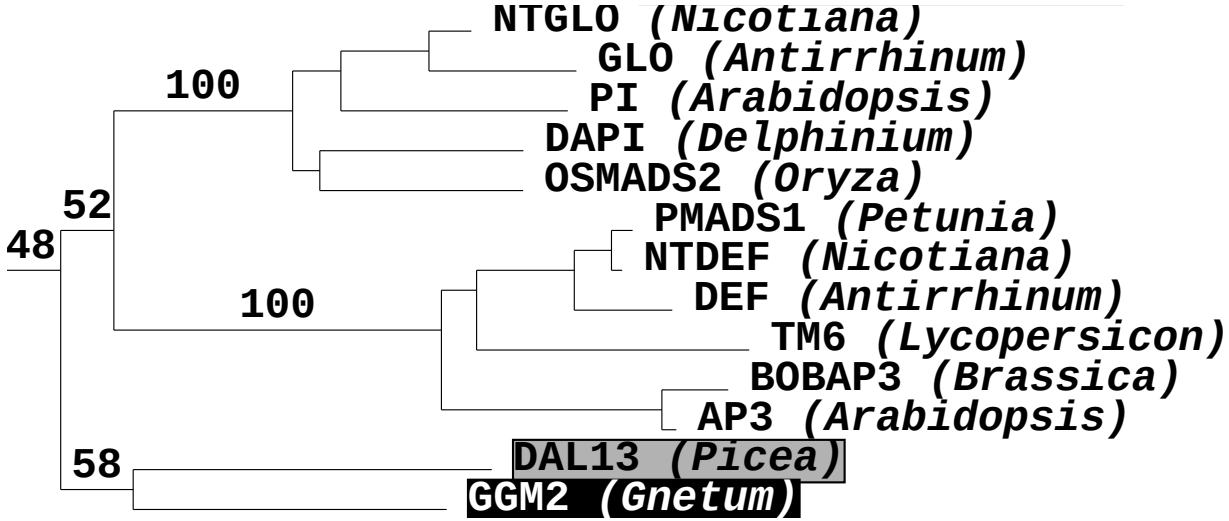
AGL6 subfamily

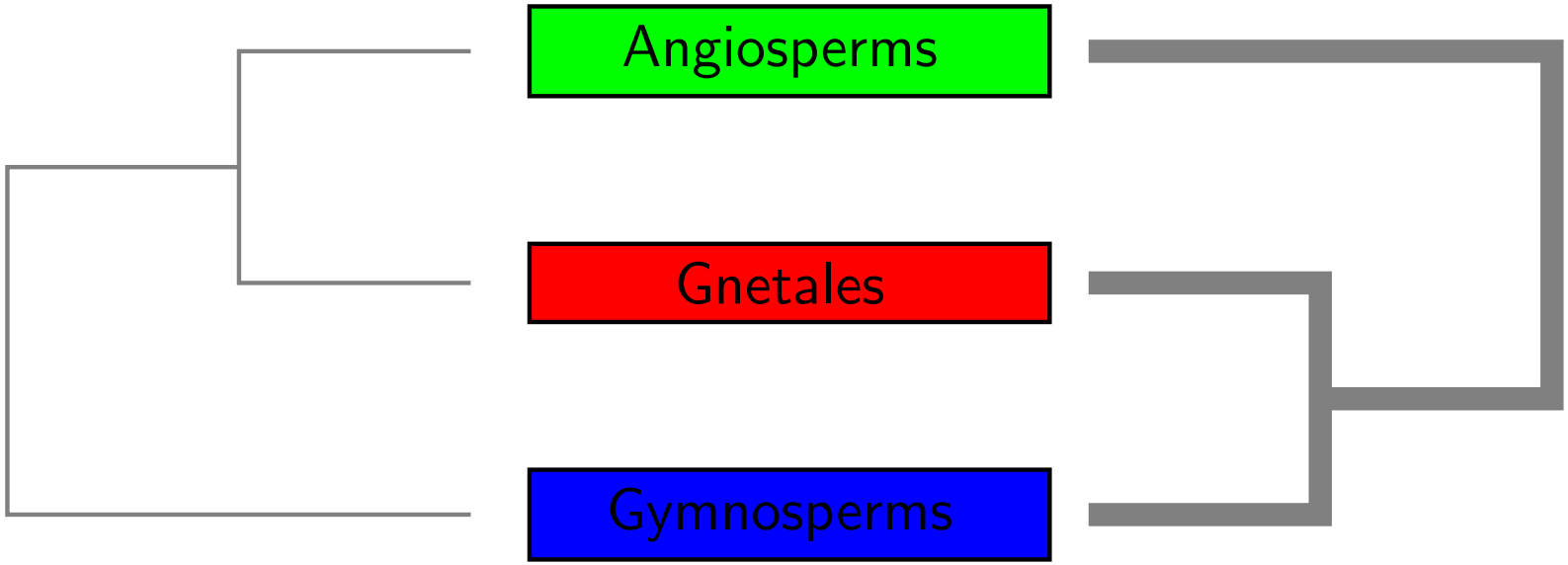


The DEF and GLO Subfamilies

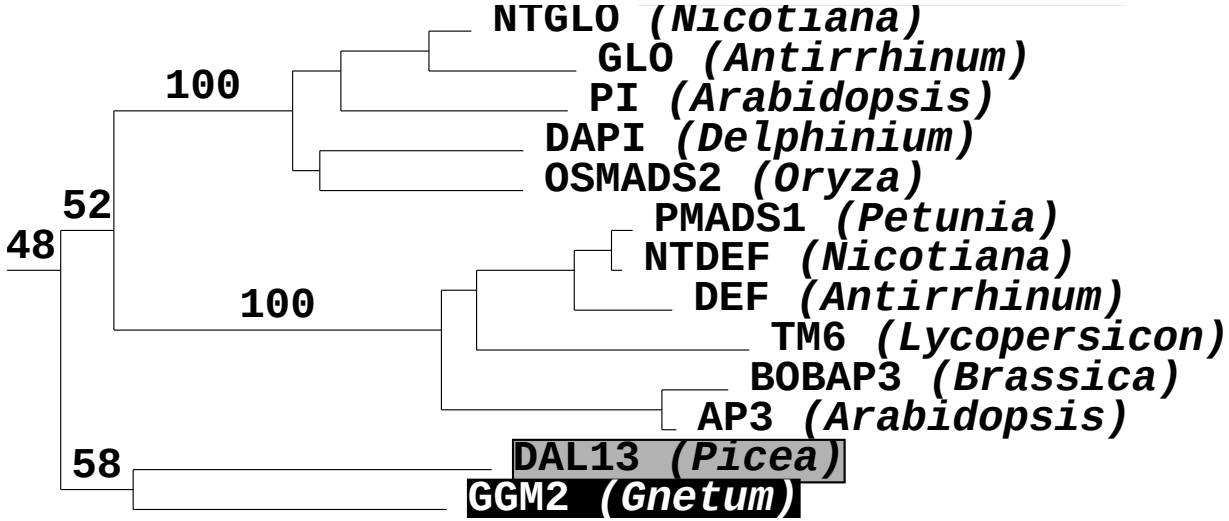


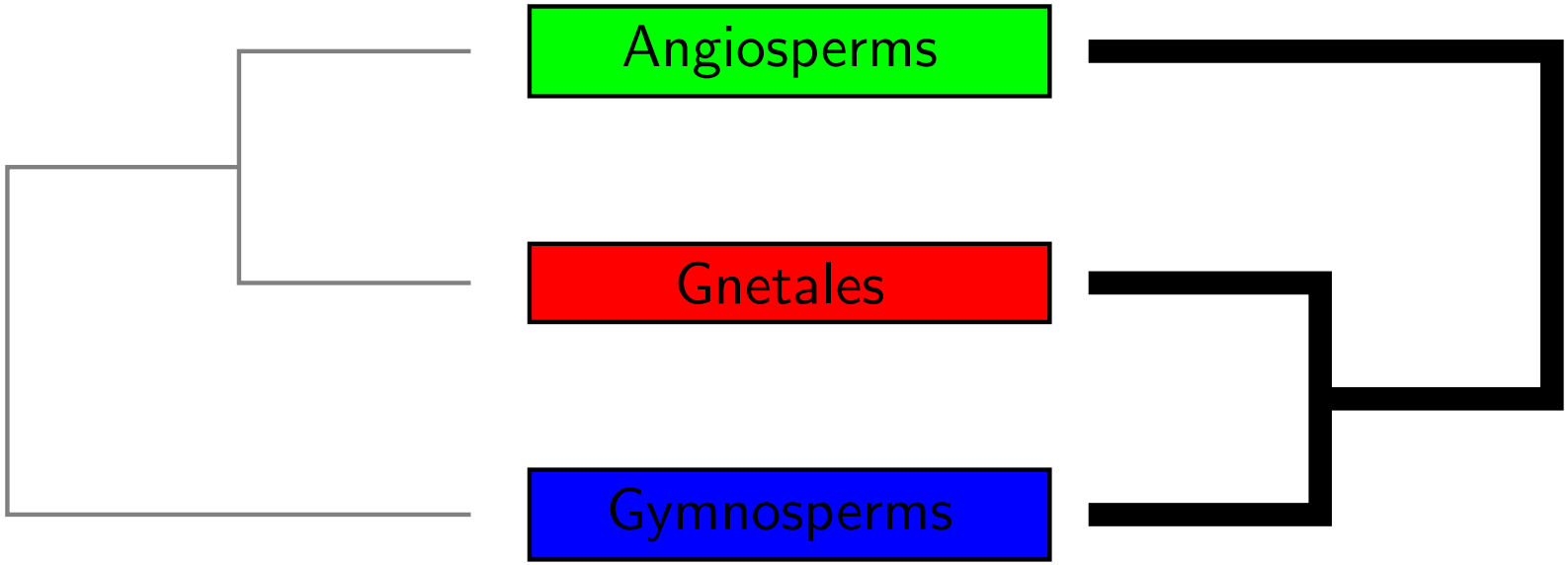
The DEF and GLO Subfamilies





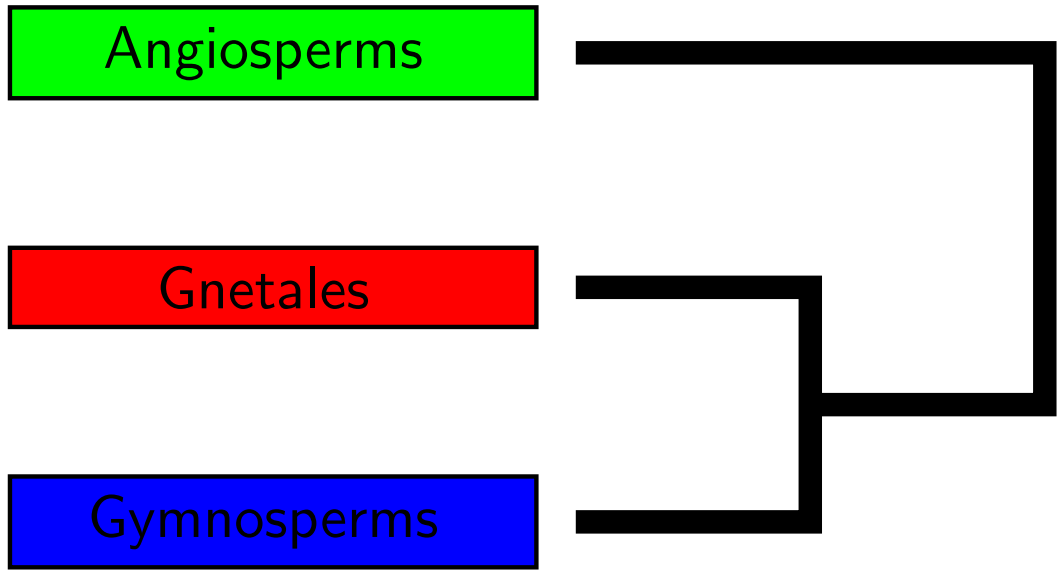
The DEF and GLO Subfamilies





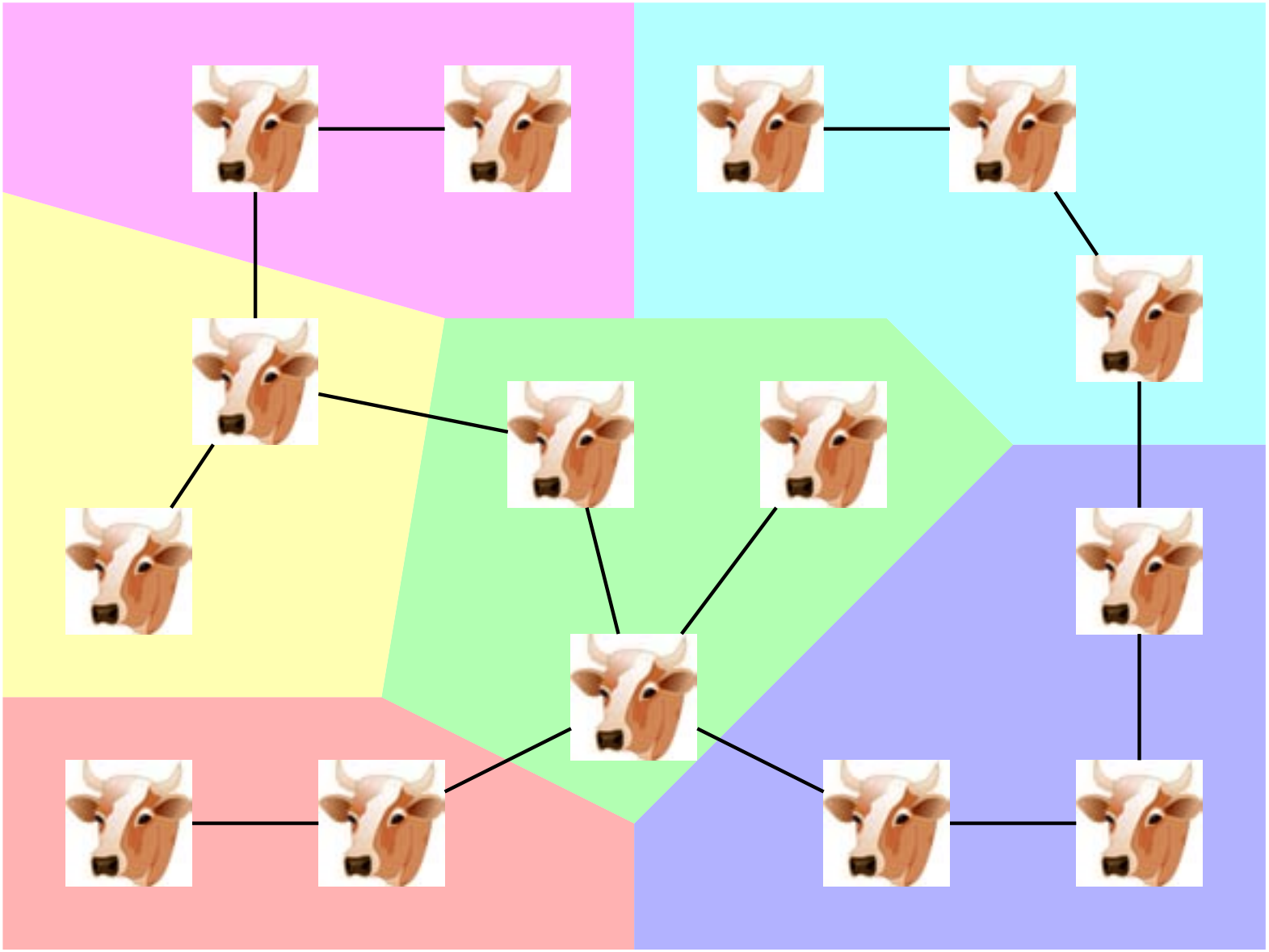
Conclusion: Land Plant Phylogeny

MADS-Box Genes Reveal That Gnetophytes Are More Closely Related to Conifers than to Flowering Plants [Winter et al., 1999].

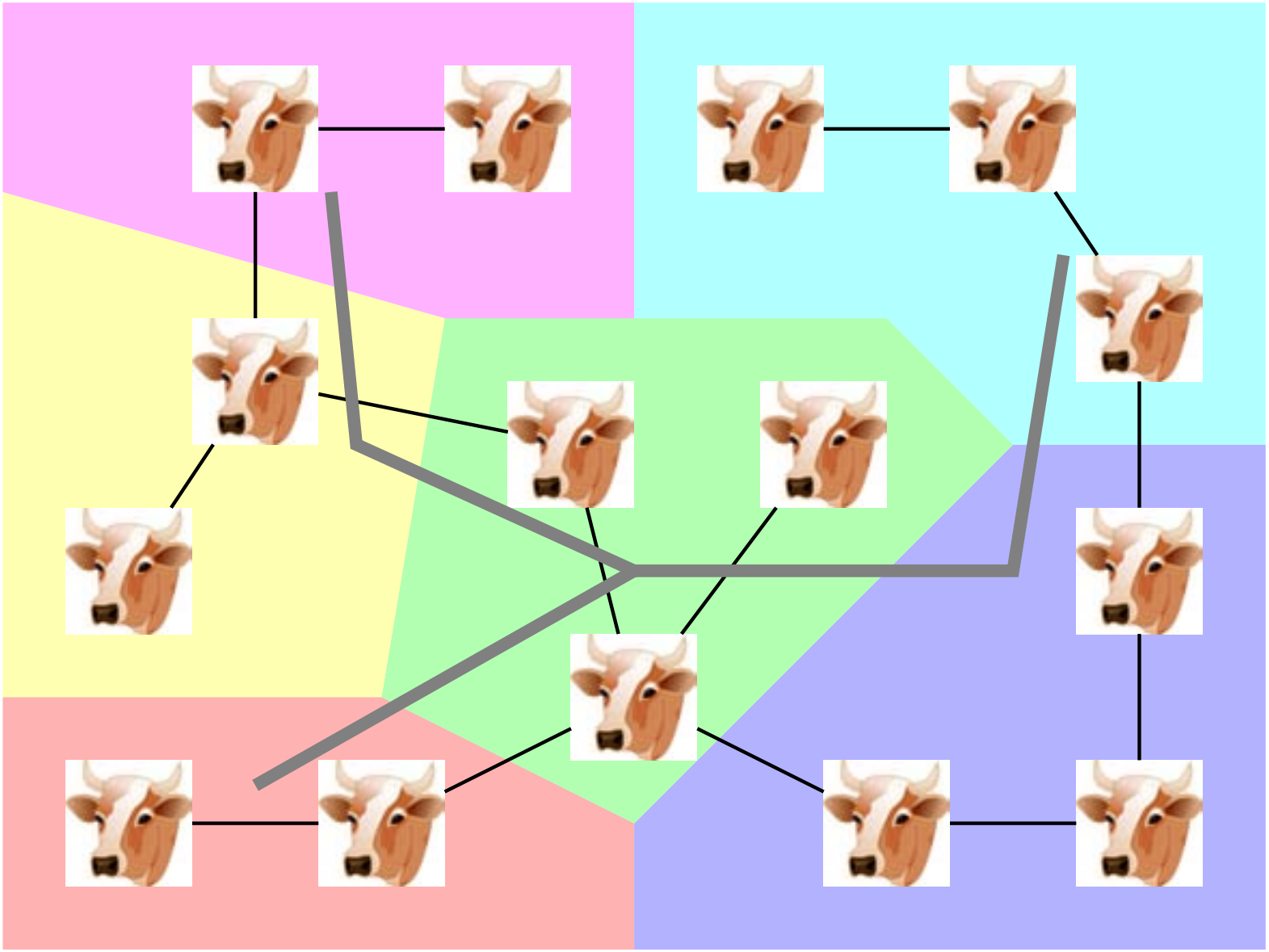


Outline

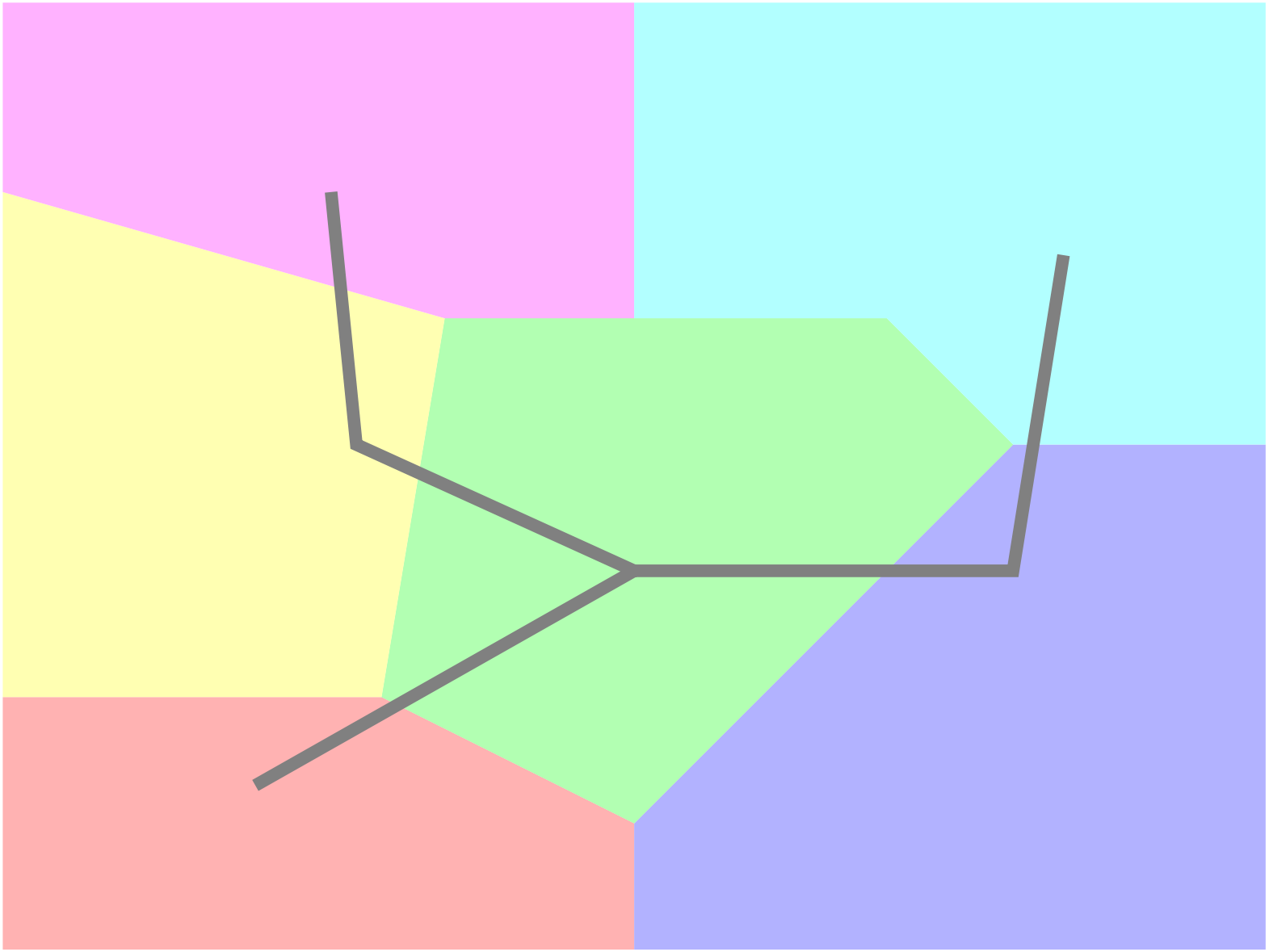
Transmission Trees



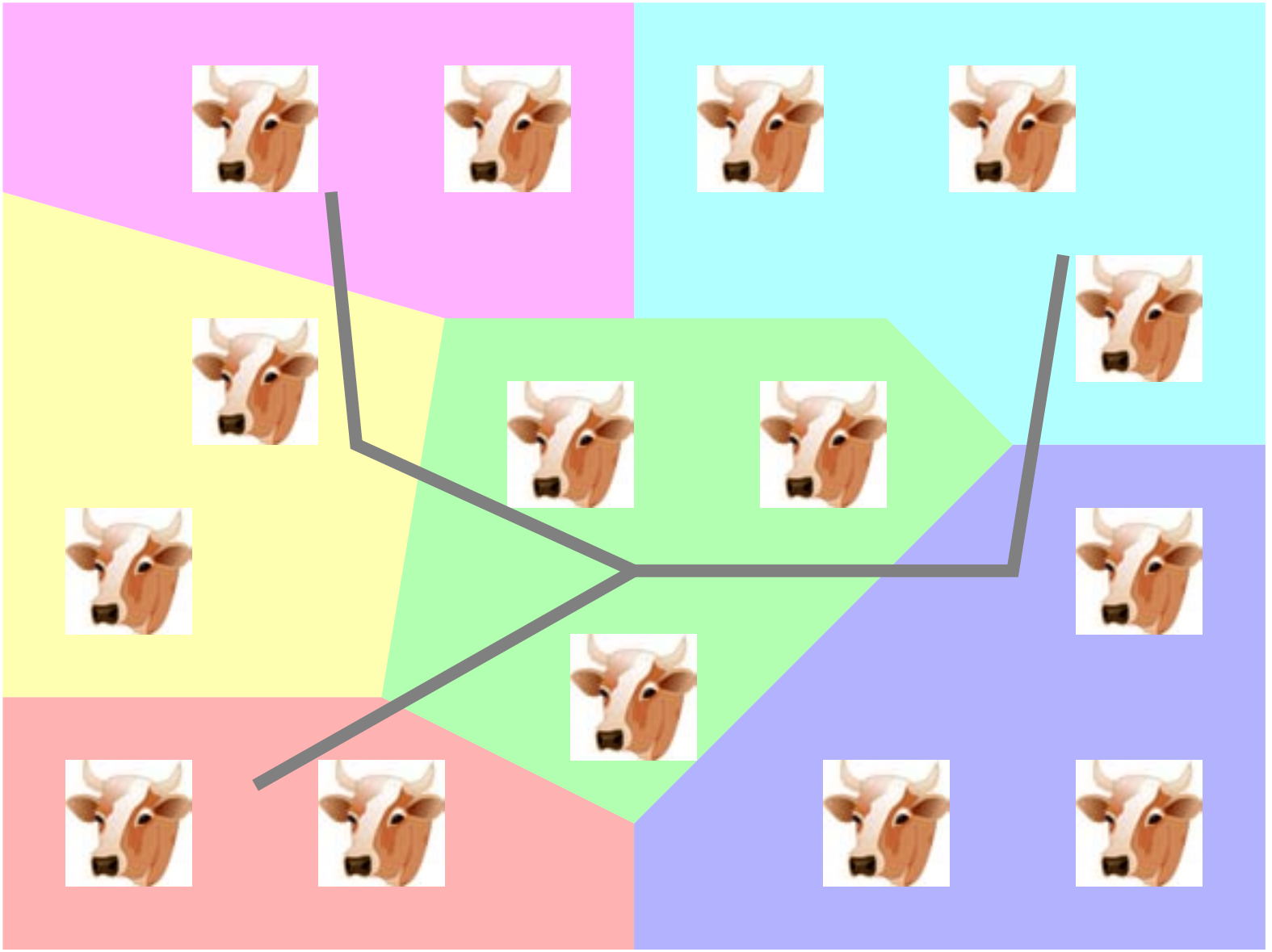
Transmission Trees



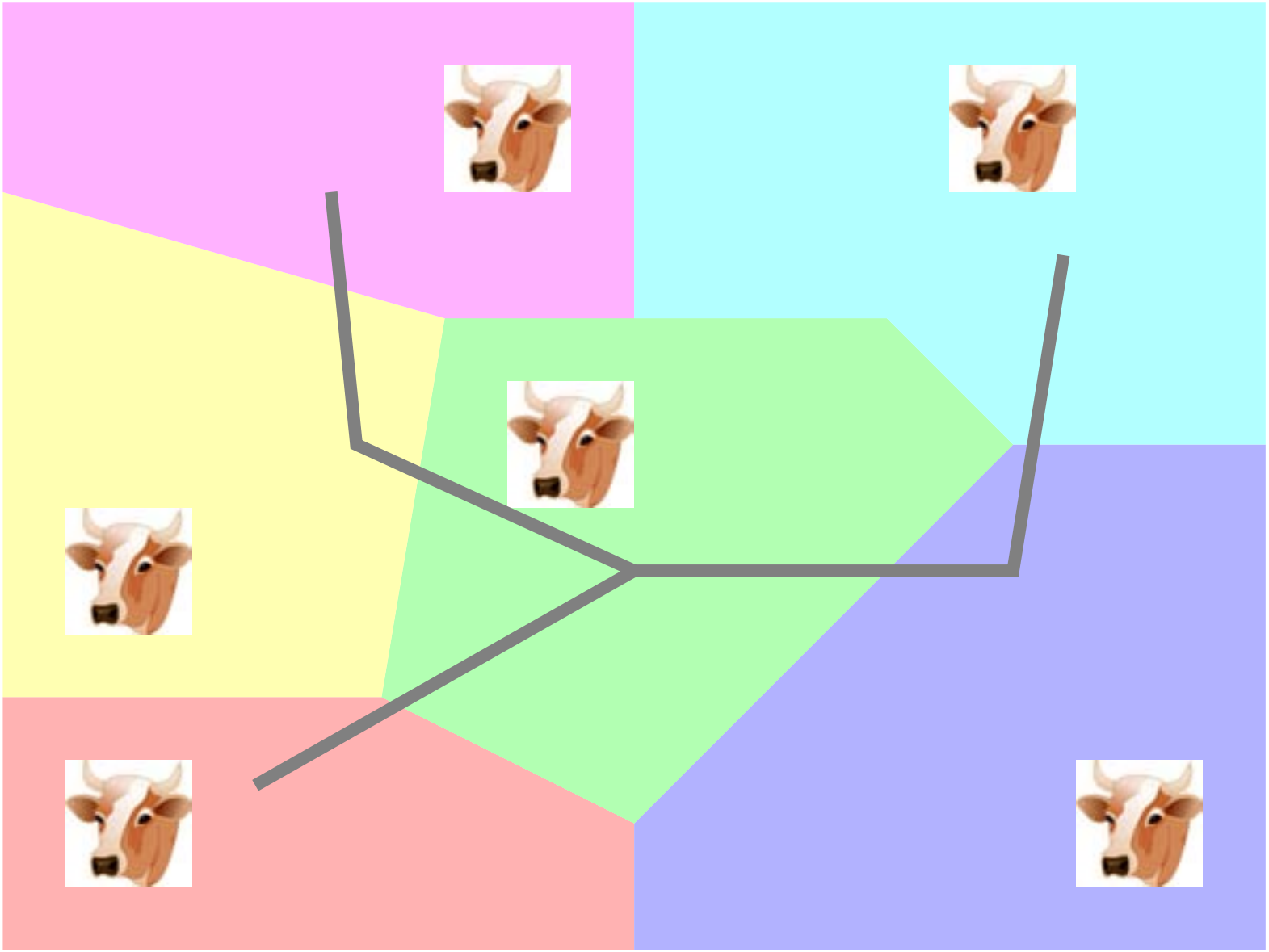
Transmission Trees



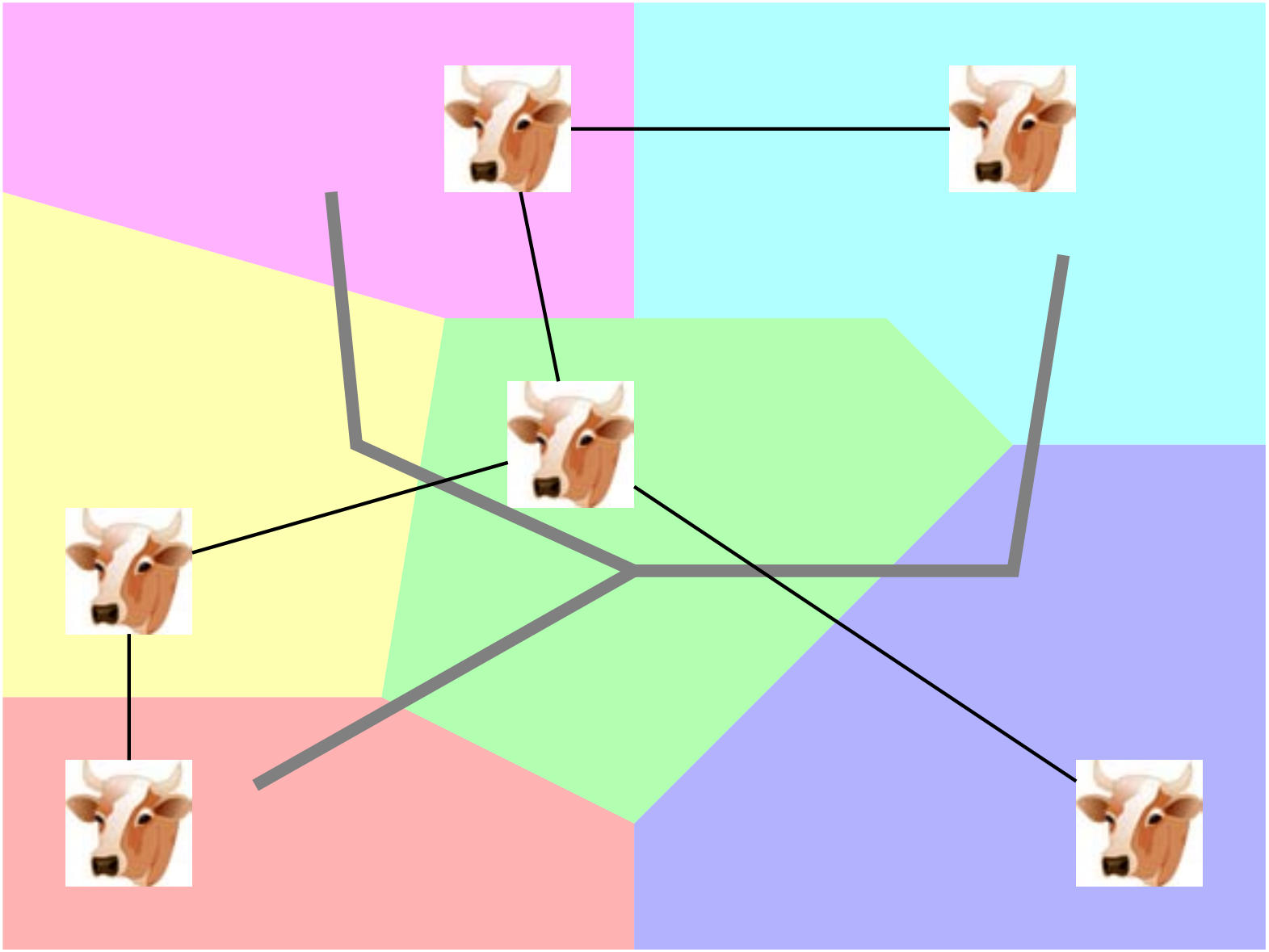
Transmission Trees



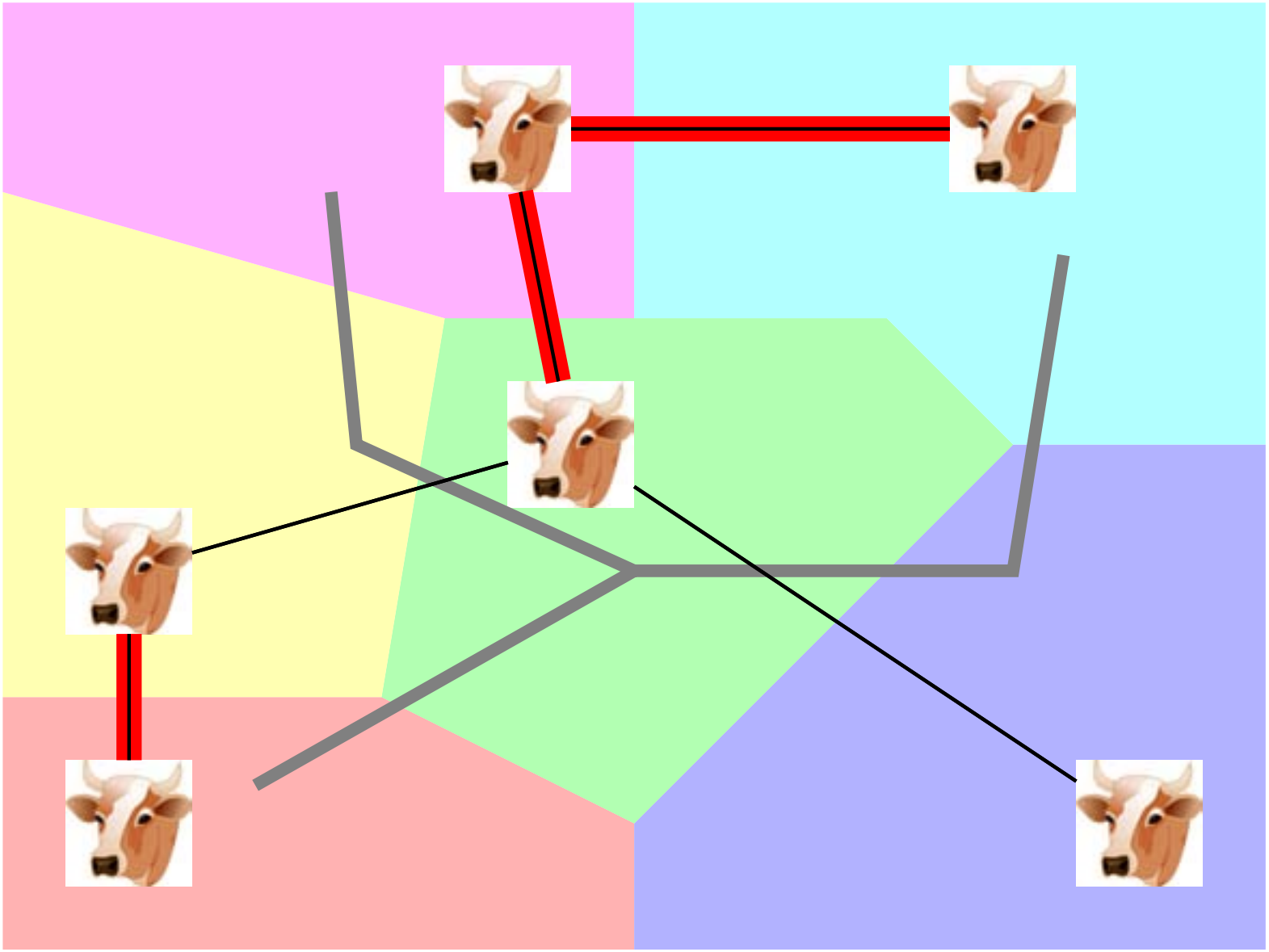
Transmission Trees



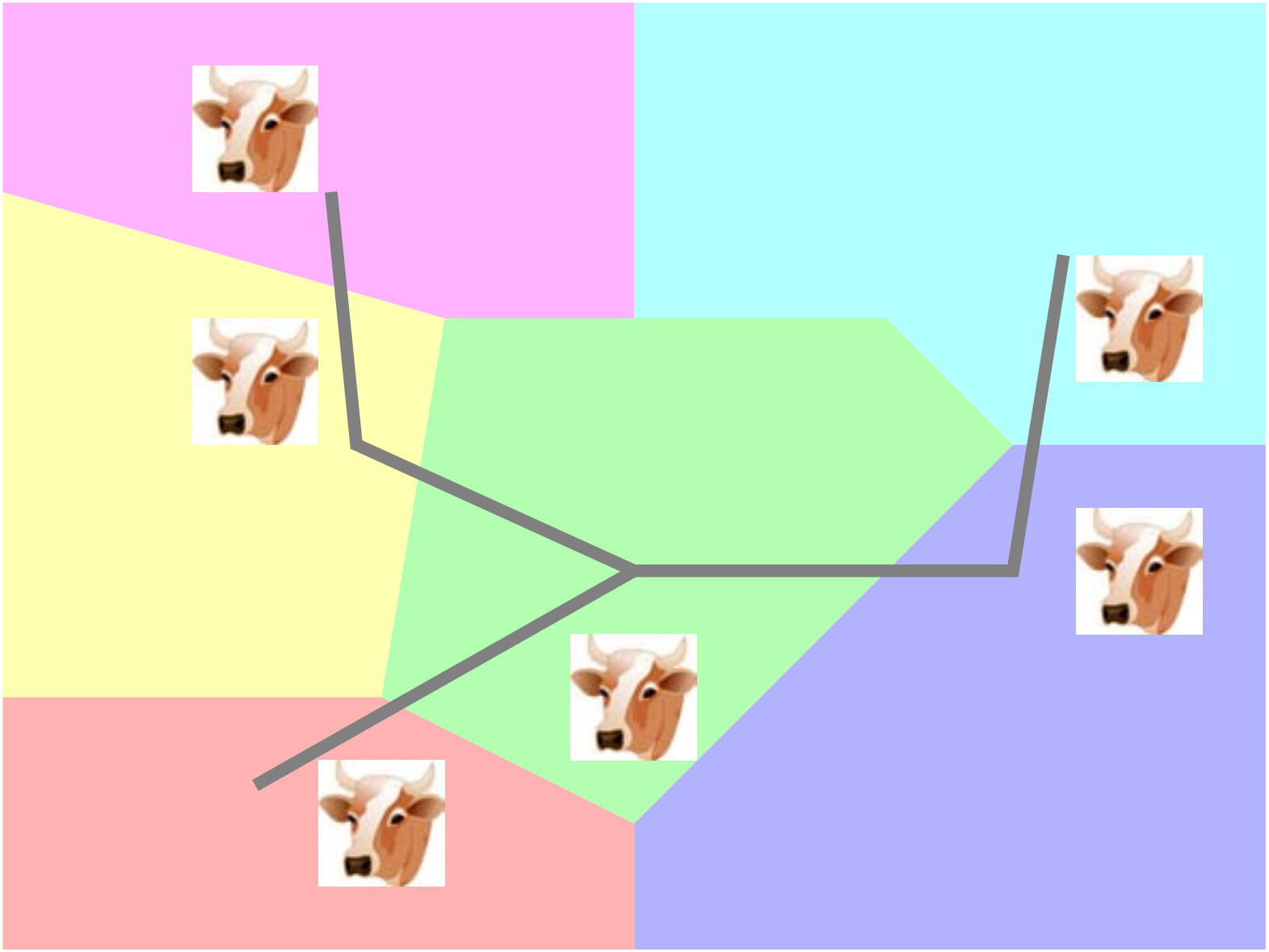
Transmission Trees



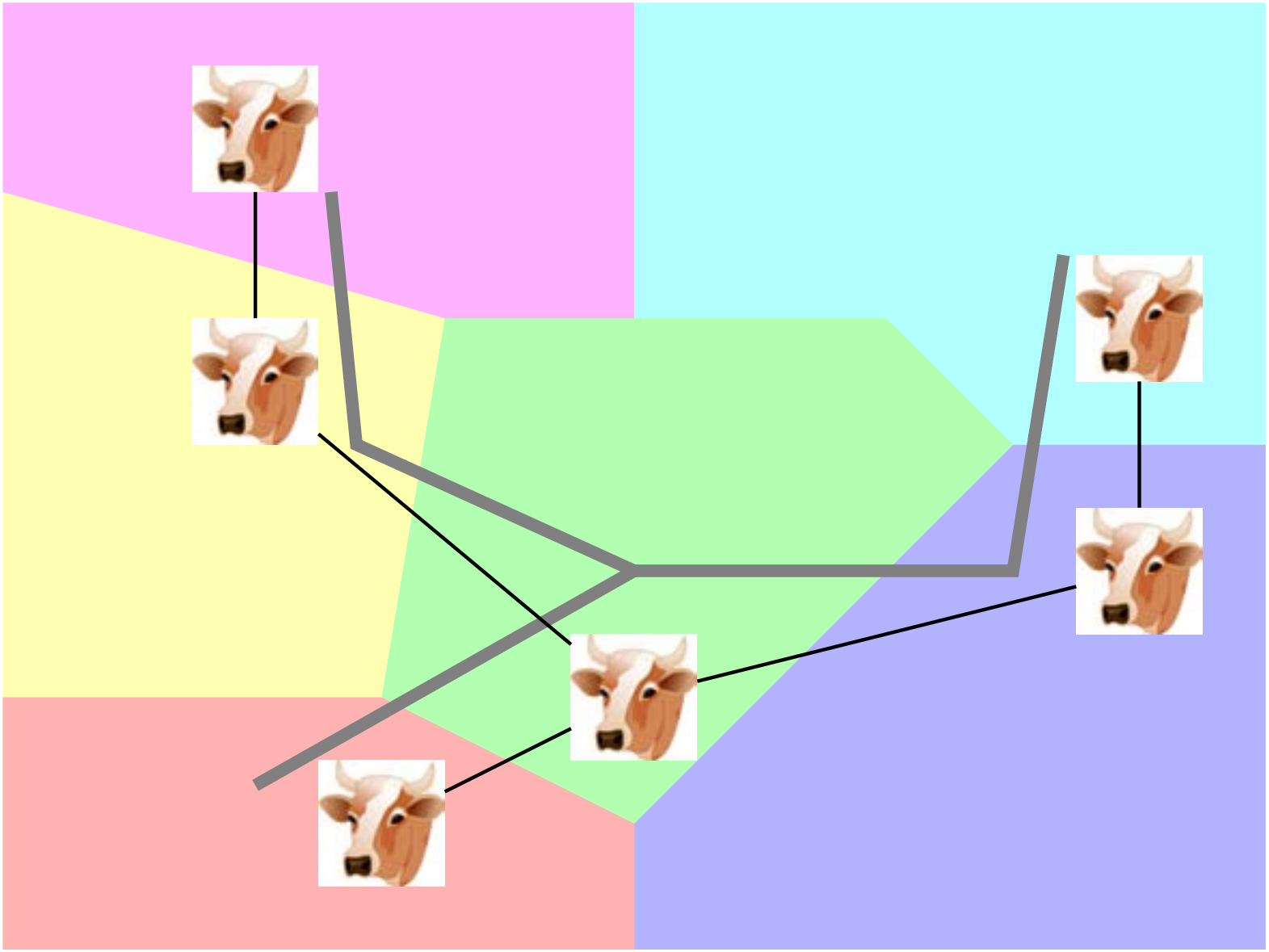
Transmission Trees



Transmission Trees



Transmission Trees

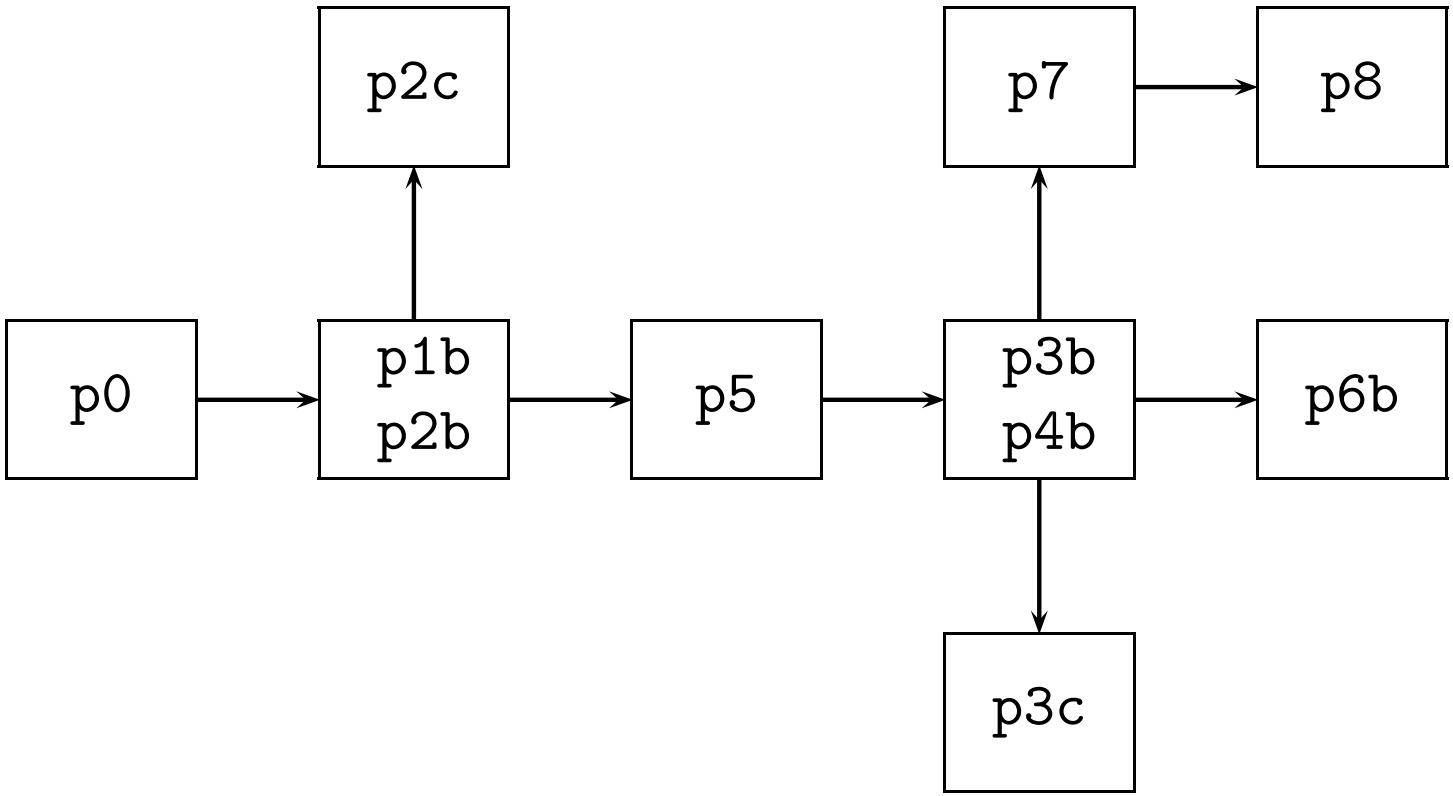


Consensus Sequence Data

site	sequences
p0 (ancestor)	2
p1b	6
p2b	7
p2c	3
p3b	8
p3c	2
p4b	2
p5	5
p6b	3
p7	8
p8	1
sum	47
combinations	967680

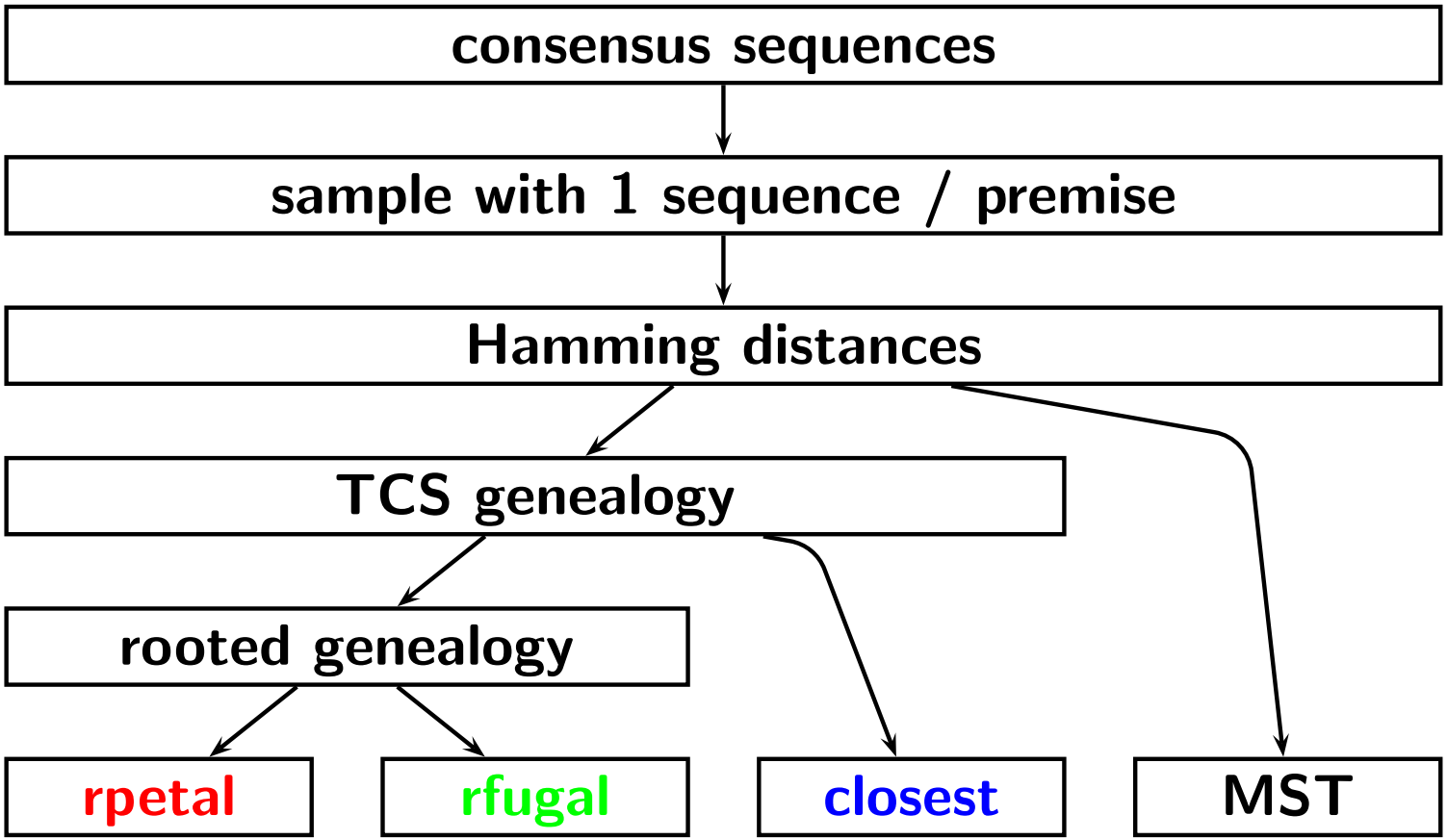
- 2007 FMD outbreak in UK
- 10 premises, 2 ancestor samples (p0)

Reference Transmission Tree



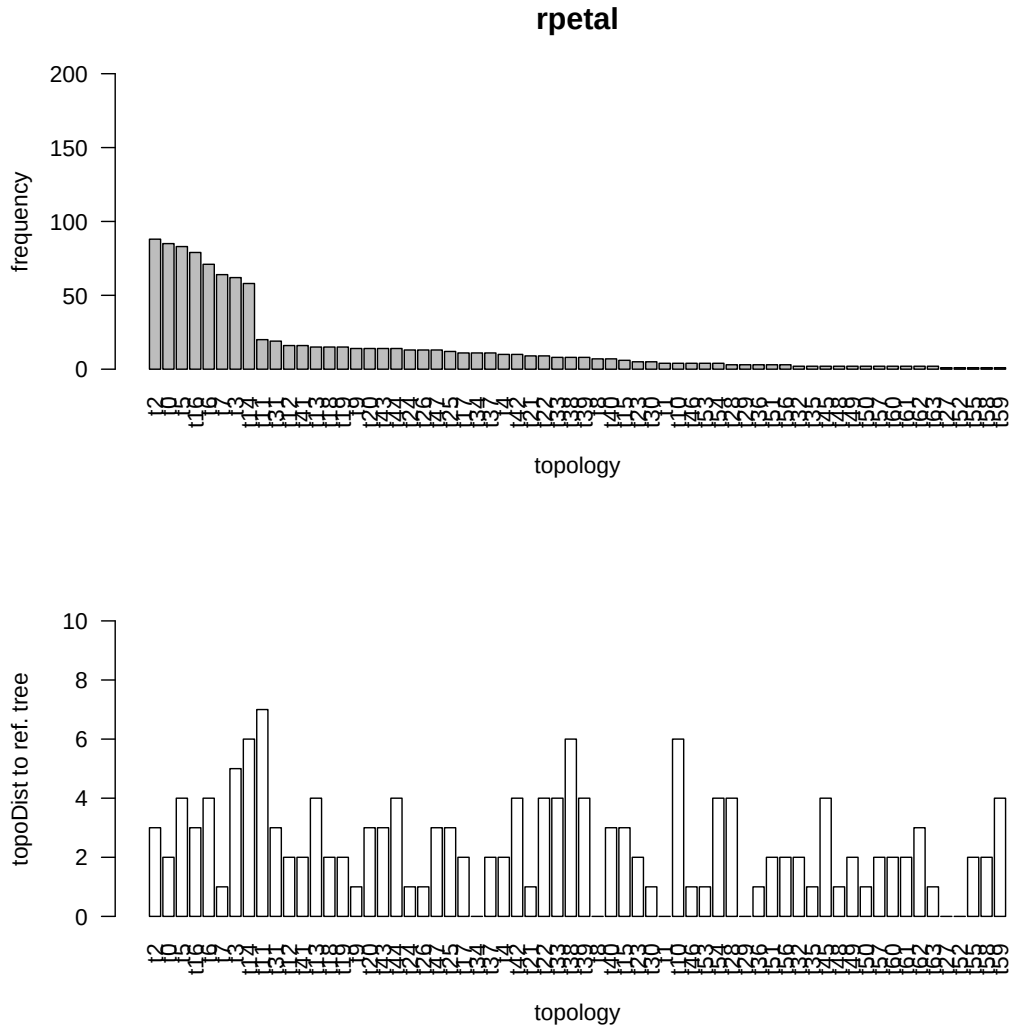
Based on a TCS geneaology of all 47 samples, and additional background knowledge.

Flowchart



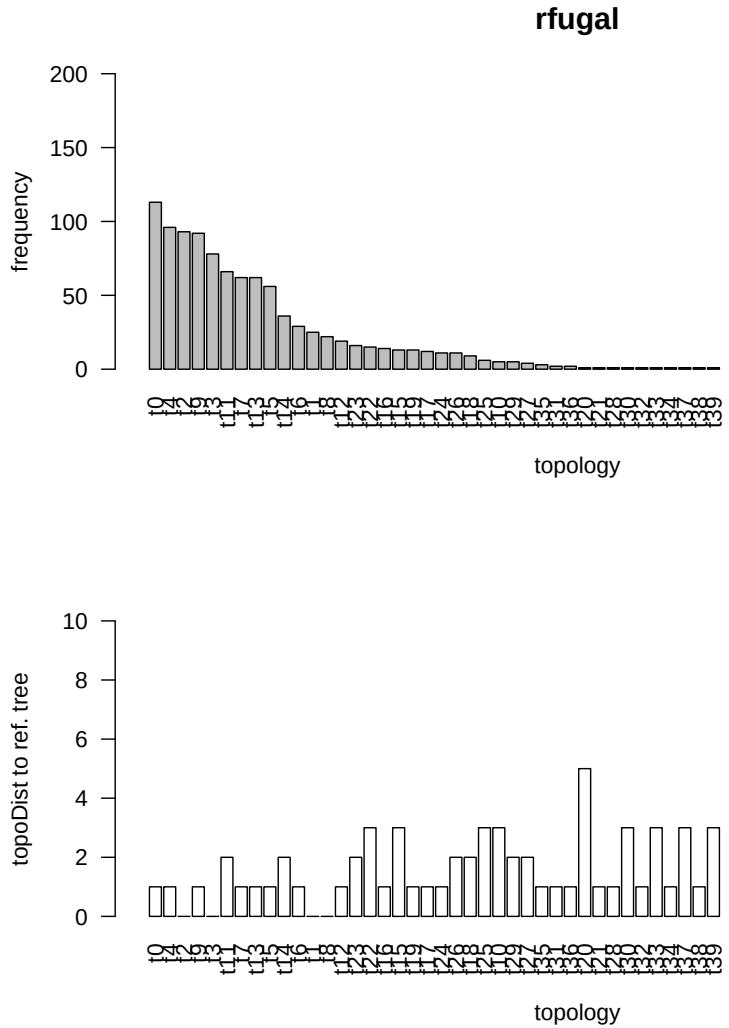
Analysis carried out for **1000 random samples** containing one consensus sequence from each site.

Results: Tree Topology



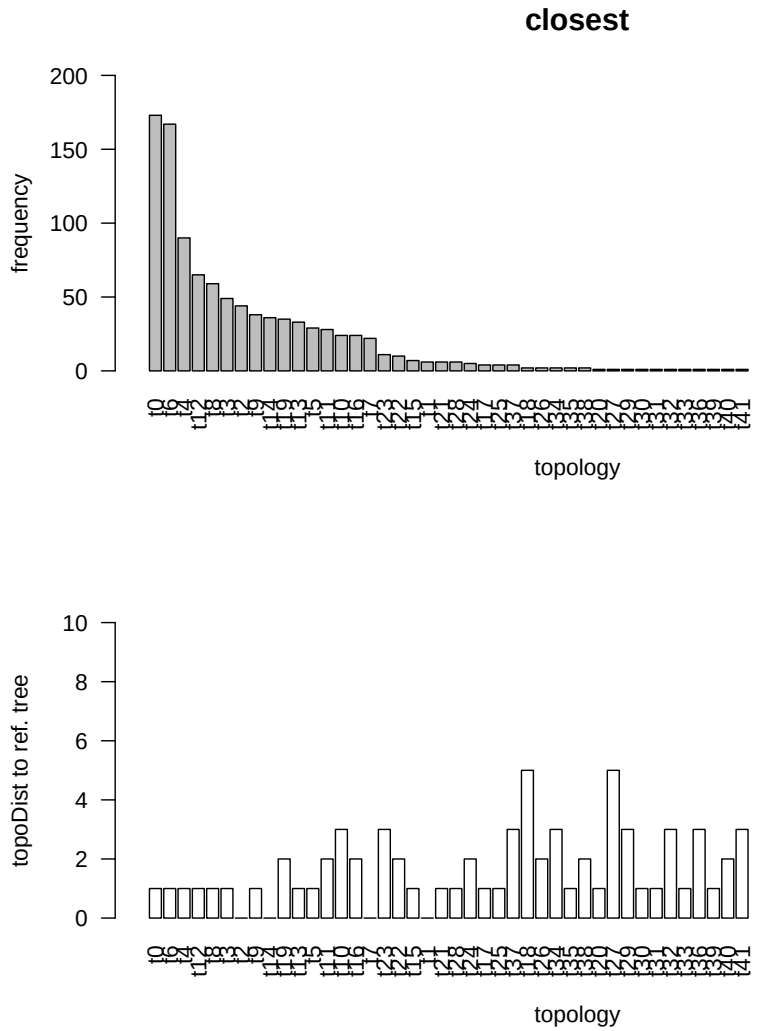
radipetal: branch nodes merged towards root (p0)

Results: Tree Topology



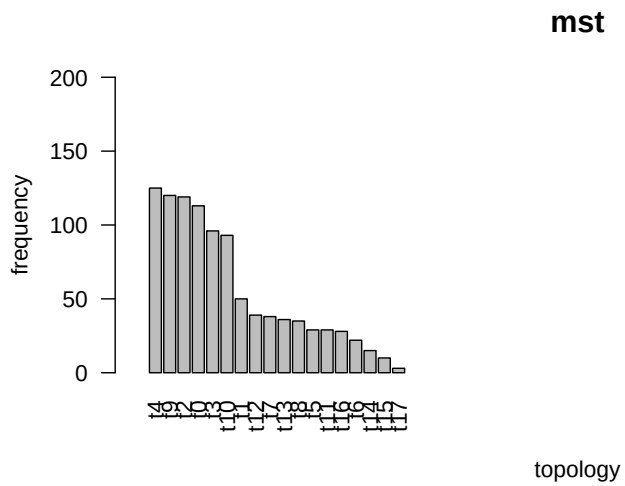
radifugal: branch nodes merged away from root (p0)

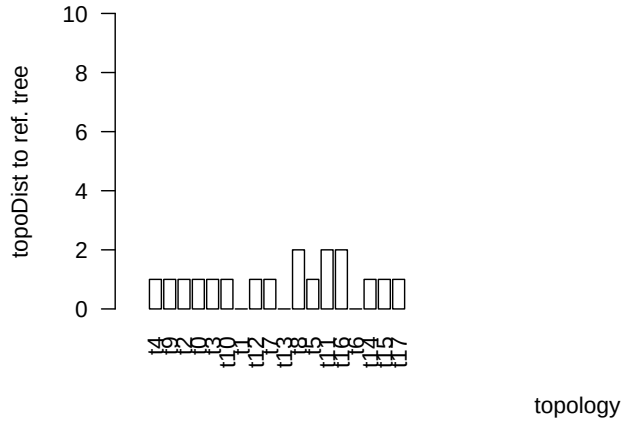
Results: Tree Topology



closest: branch nodes merged towards closest premise

Results: Tree Topology





MST

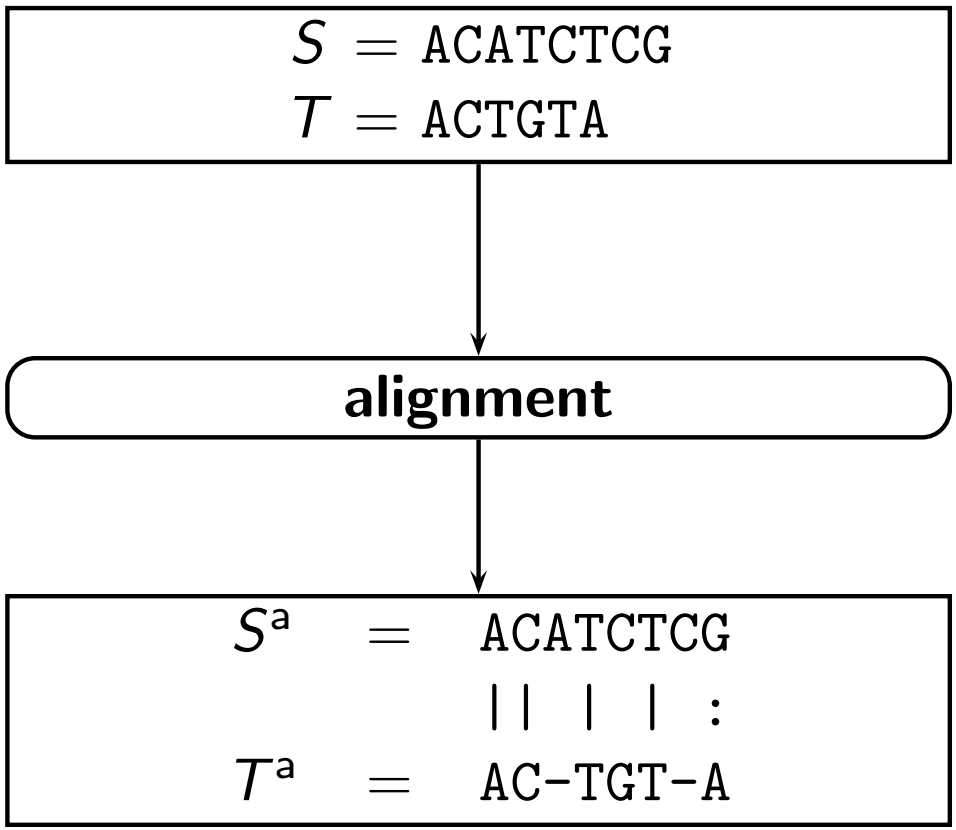
Summary: FMD Transmission Trees

- Algorithms for constructing transmission trees
  - from TCS genealogies: radipetal, radifugal, closest,
  - minimum spanning tree (MST).
- Comparison based on the 2007 outbreak.
  - closest provides TCS based transmission trees best precision.
  - MST provides even marginally better precision.
- **Outlook:**
  - Try more sophisticated distance measures.
  - Include further transmission tree reconstruction methods.
  - Larger data sets, NGS "beyond the consensus" [Wright et al., 2011]

Outline

Outline

Pairwise Alignment: Idea



Formal Definition

- **Extend** sequences S and T by inserting **gaps** to S^a and T^a.
  - aligned sequences have **equal length**: $|S^a| = |T^a|$
  - gaps cannot be paired with gaps
- Biological background: **homology**, symbols in a column should derive from same **common ancestor**.
- **Match:** column with **equal** symbols in S^a and T^a.
- **Indel:** column with a **gap symbol** in S^a or T^a.
- **Mismatch:** column with **different** symbols (non-gap) in S^a and T^a.

```
ACATCTCG
AC-TGT-A
```

Formal Definition

- **Extend** sequences S and T by inserting **gaps** to S^a and T^a.
  - aligned sequences have **equal length**: $|S^a| = |T^a|$
  - gaps cannot be paired with gaps
- Biological background: **homology**, symbols in a column should derive from same **common ancestor**.
- **Match:** column with **equal** symbols in S^a and T^a.
- **Indel:** column with a **gap symbol** in S^a or T^a.
- **Mismatch:** column with **different** symbols (non-gap) in S^a and T^a.

```
ACATCTCG
AC-TGT-A
```

Formal Definition

- **Extend** sequences S and T by inserting **gaps** to S^a and T^a.
  - aligned sequences have **equal length**: $|S^a| = |T^a|$
  - gaps cannot be paired with gaps
- Biological background: **homology**, symbols in a column should derive from same **common ancestor**.
- **Match:** column with **equal** symbols in S^a and T^a.
- **Indel:** column with a **gap symbol** in S^a or T^a.
- **Mismatch:** column with **different** symbols (non-gap) in S^a and T^a.

```
ACATCTCG
AC-TGT-A
```

Formal Definition

- **Extend** sequences S and T by inserting **gaps** to S^a and T^a.
  - aligned sequences have **equal length**: $|S^a| = |T^a|$
  - gaps cannot be paired with gaps
- Biological background: **homology**, symbols in a column should derive from same **common ancestor**.
- **Match:** column with **equal** symbols in S^a and T^a.
- **Indel:** column with a **gap symbol** in S^a or T^a.
- **Mismatch:** column with **different** symbols (non-gap) in S^a and T^a.

ACATCTCG

AC-TGT-A

Scoring of Alignments

The score $m(k)$ of column k is

- the **space penalty** $m(k) = -g$, if one symbol is the gap symbol, here: $g = 2$),
- otherwise the **pair score** $m(k) = \mu(s^{\mathrm{a}}(k), t^{\mathrm{a}}(k))$, here

$$\mu(x, y) = \begin{cases} 1, & \text{if } x = y, \\ -1, & \text{otherwise} \end{cases}$$

$S^{\mathrm{a}} =$	A	C	A	T	C	T	C	G	
$T^{\mathrm{a}} =$	A	C	-	T	G	T	-	A	
score:	$+1$	$+1$	-2	$+1$	-1	$+1$	-2	-1	$= -2$

Optimal Alignments

- **Objective:** Find the alignment with **maximal** score.
- **Problem:** The number of alignments is

$$\binom{|S|+|T|)}{|S|} \cdot \binom{|S|+|T|}{|T|}$$

- Trying out all alignments is **impossible**.
  - Recursion results in trying out all alignments.

Optimal Alignments

- **Observation:** A prefix alignment of an optimal alignment is optimal (as well).
  - Otherwise, a **contradiction** results: The optimal alignment could be improved by changing the prefix.
- **Dynamic programming:** Tabulate optimal scores of prefix alignments

```
ACATCTCG
AC-TGT-A
```

Optimal Alignments

- **Observation:** A prefix alignment of an optimal alignment is optimal (as well).
  - Otherwise, a **contradiction** results: The optimal alignment could be improved by changing the prefix.
- **Dynamic programming:** Tabulate optimal scores of prefix alignments

```
ACATCTCG
AC-TGT-A
```

Optimal Alignments

- **Observation:** A prefix alignment of an optimal alignment is optimal (as well).
  - Otherwise, a **contradiction** results: The optimal alignment could be improved by changing the prefix.
- **Dynamic programming:** Tabulate optimal scores of prefix alignments

```
ACATCTCG
AC-TGT-A
```

Optimal Alignments

- **Observation:** A prefix alignment of an optimal alignment is optimal (as well).
  - Otherwise, a **contradiction** results: The optimal alignment could be improved by changing the prefix.
- **Dynamic programming:** Tabulate optimal scores of prefix alignments

```
ACATCTCG
AC-TGT-A
```

Optimal Alignments

- **Observation:** A prefix alignment of an optimal alignment is optimal (as well).
  - Otherwise, a **contradiction** results: The optimal alignment could be improved by changing the prefix.
- **Dynamic programming:** Tabulate optimal scores of prefix alignments

```
ACATCTCG
AC-TGT-A
```

Optimal Alignments

- **Observation:** A prefix alignment of an optimal alignment is optimal (as well).
  - Otherwise, a **contradiction** results: The optimal alignment could be improved by changing the prefix.
- **Dynamic programming:** Tabulate optimal scores of prefix alignments

```
ACATCTCG
AC-TGT-A
```

Optimal Alignments

- **Observation:** A prefix alignment of an optimal alignment is optimal (as well).
  - Otherwise, a **contradiction** results: The optimal alignment could be improved by changing the prefix.
- **Dynamic programming:** Tabulate optimal scores of prefix alignments

```
ACATCTCG
AC-TGT-A
```

Optimal Alignments

- **Observation:** A prefix alignment of an optimal alignment is optimal (as well).
  - Otherwise, a **contradiction** results: The optimal alignment could be improved by changing the prefix.
- **Dynamic programming:** Tabulate optimal scores of prefix alignments

```
ACATCTCG
AC-TGT-A
```

Optimal Alignments

- **Observation:** A prefix alignment of an optimal alignment is optimal (as well).
  - Otherwise, a **contradiction** results: The optimal alignment could be improved by changing the prefix.
- **Dynamic programming:** Tabulate optimal scores of prefix alignments

```
ACATCTCG
AC-TGT-A
```

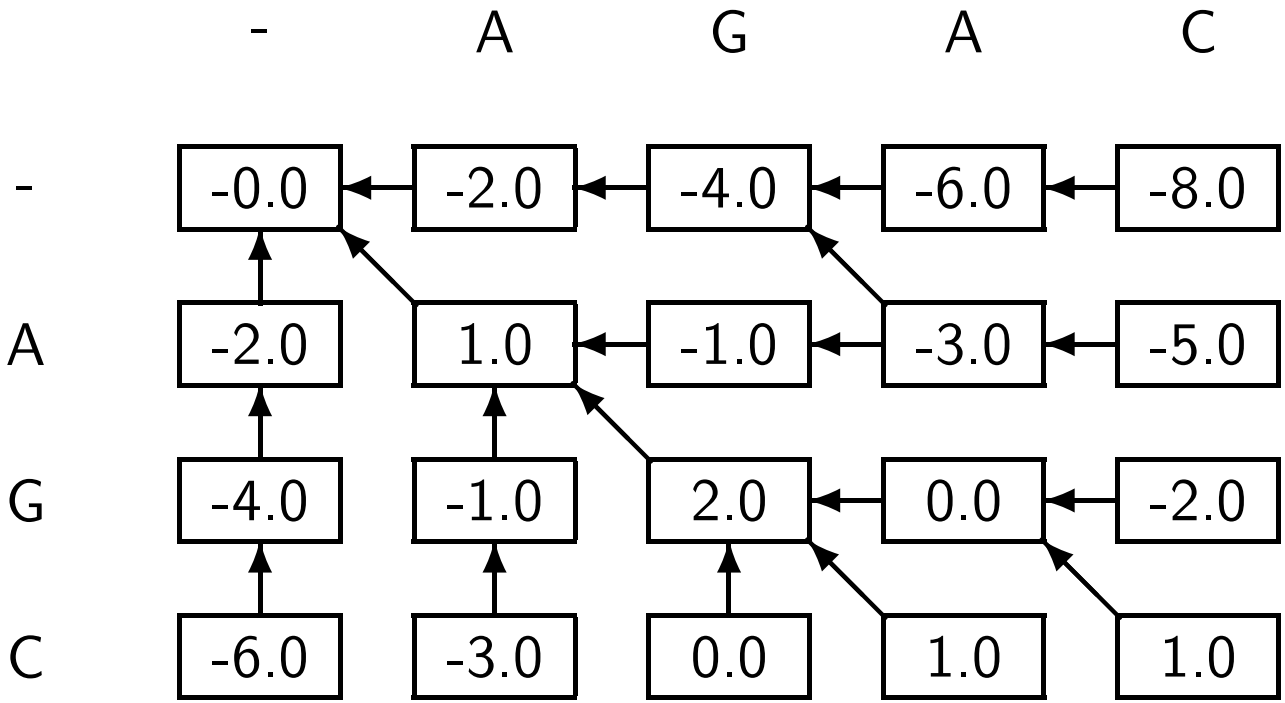
Table of Prefix-Alignment Scores

	-	A	G	A	C
-	0.0	-2.0	-4.0	-6.0	-8.0
A	-2.0	1.0	-1.0	-3.0	-5.0
G	-4.0	-1.0	2.0	0.0	-2.0
C	-6.0	-3.0	0.0	1.0	1.0

The **optimal** alignment score is 1.0.

☞Notice O(n^2) complexity.

Backtracking the Alignment



AG-C

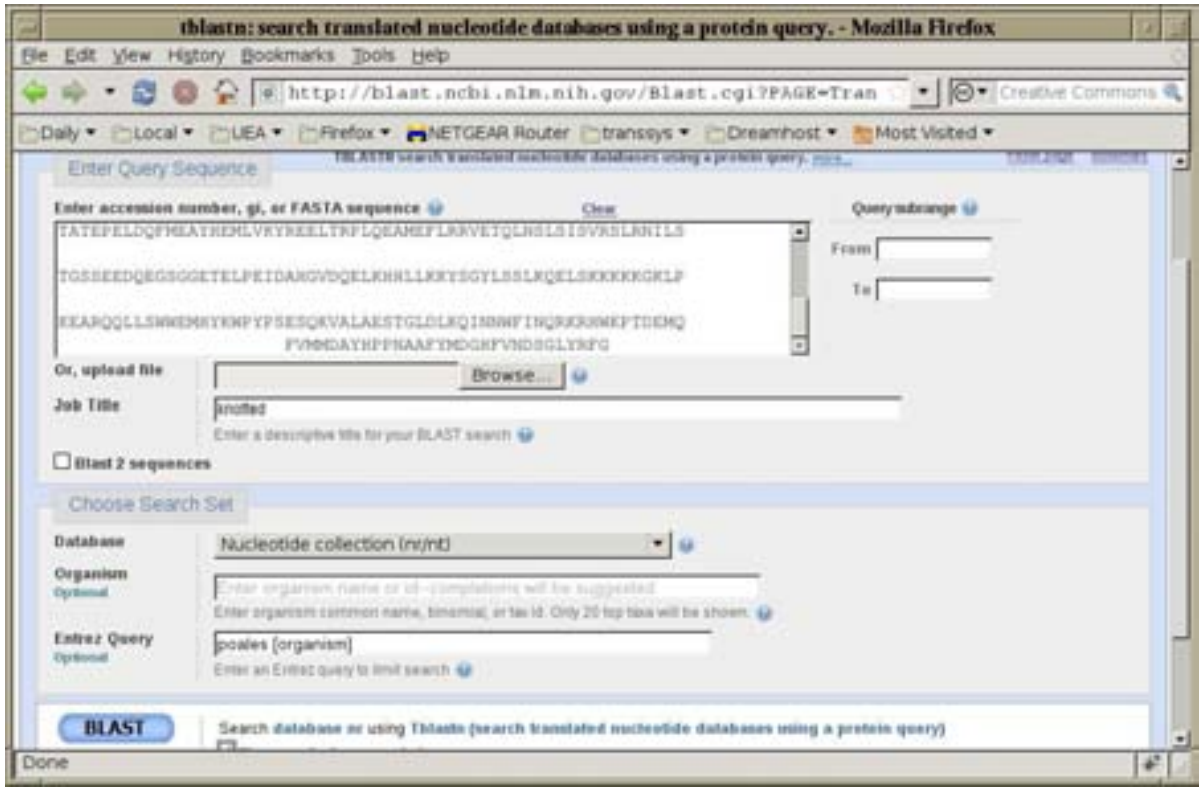
AGAC

Outline

BLAST: Basic Local Alignment Search Tool

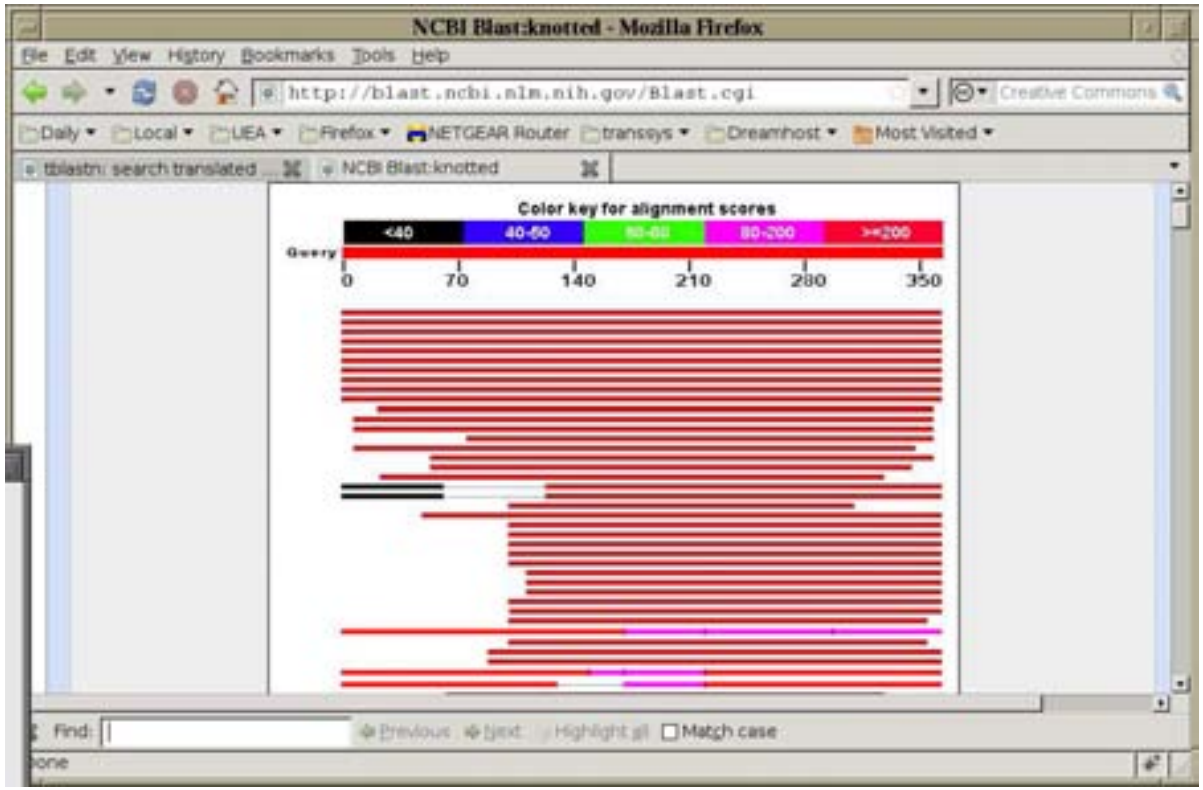
- **Objective**: Given a **query sequence**, find **similar sequences** in a database.
- Size of database prohibits pairwise alignment of query to all entries.
- Algorithm outline [Altschul et al., 1997]:
  1. **Scan** for **hits**, i.e. gapless short word alignments exceeding a threshold score.
  2. **Extend** hits maximally to obtain **HSPs** (high scoring pairs.
  3. **Combine** HSPs to (gapped) alignments.
- E-values indicate **expected number** of HSPs with given score.
  - Interesting HSPs have $E \ll 1$.

BLAST: Search Engine for Sequences



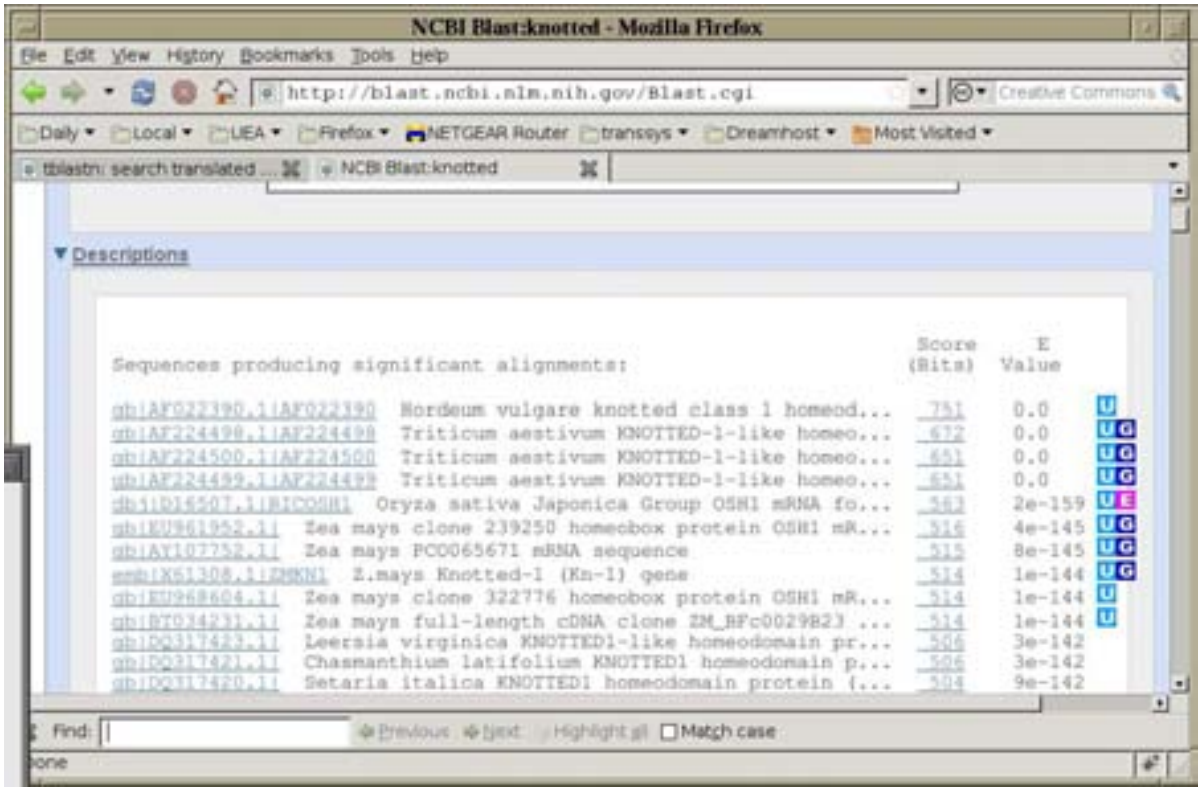
http://blast.ncbi.nlm.nih.gov/

BLAST: Search Engine for Sequences



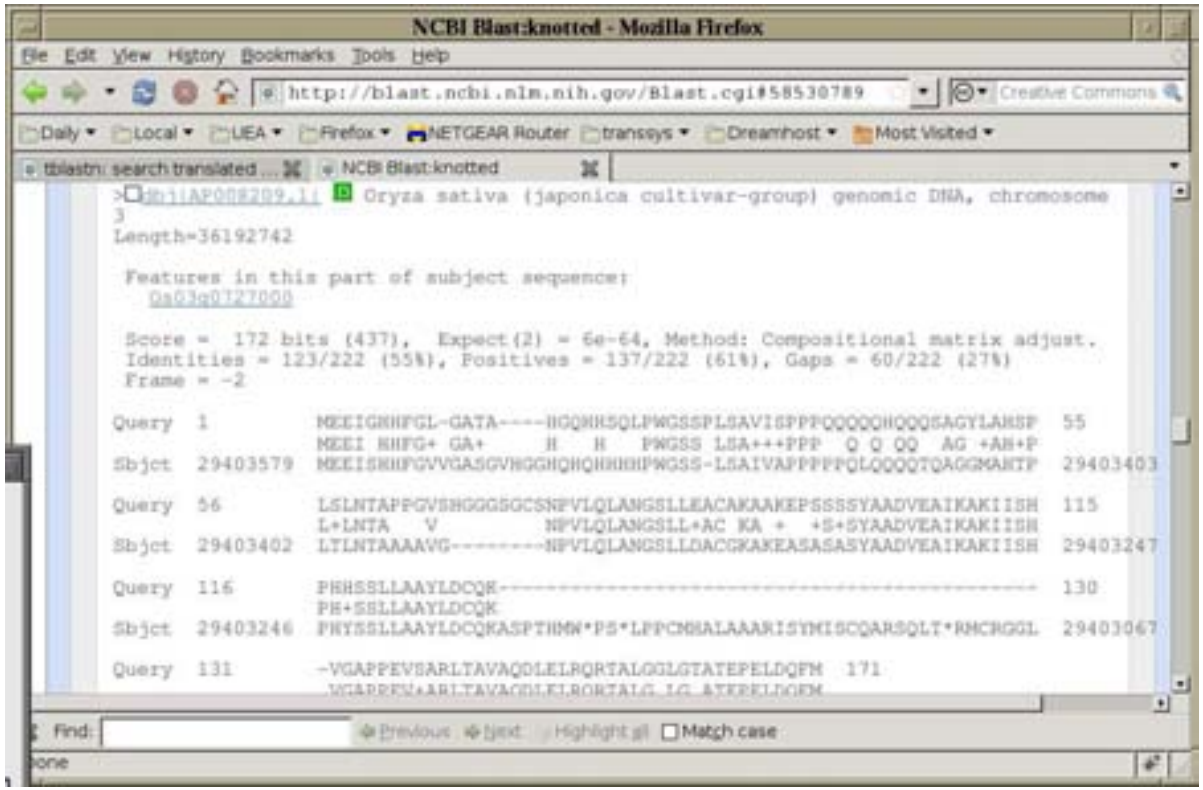
http://blast.ncbi.nlm.nih.gov/

BLAST: Search Engine for Sequences



`http://blast.ncbi.nlm.nih.gov/`

BLAST: Search Engine for Sequences



http://blast.ncbi.nlm.nih.gov/

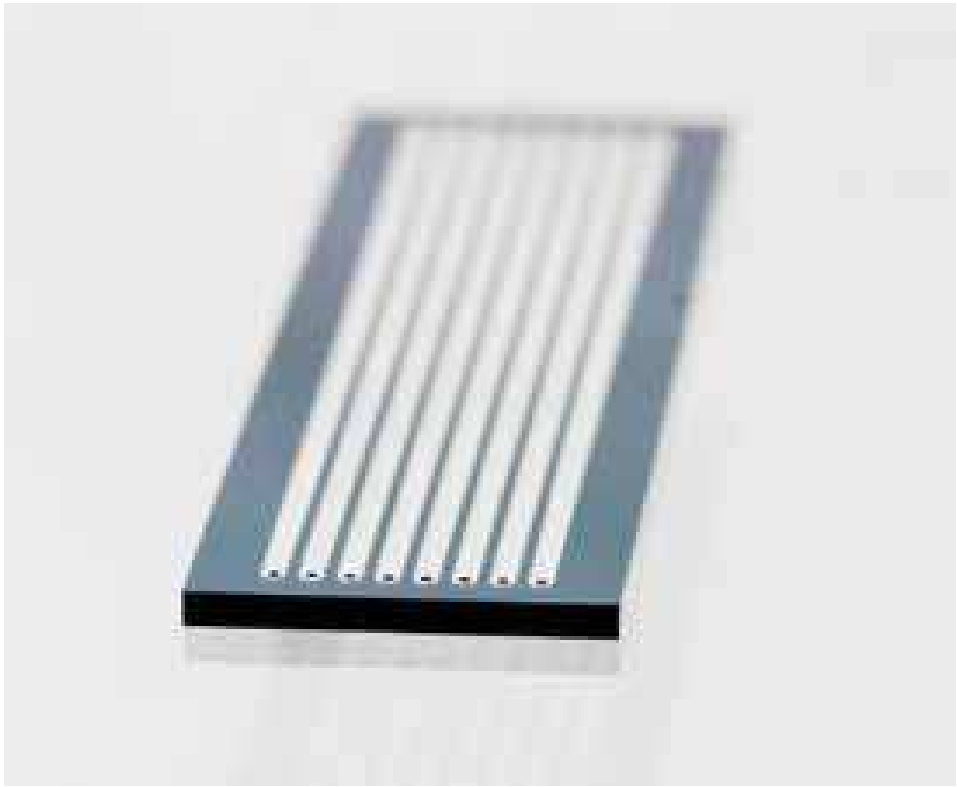
Outline

"Next Generation" Sequencing

- Long DNA sequences cannot be read like a tape.
- Short fragments from random genomic locations can be sequenced.
- NGS generates **very large** number of (very) **short** sequencing reads.

http://www.illumina.com/systems/miseq.ilmn

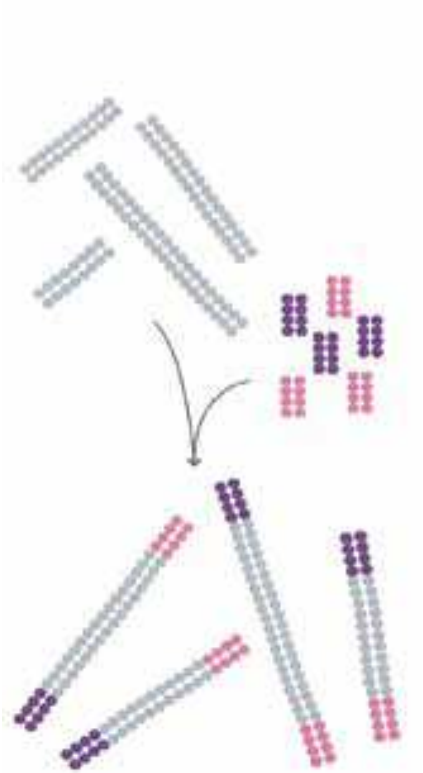
Illumina NGS Sequencing



Massive numbers of sequencing reactions take place in one **flow cell**.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

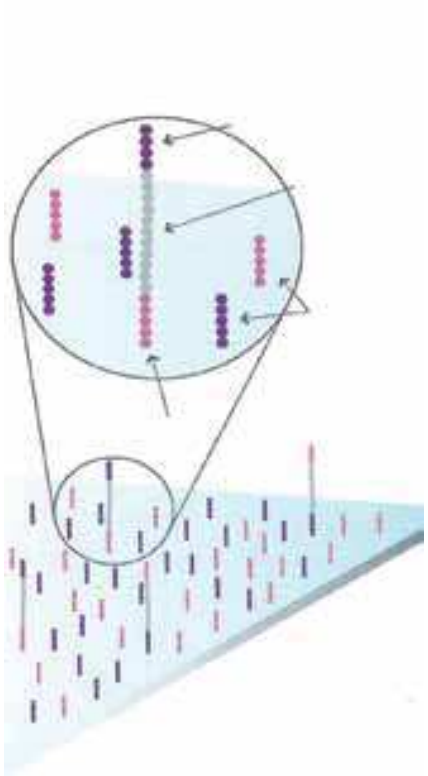
Illumina NGS Sequencing



DNA is **fragmented** and **adapters** are ligated.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

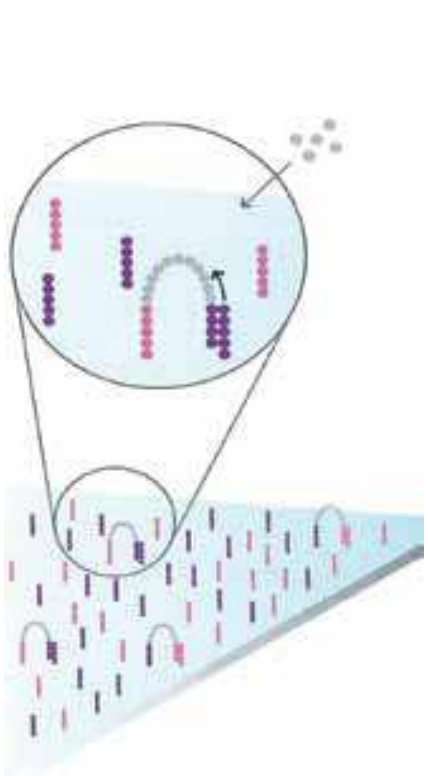
Illumina NGS Sequencing



Fragments (with adapters) are **attached** to the slide in a flow cell.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

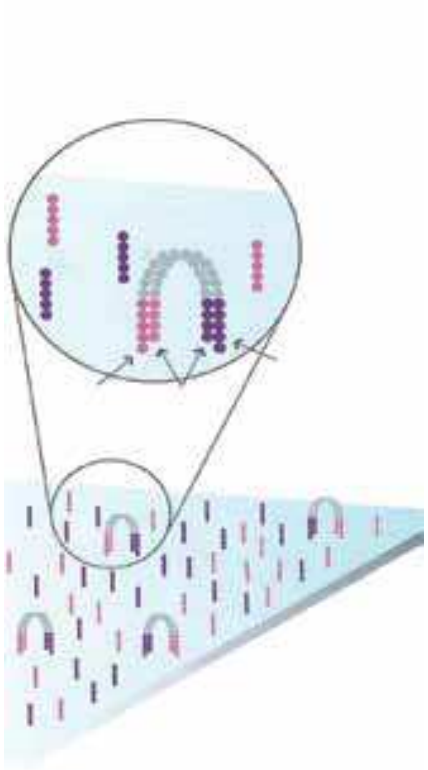
Illumina NGS Sequencing



The slide is studded with **primers**, facilitating **bridge amplification** . . .

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

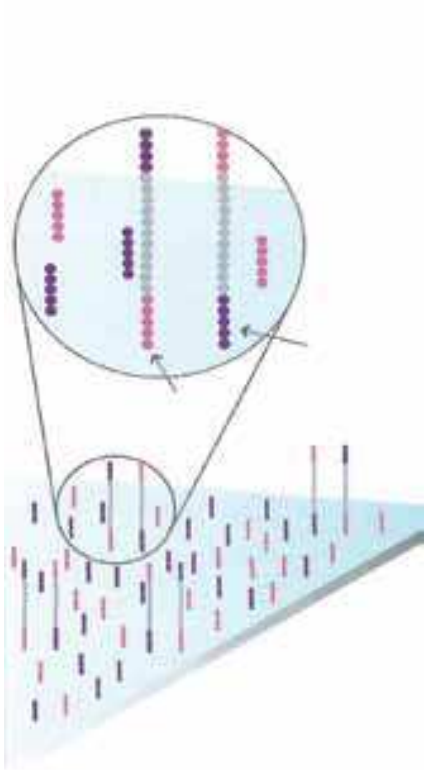
Illumina NGS Sequencing



. . . resulting in **double stranded fragments** . . .

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

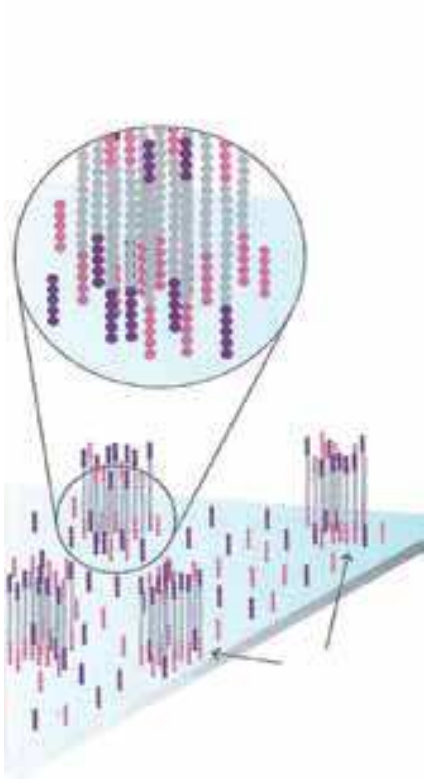
Illumina NGS Sequencing



. . . which are then **denatured**.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

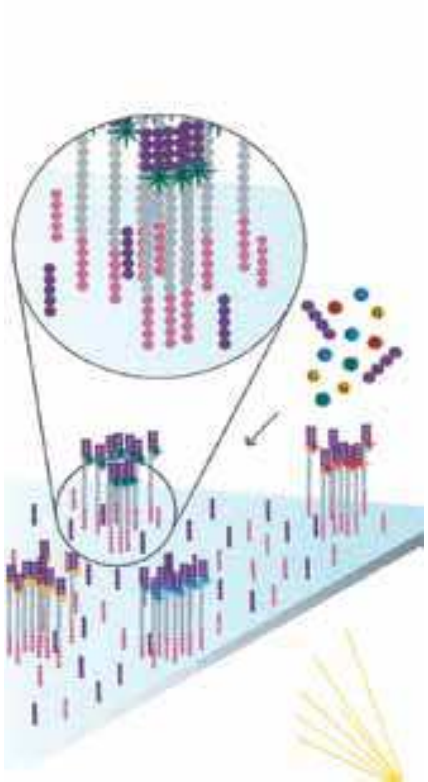
Illumina NGS Sequencing



Multiple rounds of amplification result in a **cluster** from each initial fragment.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

Illumina NGS Sequencing



Reversible terminator nucleotides are added.

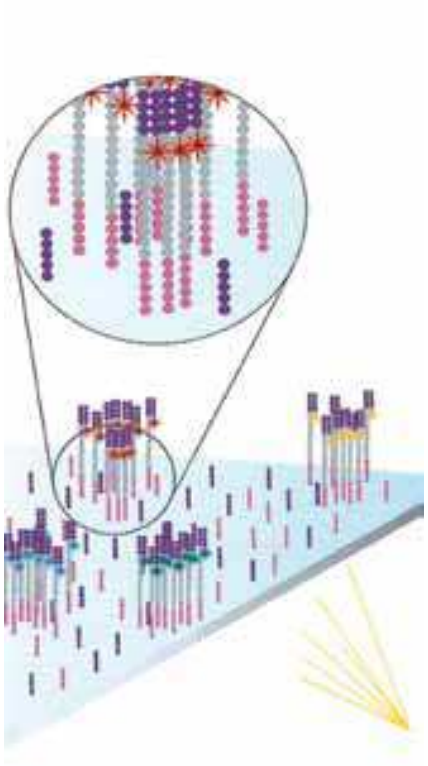
`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

Illumina NGS Sequencing

Incorporated nucleotide fluoresce at different wave lengths.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

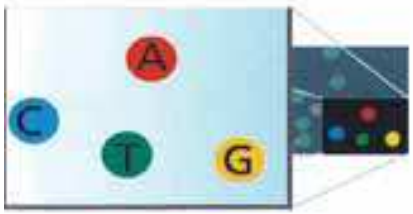
Illumina NGS Sequencing



After removal of the terminator, the next nucleotide is added ...

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

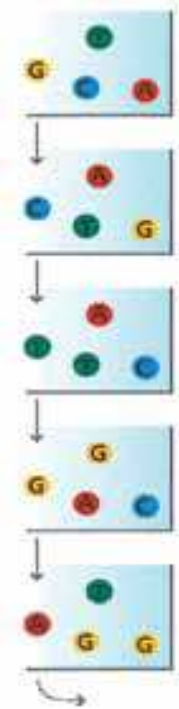
Illumina NGS Sequencing



. . . and the fluorescent light is imaged.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

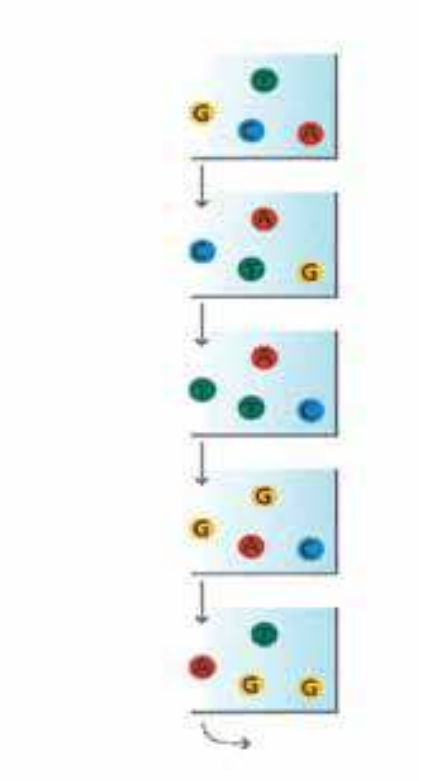
Illumina NGS Sequencing



Each image yields one base **for each cluster**.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

Illumina NGS Sequencing

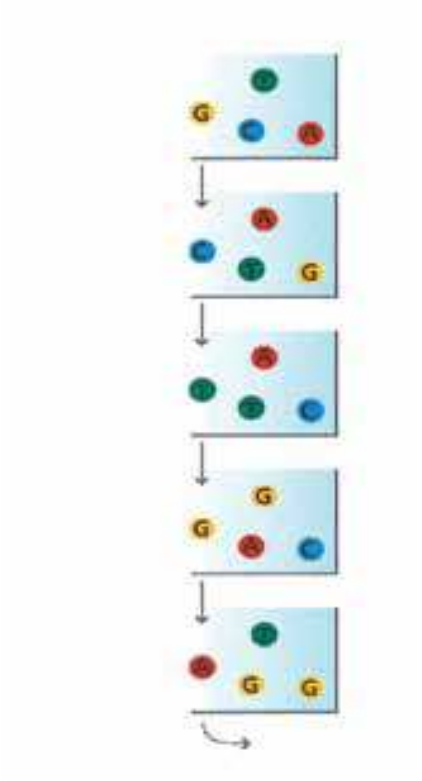


```
read0 =
read1 =
read2 =
read3 =
```

Each image yields one base **for each cluster**.

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illumina NGS Sequencing

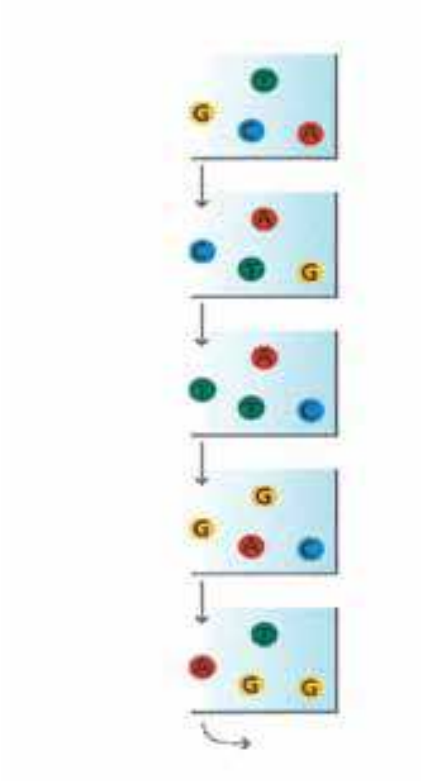


```
read0 = G
read1 = T
read2 = C
read3 = A
```

Each image yields one base **for each cluster**.

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illumina NGS Sequencing

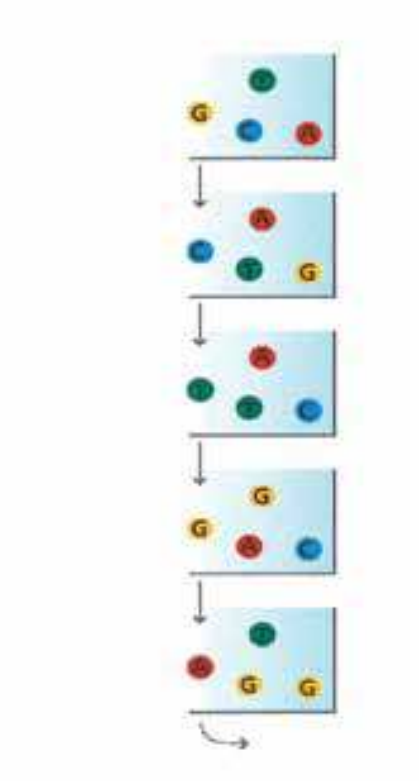


```
read0 = GC
read1 = TA
read2 = CT
read3 = AG
```

Each image yields one base **for each cluster**.

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illumina NGS Sequencing

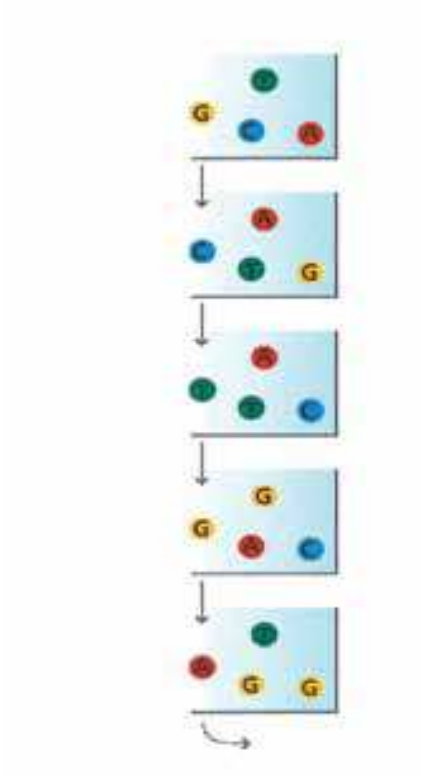


```
read0 = GCT
read1 = TAA
read2 = CTT
read3 = AGC
```

Each image yields one base **for each cluster**.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

Illumina NGS Sequencing

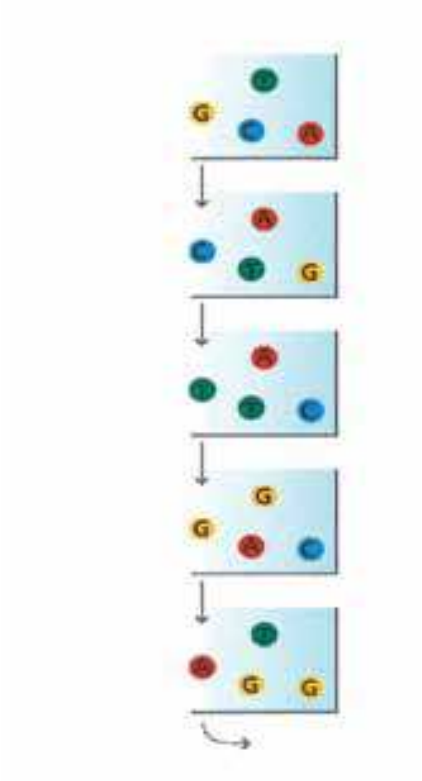


```
read0 = GCTG
read1 = TAAG
read2 = CTTA
read3 = AGCC
```

Each image yields one base **for each cluster**.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

Illumina NGS Sequencing

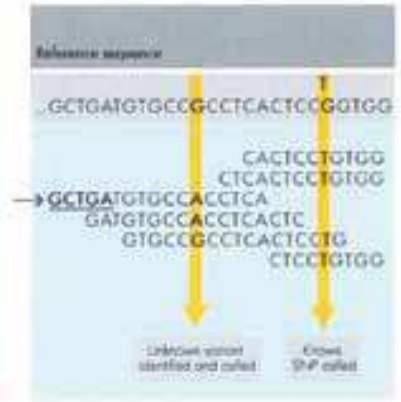


```
read0 = GCTGA
read1 = TAAGT
read2 = CTTAG
read3 = AGCCG
```

Each image yields one base **for each cluster**.

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Illumina NGS Sequencing



Reads are further processed, e.g. in sequence assembly.

`http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`

Applications of "Next Generation" Sequencing

- Novel genome sequencing
- Re-sequencing to discover **genomic variation**
  - Single nucleotide polumorphisms (SNPs), and their association to pheonotypic traits,
  - Evolution of genomic variation patterns.
- Metagenomics
- *-Seq techniques
  - gene expression measurement: RNA-Seq
  - binding sites: ChIP-Seq
  - microRNA-Seq

Mapping NGS Reads

- Task: align billions of to a known reference genome.
- Not feasible using dynamic programming.
- Feasible using advanced **indexing** of the reference.
  - e.g. Burrows-Wheeler transform
  - Pigeonhole principle

```
GACTAGAGTAGACGATGAGACCCATGACA

GGC  GAGTAGACGAT  GACCCATGATA
GGCT GAGTAGCCGATG   CCCATGACA
GGCT  AGTAGCCGATGAG CTCATGACA
GGCTAG GTAGACGATGAGA  CATGACA
GGCTAGA  AGACGATGAGA   ATGACA
GGCTAGAG AGCCGATGAGACC ATGACA
GGCTAGAGT GACGATGAGACCC
          CCGATGAGACCCAT
```

Finding Single Nucleotide Polymorphisms (SNPs)

GACTAGAGTAGACGATGAGACCCATGACA

GGC GAGTAGACGAT GACCCATGATA
GGCT GAGTAGCCGATG CCCATGACA
GGCT AGTAGCCGATGAG CTCATGACA
GGCTAG GTAGACGATGAGA CATGACA
GGCTAGA AGACGATGAGA ATGACA
GGCTAGAG AGCCGATGAGACC ATGACA
GGCTAGAGT GACGATGAGACCC
CCGATGAGACCCAT

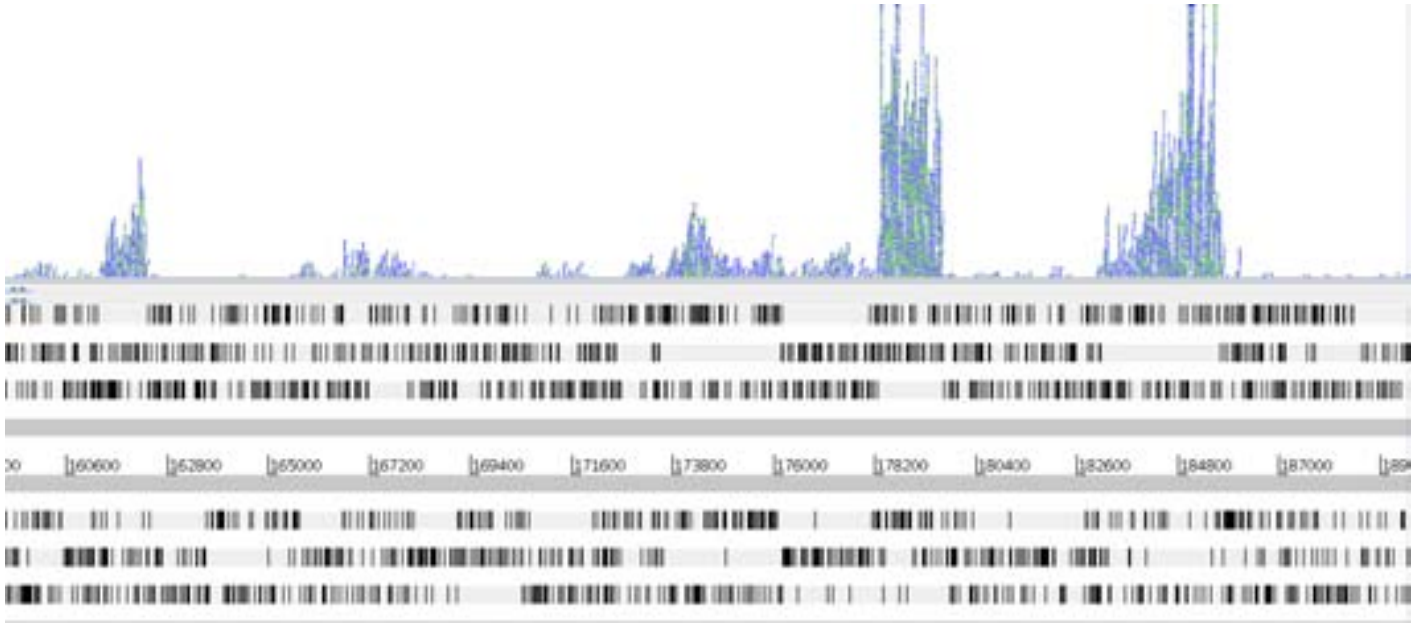
Finding Single Nucleotide Polymorphisms (SNPs)

GACTAGAGTAGACGATGAGACCCATGACA

GGC GAGTAGACGAT GACCCATGATA
GGCT GAGTAGCCGATG CCCATGACA
GGCT AGTAGCCGATGAG CTCATGACA
GGCTAG GTAGACGATGAGA CATGACA
GGCTAGA AGACGATGAGA ATGACA
GGCTAGAG AGCCGATGAGACC ATGACA
GGCTAGAGT GACGATGAGACCC
CCGATGAGACCCAT

Aligning Reads to Reference Sequences

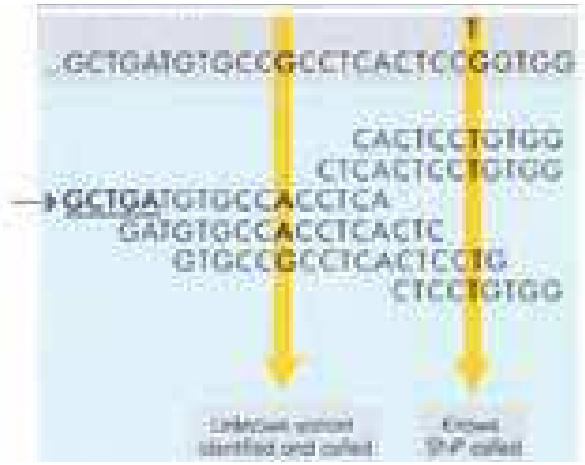
Example: RNA-Seq



Assembling NGS Reads

```
GCTGATGTGCCGCCTCACTCCGGTGG

                    CACTCCGGTGG
                 CTCACTCCTGTGG
GCTGATGTGCCACCTCA
   GATGTGCCGCCTCACTC
      GTGCCACCTCACTCCGG
                     CTCCGGTGG
```



- **Many copies** of a genome are **fragmented**
- Each base has **quality**, giving its **probability** of being correct.

NGS Assembly: Example

C A G A G C A
99 99 99 98 95 96 93

C A G A G C A G A C A
99 99 99 99 99 99 97 95 94 96 89

A G A C A A C T A A G T
99 99 99 99 99 45 26 57 87 85 84 78

A A G T G C T A T C A
99 99 99 99 99 98 99 96 91 88 82

C T A T C A A C T
99 99 99 99 99 99 96 94 95

T A T C A A C T A G
99 99 99 99 99 94 97 95 91 88

A A C T A G
99 99 99 98 91 93

NGS Assembly: Example

C A G A G C A
99 99 99 98 95 96 93

C A G A G C A G A C A
99 99 99 99 99 99 97 95 94 96 89

A G A C A A C T A A G T
99 99 99 99 99 45 26 57 87 85 84 78

A A G T G C T A T C A
99 99 99 99 99 98 99 96 91 88 82

C T A T C A A C T
99 99 99 99 99 99 96 94 95

T A T C A A C T A G
99 99 99 99 99 94 97 95 91 88

A A C T A G
99 99 99 98 91 93

C A G A G C A G A C A A C T A A G T G C T A T C A A C T A G

NGS Assembly: Example

C A G A G C A
99 99 99 98 95 96 93

C A G A G C A G A C A
99 99 99 99 99 99 97 95 94 96 89

A G A C A A C T A A G T
99 99 99 99 99 45 26 57 87 85 84 78

A A G T G C T A T C A
99 99 99 99 99 98 99 96 91 88 82

C T A T C A A C T
99 99 99 99 99 99 96 94 95

T A T C A A C T A G
99 99 99 99 99 94 97 95 91 88

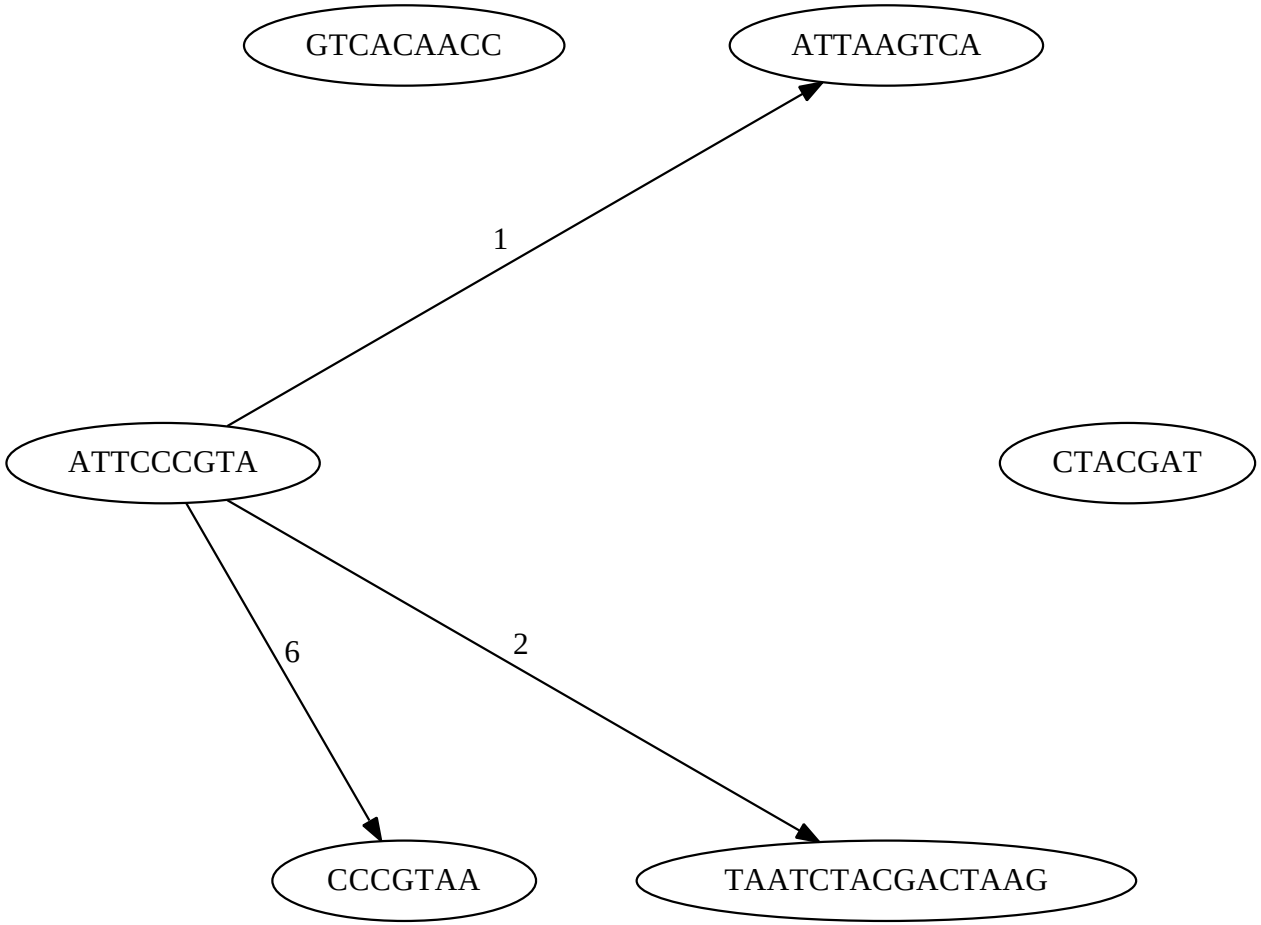
A A C T A G
99 99 99 98 91 93

C A G A G C A G A C A A N T A A G T G C T A T C A A C T A G

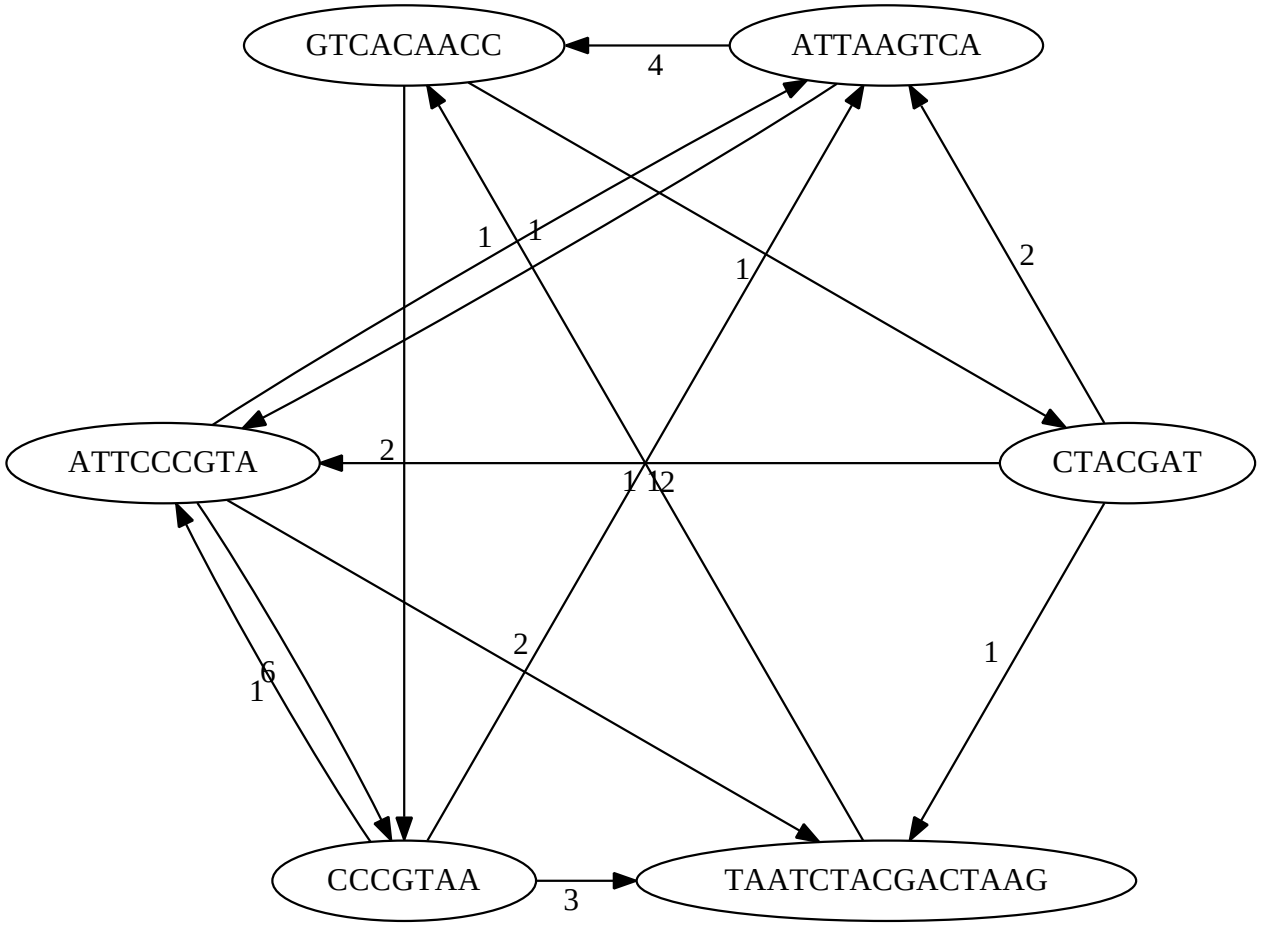
NGS Sequence Assembly

- Assembly depends on **overlaps** among reads.
- **Quality** of bases must be taken into account.
- Reads that are too short are **not informative**.
- **Repetitive sequences** make assembly difficult.
- Insufficient **depth** results in **multiple contigs**.
- Sufficient depth is a key success factor:
  - **Joining of contigs** depends on sufficient overlap (N_{50} value).
  - Resolving low quality bases depends on depth.
  - Depth does not help resolve repetitive sequences.

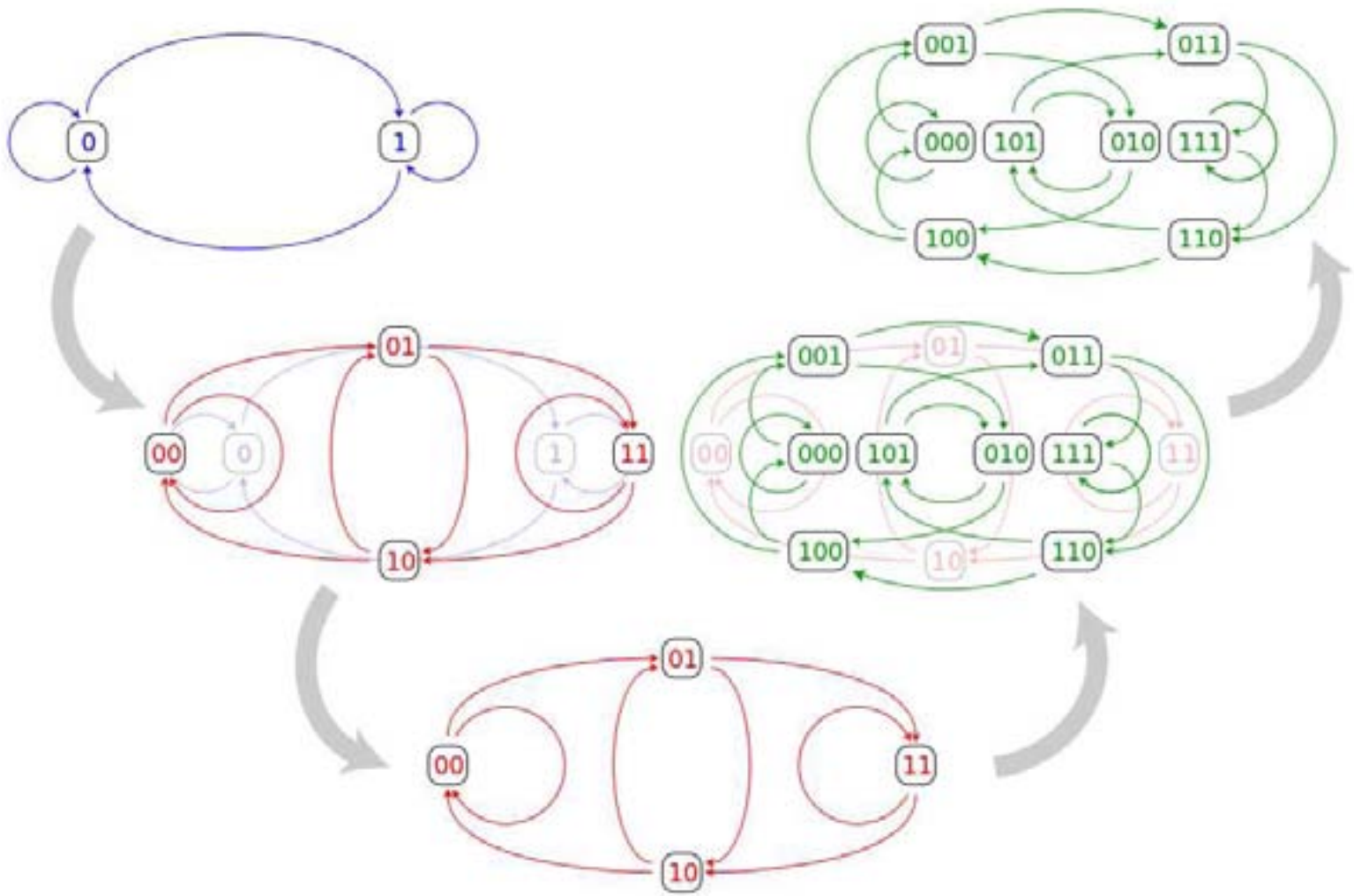
NGS Assembly: Overlap Approach



NGS Assembly: Overlap Approach

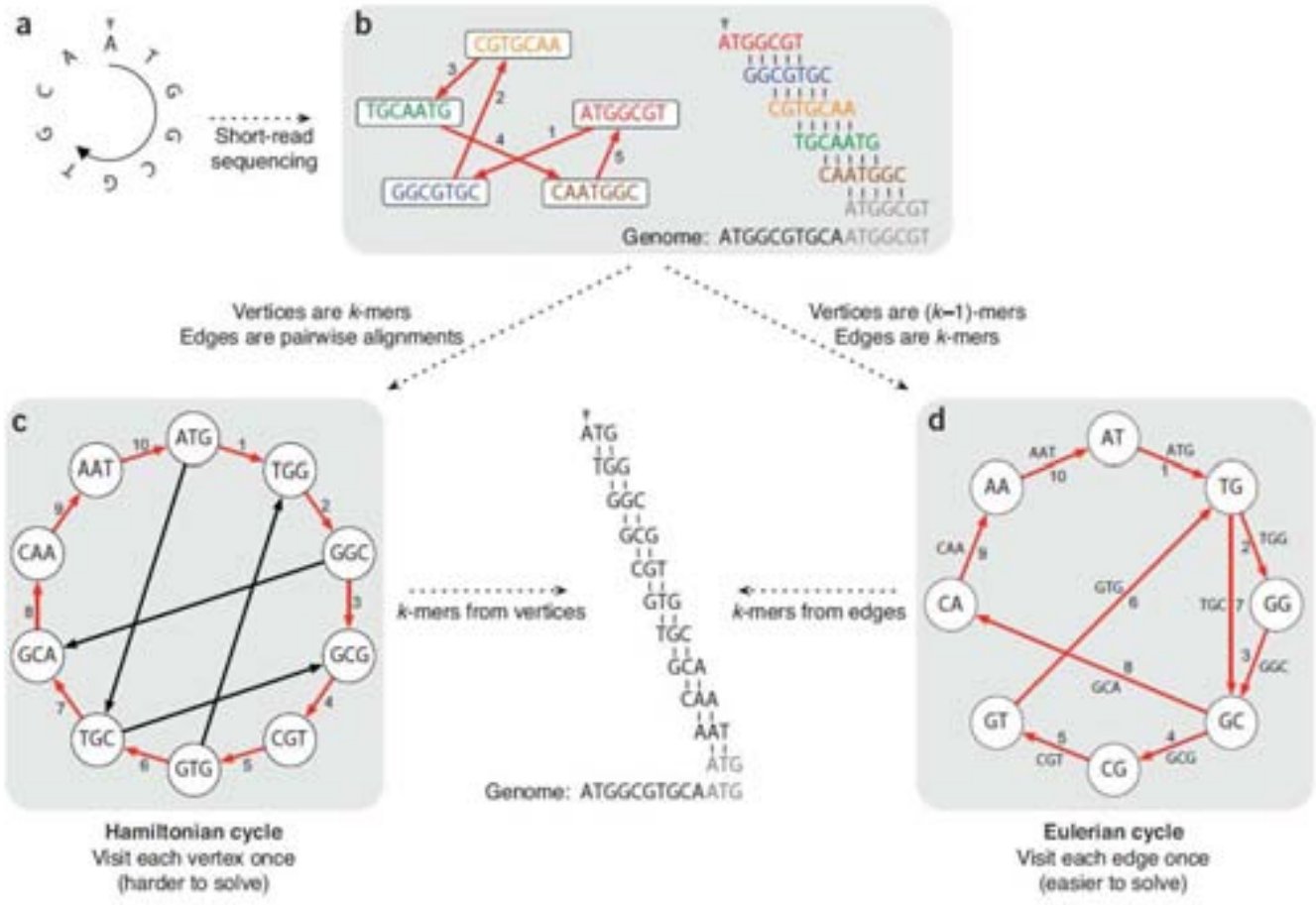


De Bruijn Graph



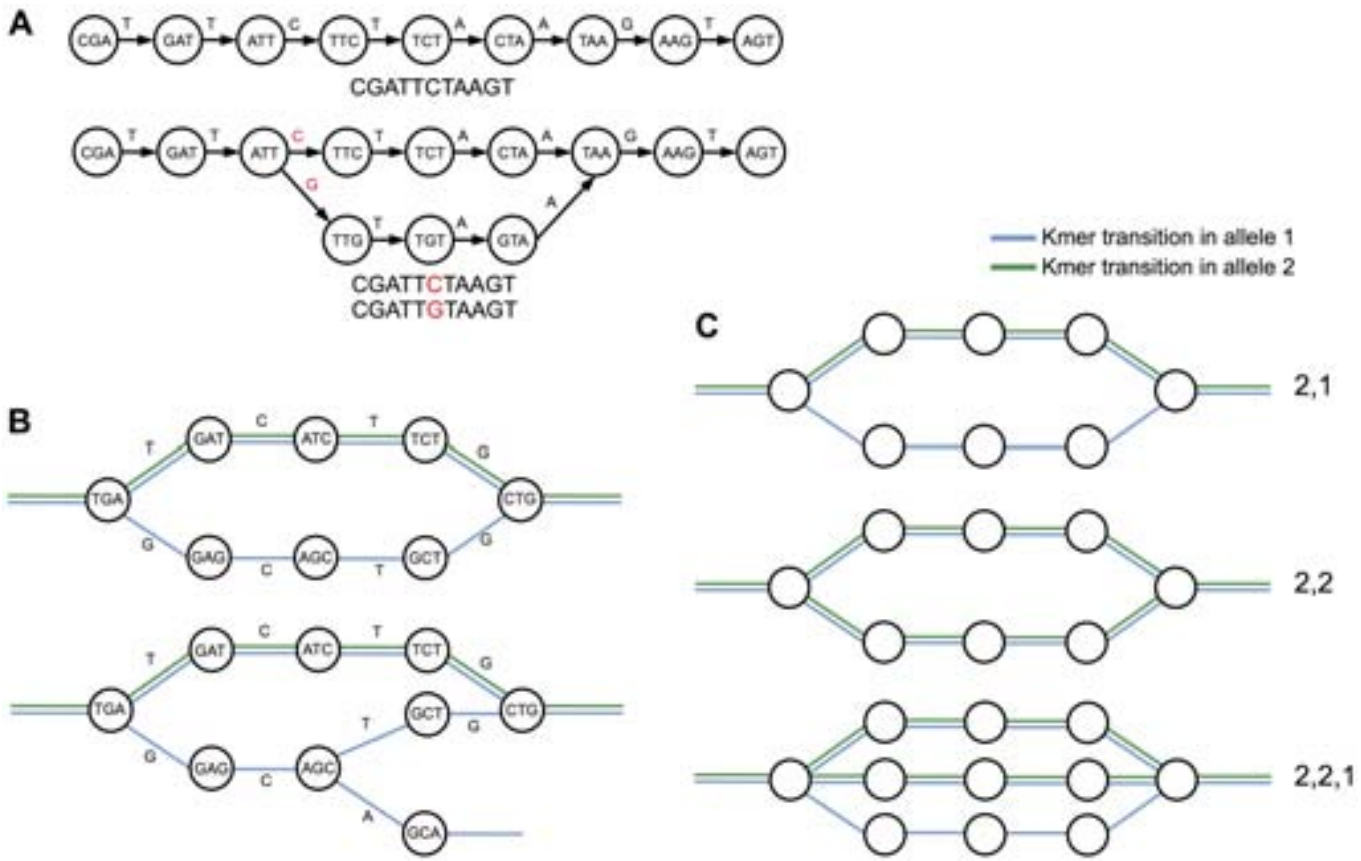
http://commons.wikimedia.org/wiki/File:DeBruijn-as-line-digraph.svg

NGS Assembly: De Bruijn Approach



Compeau, Pevzner & Tesler, Nature Computational Biology 29 (2011): 987–991, Fig. 3

Polymorphisms and de Bruijn Assembly



[Leggett et al., 2013, Fig. 1]

Summary ¡: NGS Data Analysis

- **Mapping** to a reference sequence, using **indexing**
  - resequencing
  - detection of SNPs and other variants,
  - identification of genes (RNA-seq).
- **De novo assembly** of genomes or transcriptomes.
  - Resource intensive (particularly memory)
  - Overlap assembly: feasible with smaller sets
  - De Bruijn graph assembly of k-mers
- NGS metagenomics . . .

⚠ Software has limitations and is evolving rapidly. ⚠

Outline

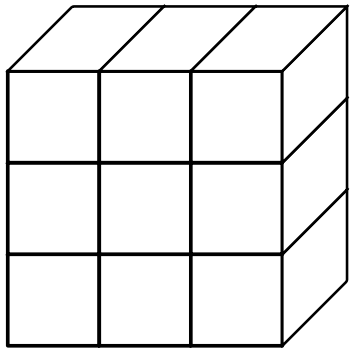
Multiple Alignment

- Extend pairwise approach?
  - 2 sequences: table of n^2 prefix alignments
  - 3 sequences: table of n^3 prefix alignments
  - Warning: Very large numbers ahead

- Aligning 100 sequences of 300 symbols: about 10^{170} prefix alignments.
  - How much computing time does the universe have?

Multiple Alignment

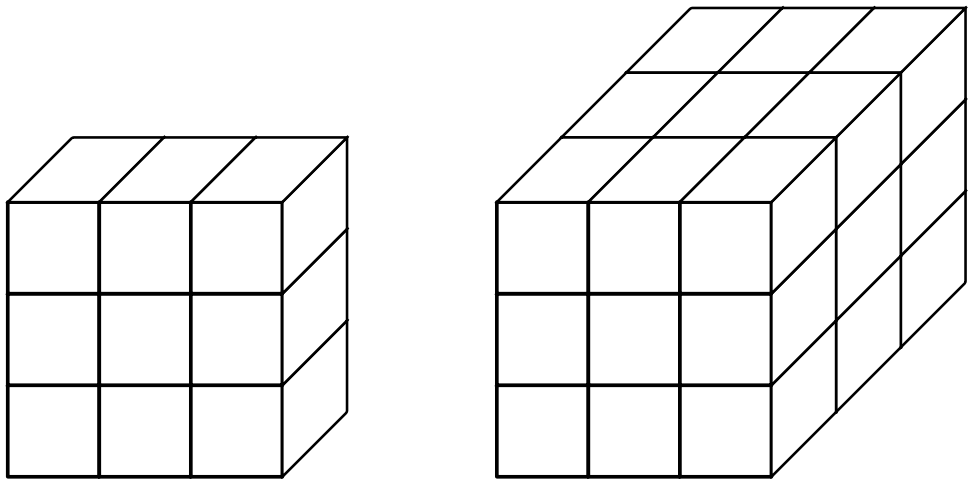
- Extend pairwise approach?
  - 2 sequences: table of n^2 prefix alignments
  - 3 sequences: table of n^3 prefix alignments
  - Warning: Very large numbers ahead



- Aligning 100 sequences of 300 symbols: about 10^{170} prefix alignments.
  - How much computing time does the universe have?

Multiple Alignment

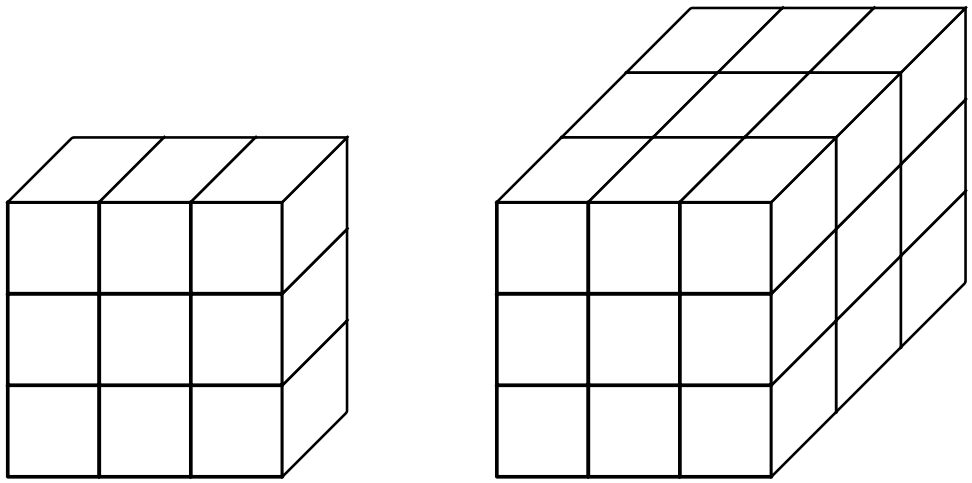
- Extend pairwise approach?
  - 2 sequences: table of n^2 prefix alignments
  - 3 sequences: table of n^3 prefix alignments
  - Warning: Very large numbers ahead



- Aligning 100 sequences of 300 symbols: about 10^{170} prefix alignments.
  - How much computing time does the universe have?

Multiple Alignment

- Extend pairwise approach?
  - 2 sequences: table of n^2 prefix alignments
  - 3 sequences: table of n^3 prefix alignments
  - Warning: Very large numbers ahead



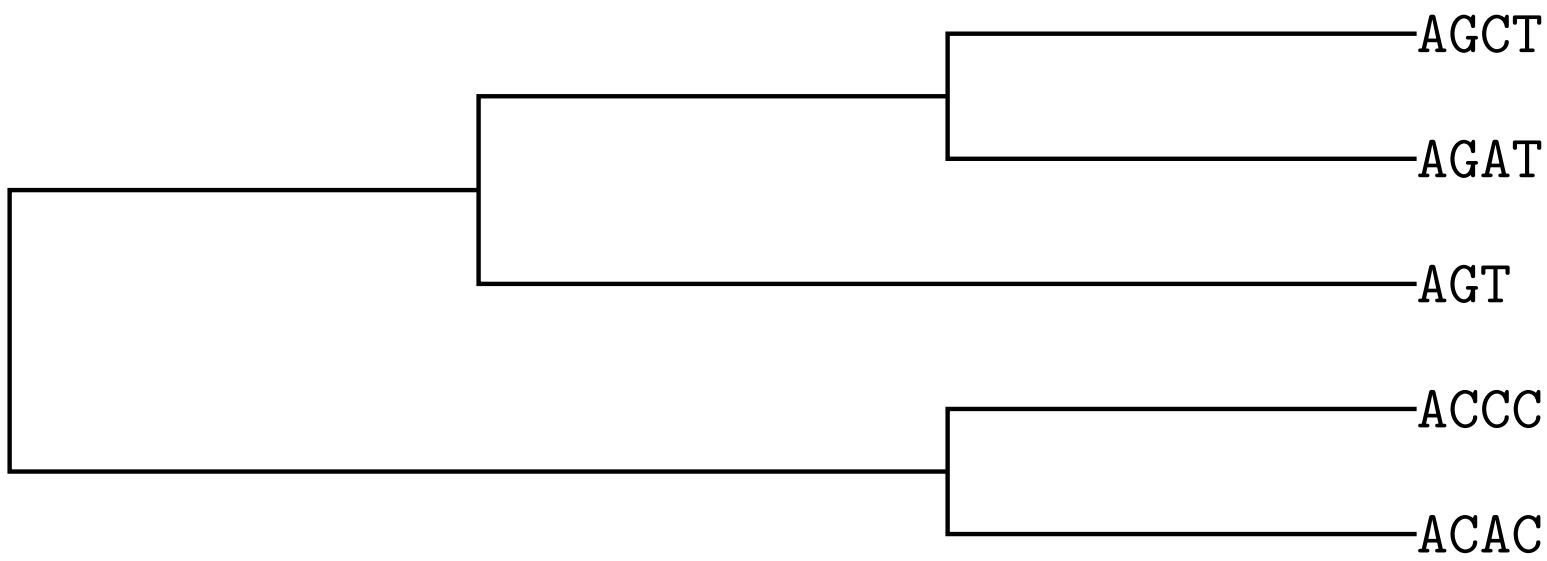
- Aligning 100 sequences of 300 symbols: about 10^{170} prefix alignments.
  - How much computing time does the universe have?

Progressive Multiple Alignment

- Compute **all** pairwise alignments
- Use **alignment dissimilarities** to produce a **guilde tree**.
- Align most similar pair of sequences and **merge** them into a **profile**.
- **Progressively** align profiles.
- **Result:** All sequences aligned (and merged into one profile).
- Programs `clustal`, `muscle`

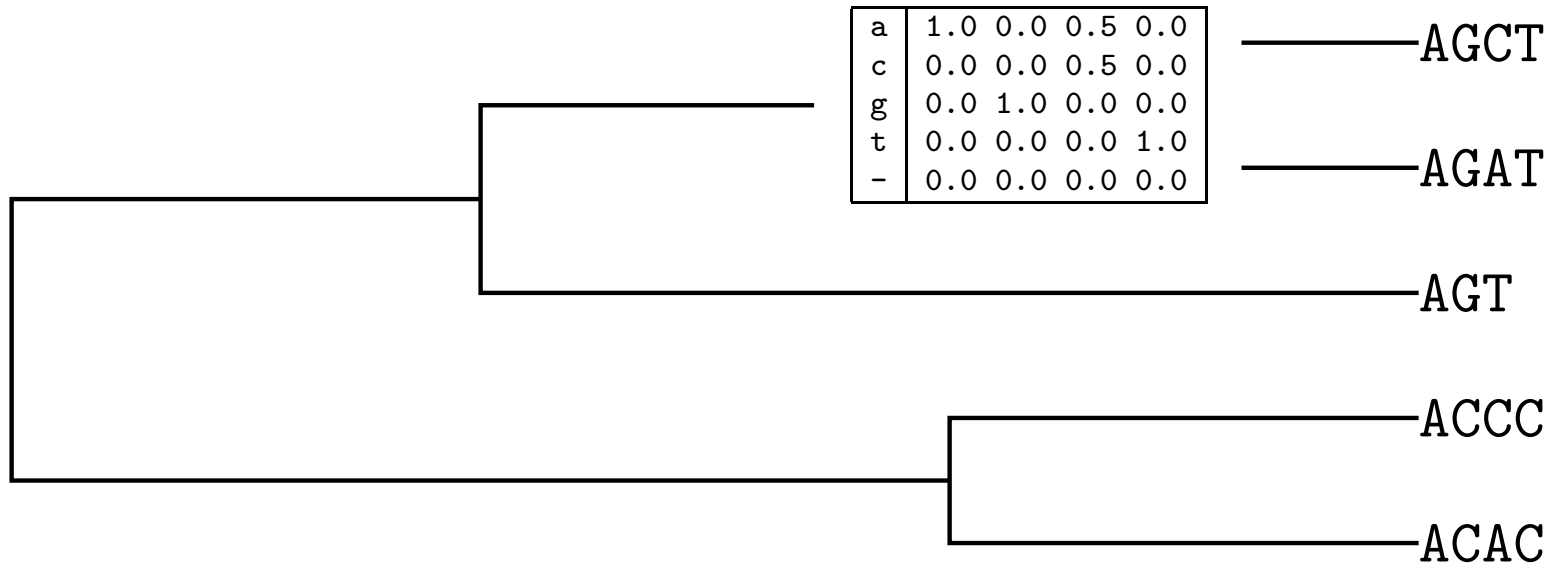
More Uses of Continuous Sequences

- Profile searches (mostly superseded by HMMs)
- Progressive multiple alignment



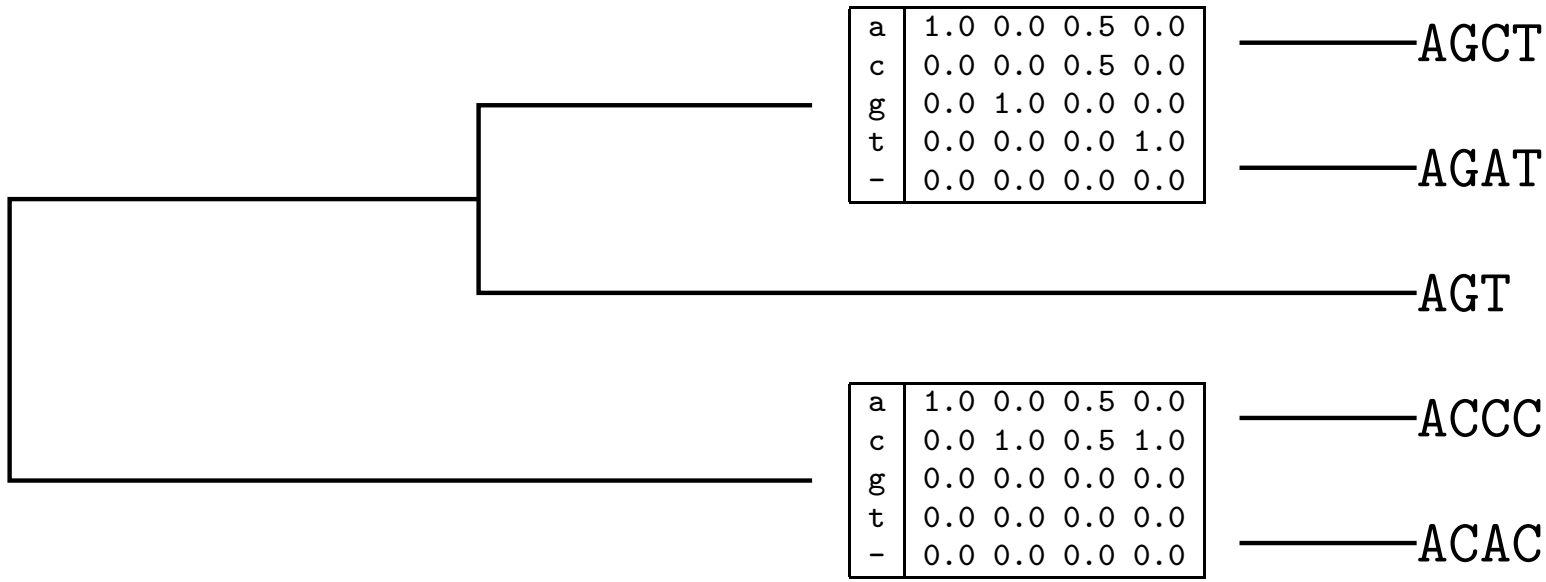
More Uses of Continuous Sequences

- Profile searches (mostly superseded by HMMs)
- Progressive multiple alignment



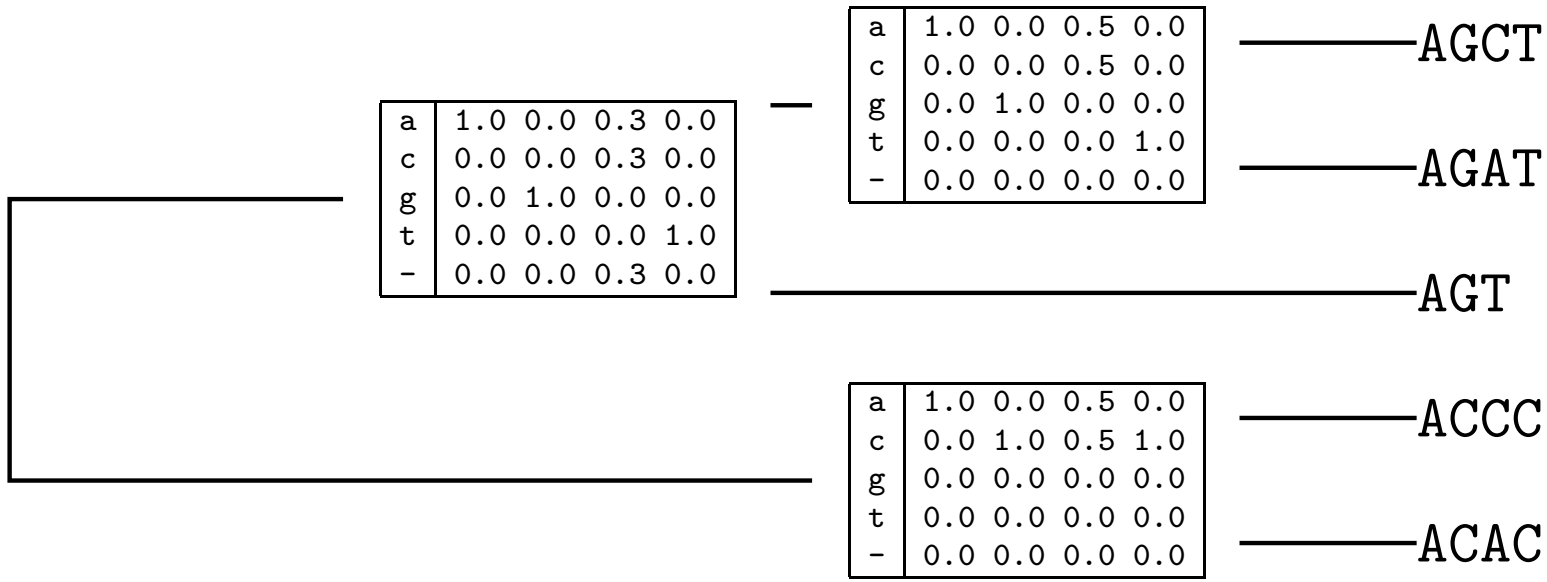
More Uses of Continuous Sequences

- Profile searches (mostly superseded by HMMs)
- Progressive multiple alignment



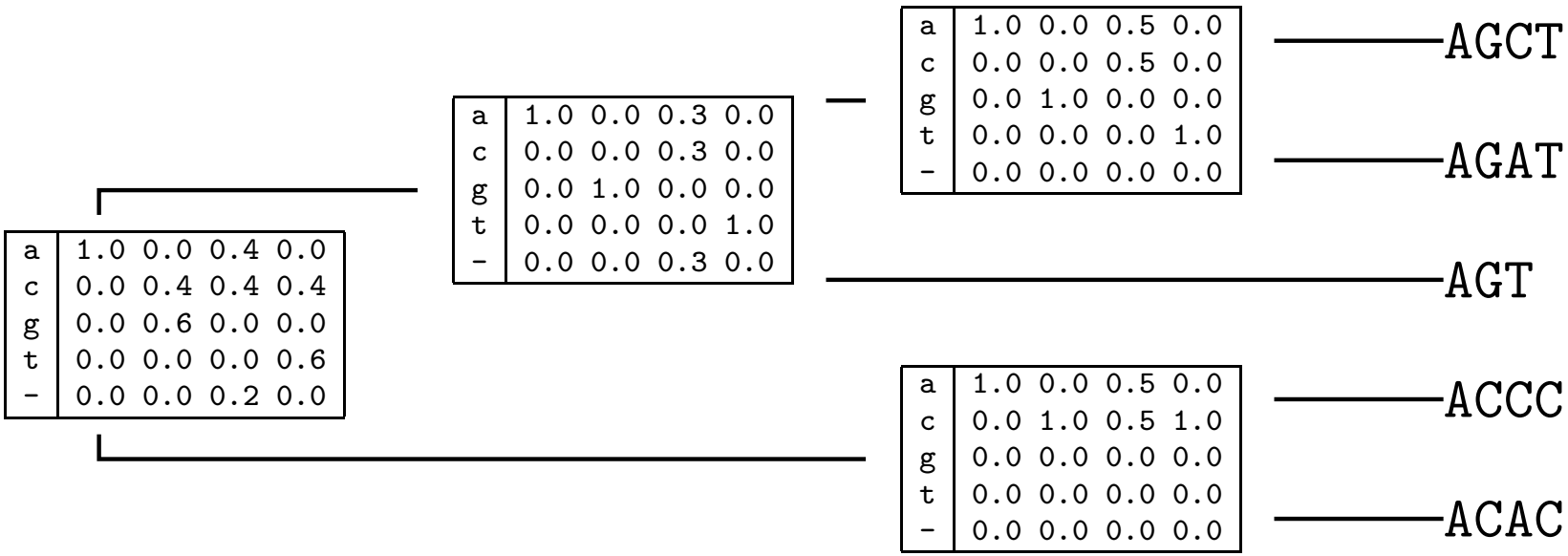
More Uses of Continuous Sequences

- Profile searches (mostly superseded by HMMs)
- Progressive multiple alignment

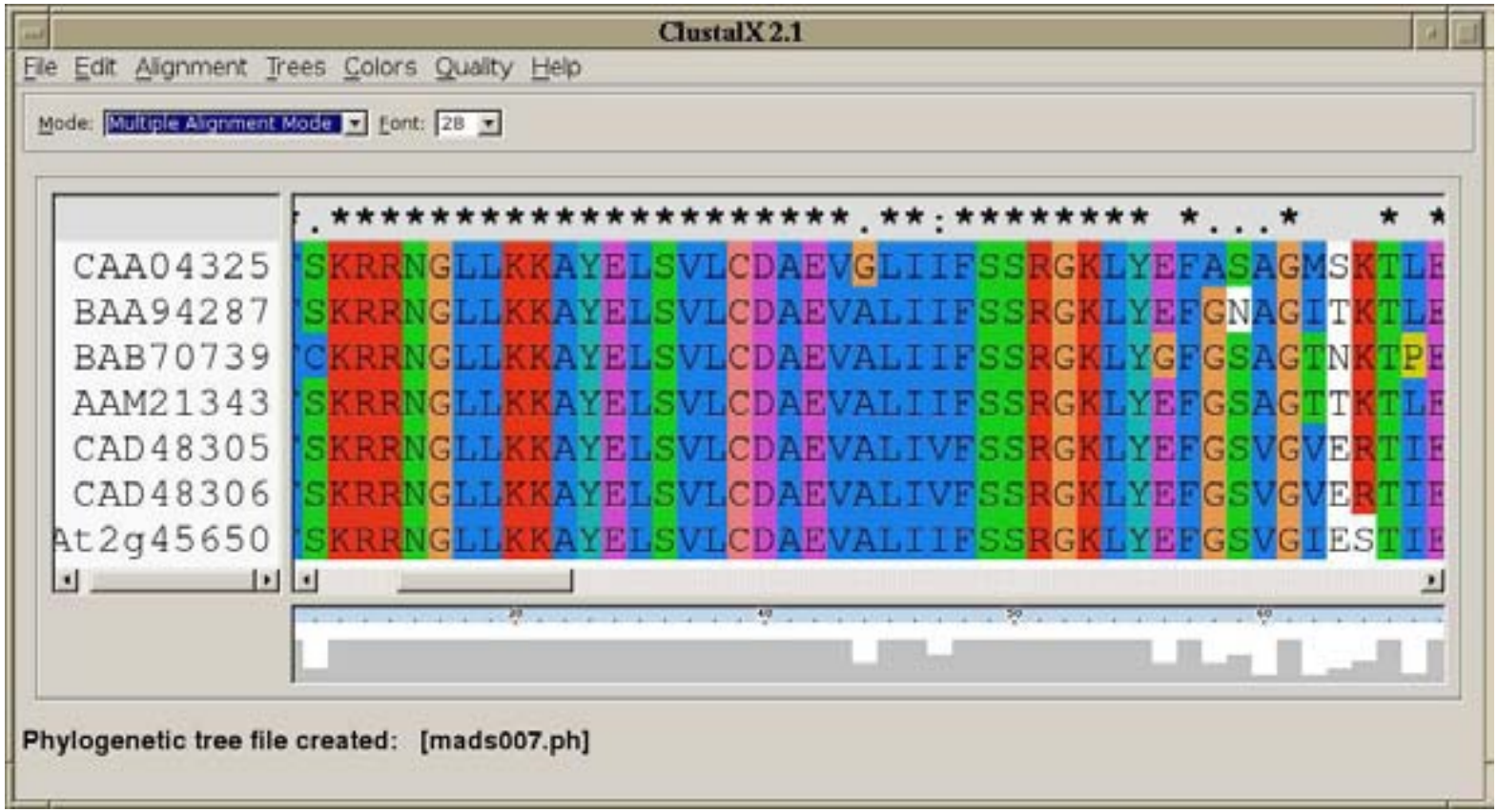


More Uses of Continuous Sequences

- Profile searches (mostly superseded by HMMs)
- Progressive multiple alignment



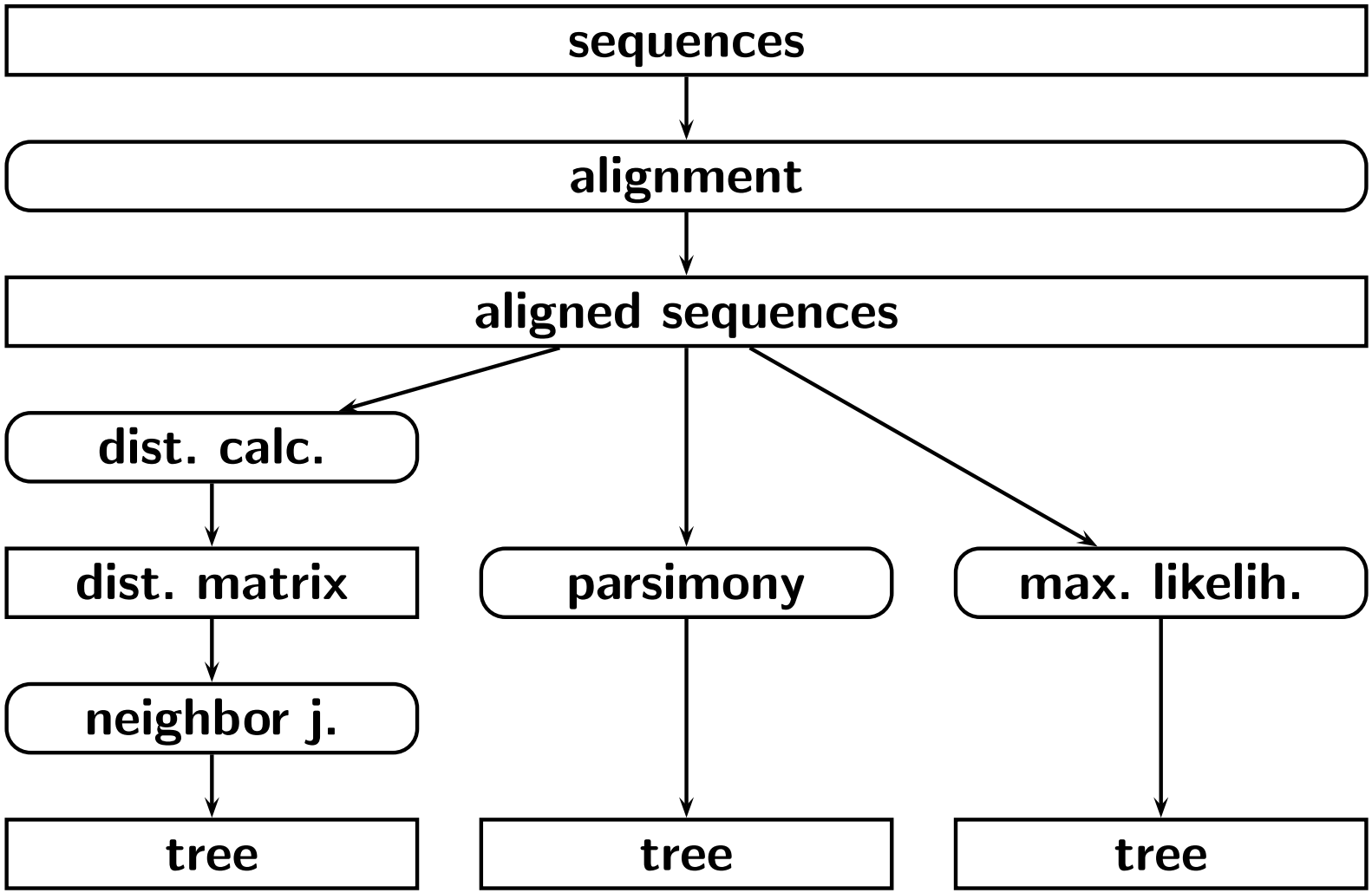
Multiple Alignment



(program: `clustalx`)

`http://www.clustal.org/`

Overview of Molecular Phylogeny



Acknowledgements

- Kai-Uwe Winter, Thomas Münster, Luzie U. Wingen, Günter Theißen, Heinz Saedler
- Begoña Valdazo-Gonzalez, Nick Knowles, Don King
- Jan Gewehr, Thomas Martinetz, Daniel Polani, Simon Moxon, Vincent Moulton
- Anyela Camargo, Alessandra Devoto, John Turner

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997).
Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.
Nucleic Acids Research, 25:3389–3402.
http://nar.oupjournals.org/cgi/content/full/25/17/3389.

Crosswell, L. C. and Thornton, J. M. (2012).
ELIXIR: A distributed infrastructure for European biological data.
Trends in Biotechnology, 30:241–242.

Langton, C. G. (1992).
Preface.
In Langton, C. G., Taylor, C., Farmer, J. D., and Rasmussen, S., editors, *Artificial Life II*, volume X of *Santa Fe Institute Studies in the Sciences of Complexity, Proceedings*, pages xiii–xviii, Redwood City, CA. Addison-Wesley.

Leggett, R. M., Ramirez-Gonzalez, R. H., Verweij, W., Kawashima, C. G., Iqbal, Z., Jones, J. D., Caccamo, M., and MacLean, D. (2013).
Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de bruijn graphs.
PLoS One, 8:e60058.

Winter, K.-U., Becker, A., Münster, T., Kim, J. T., Saedler, H., and Theißen, G. (1999).
MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants.
Proceedings of the National Academy of Sciences, USA, 96:7342–7347.

Wright, C. F., Morelli, M. J., Thébaud, G., Knowles, N. J., Merzyk, P., Paton, D. J., Haydon, D. T., and King, D. P. (2011).
Beyond the consensus: Dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing.
Journal of Virology, 85:2266–2275.

Thank You

for your attention and participation