Analysis of DNA sequences

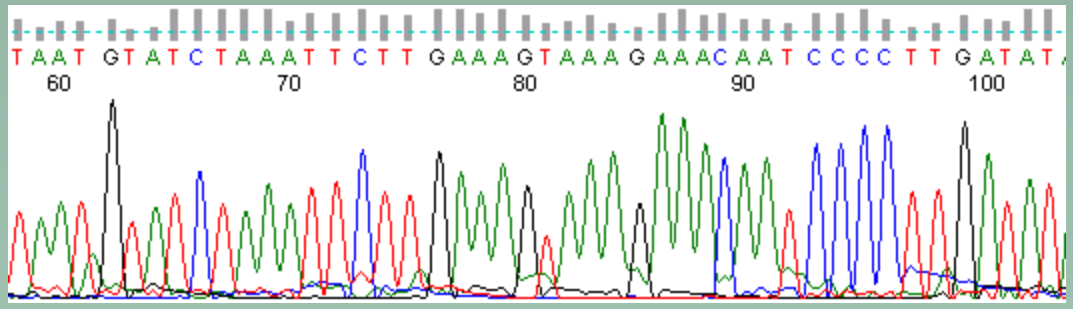
Tomáš Fér
Department of Botany, Charles University, Prague
tomas.fer@centrum.cz
http://botany.natur.cuni.cz/fer/markers/practicals/DNA.htm

Analysis of sequences

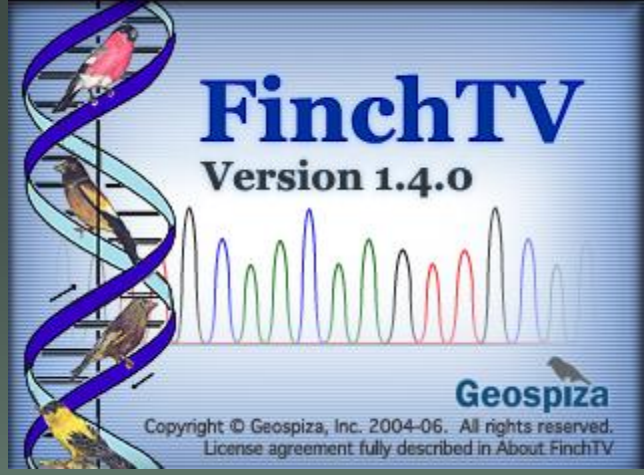
- sequencing reaction (BigDye v3.1 or v1.1)
- run on an automated sequencer
- sequence control in BLAST
- contig → consensus sequence
- alignment and its editing
- alignment trimming (removing equivocal parts)
- indel coding (optional)
- input format for various software (PAUP, MrBayes, TCS…)

Products of sequencing reaction

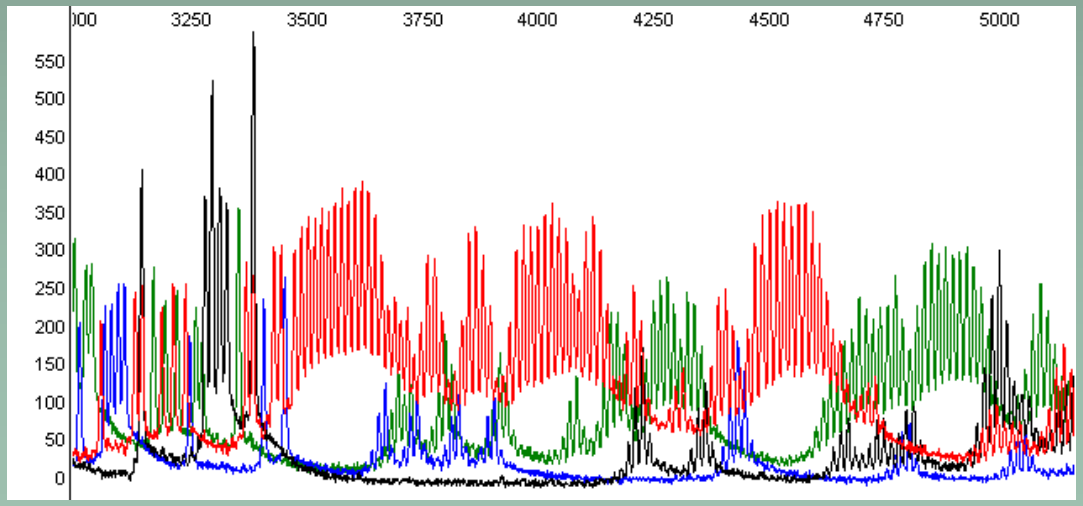
- capillary gel electrophoresis – sequencer
- primary data – laser emission detected by camera
- automated analysis (sequencer software)
- chromatogram with *base calling* (and with quality information)



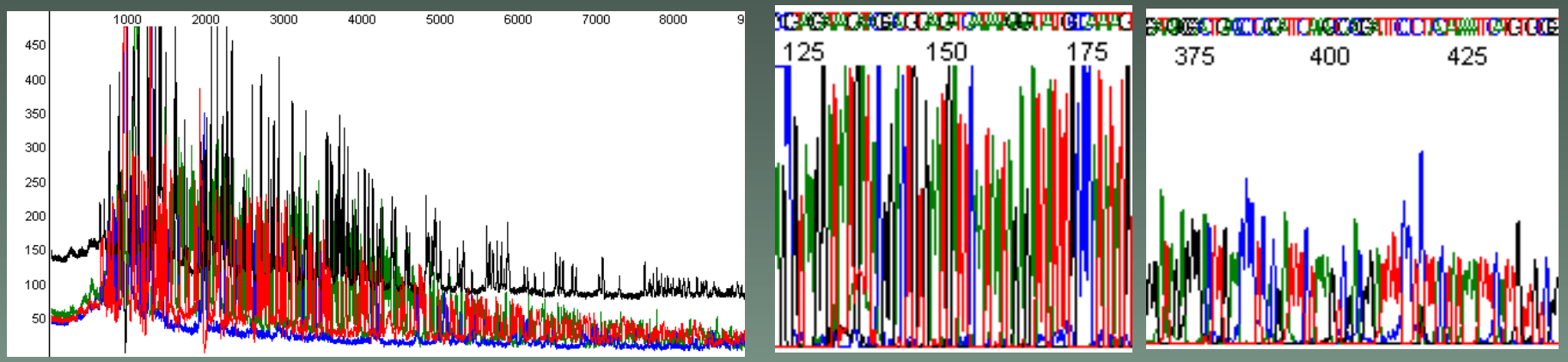
Sequence viewing



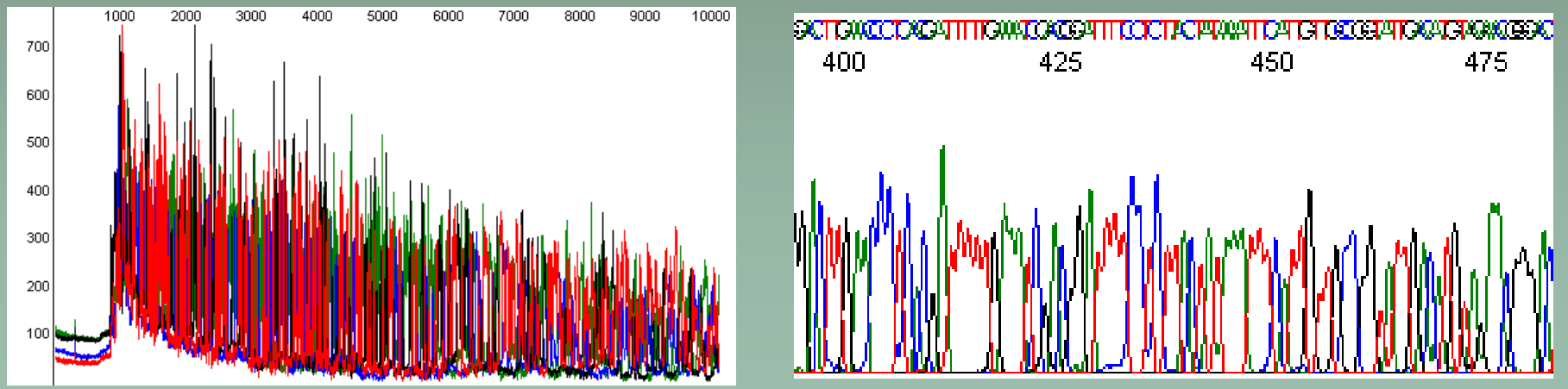
- e.g., software FinchTV
- viewing, printing…
- *base calling* edit
- viewing of primary data (*raw data*)



Effect of template quantity

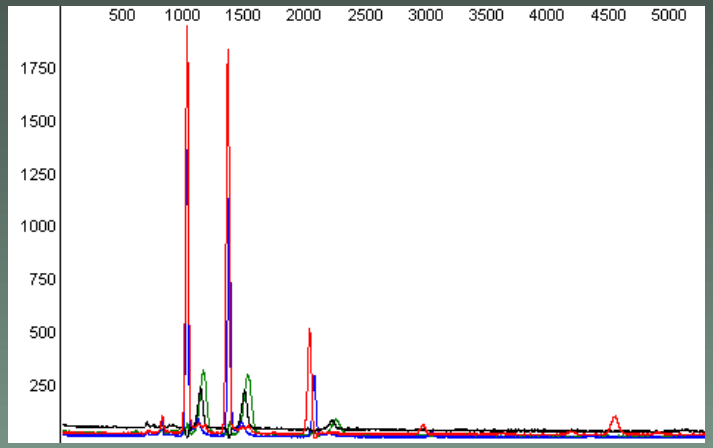


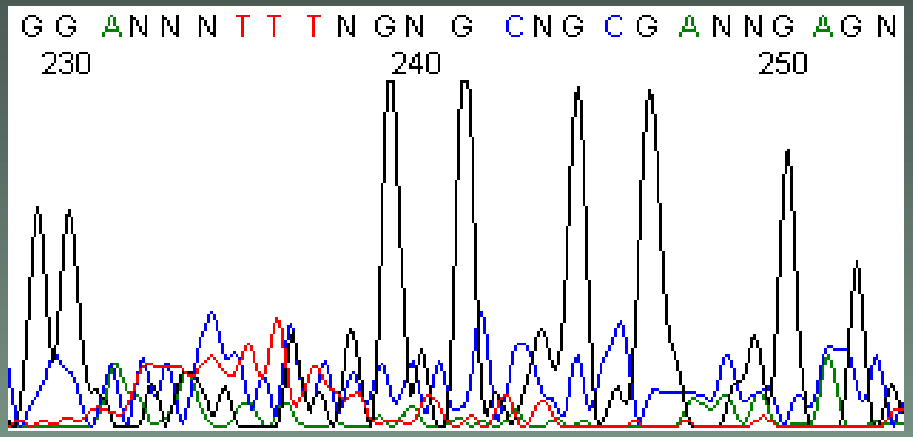
too much template DNA



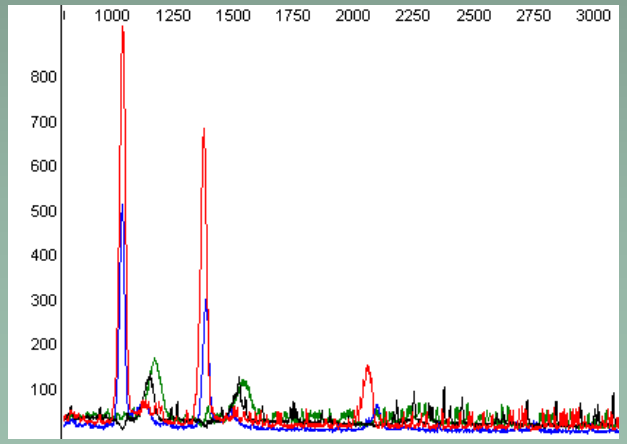
correct template/primer ratio

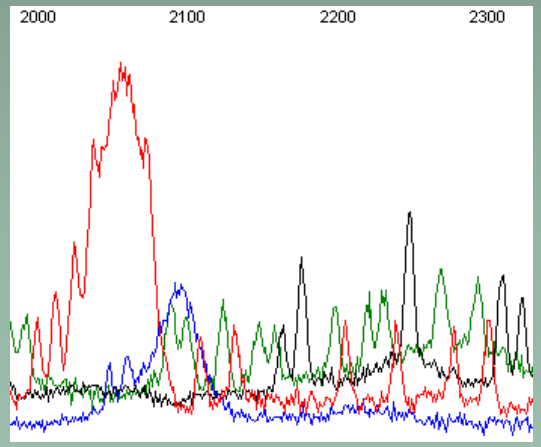
Bad sequences

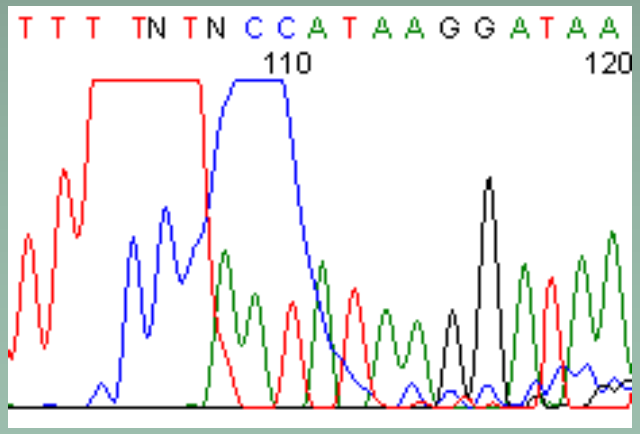




no template (or „discarded“ products)

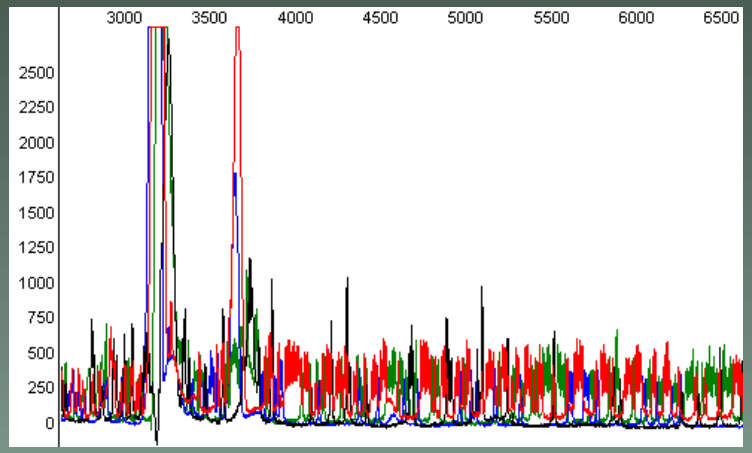


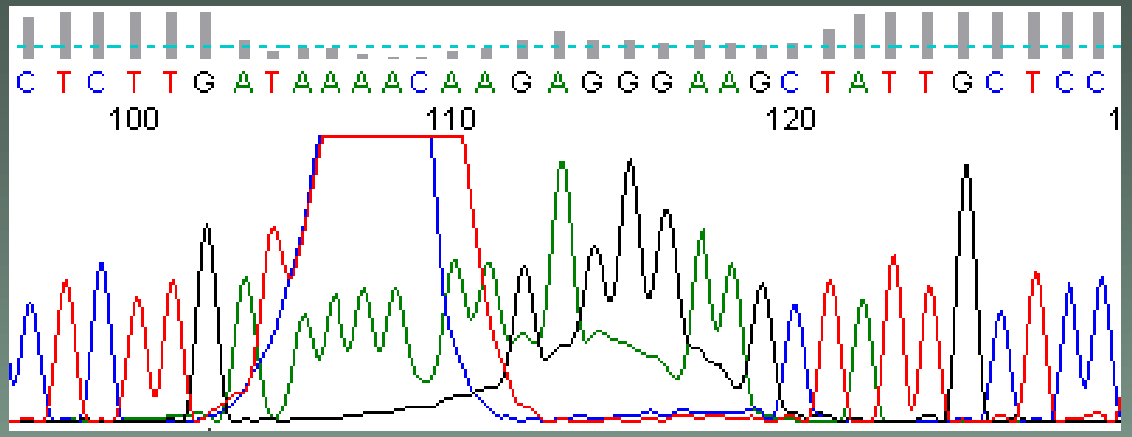




weak reaction

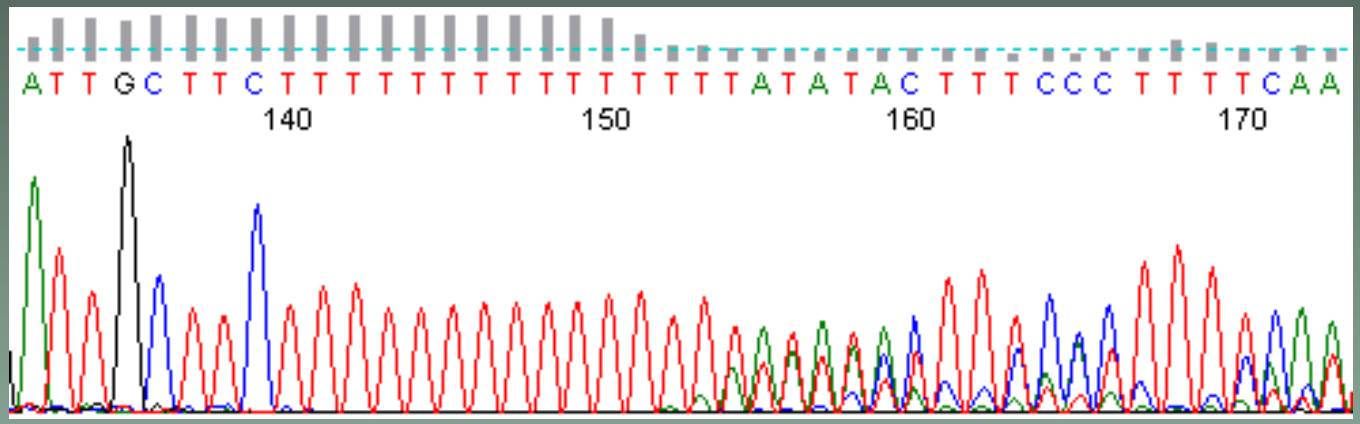
Bad sequences

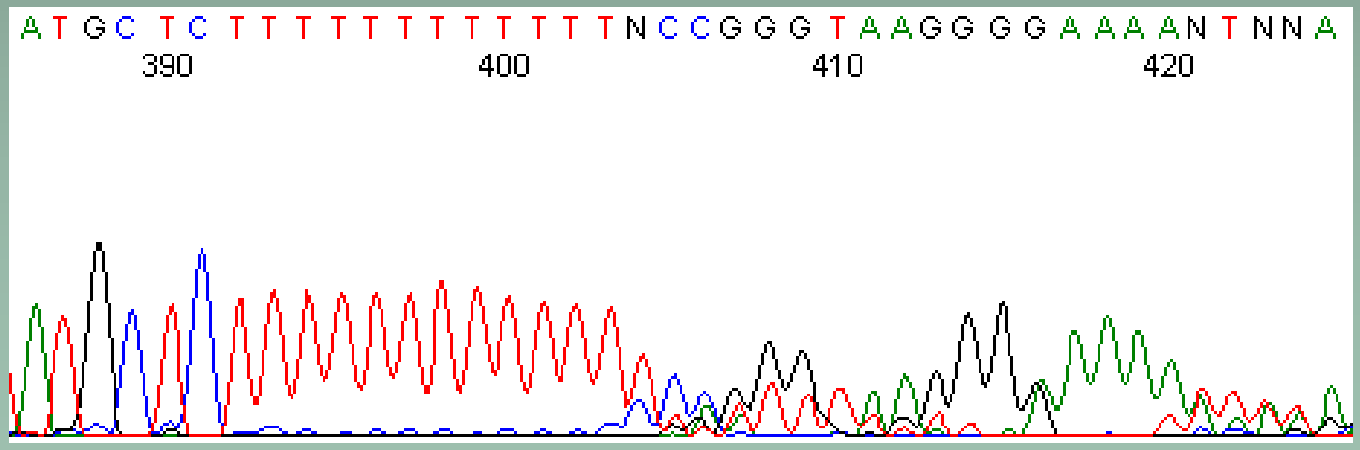




poorly purified reaction

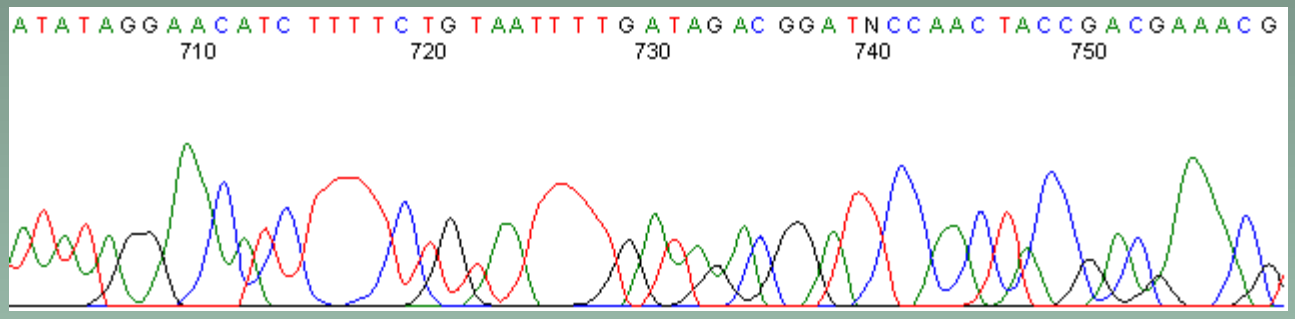
Problem with mononucleotide repeats



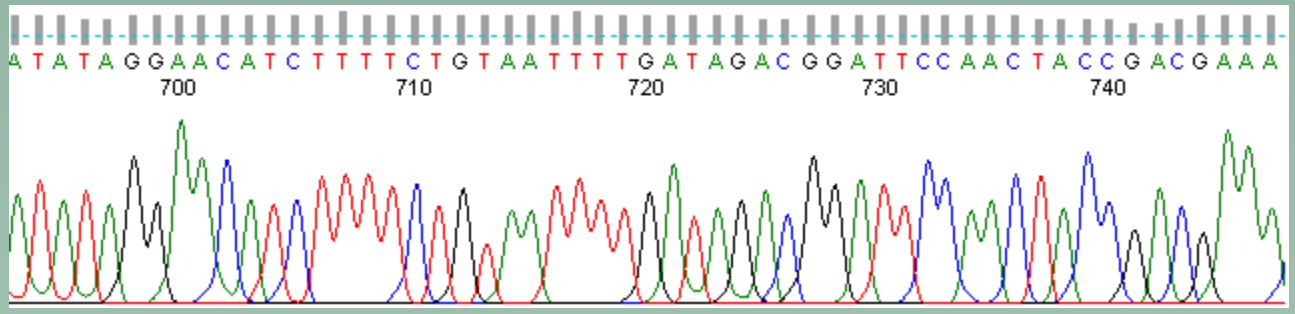


Chromatogram improvement

- use alternativ *base calling* algorithm
- e.g., PeakTrace Online (paid!) http://www.nucleics.com/peaktrace-sequencing/index.php



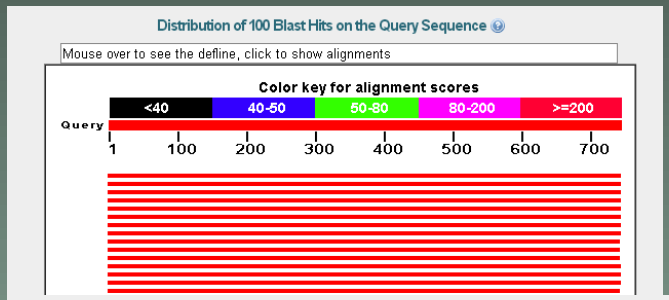
ABI basecaller



PeakTrace basecaller

BLAST (Basic Local Alignment Search Tool)

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



Curcuma colorata tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast

Sequence ID: gb|DQ471978.1| Length: 900 Number of Matches: 1

Related Inform

Range 1: 95 to 830 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
1328 bits(1472)	0.0	736/736(100%)	0/736(0%)	Plus/Minus

```
Query  1    TTAAAGGGACTAAAAAGTAGTAAAAAATTTAACTAGTAAGCGTAAGGATAATGATACGGA  60
Sbjct  830  TTAAAGGGACTAAAAAGTAGTAAAAAATTTAACTAGTAAGCGTAAGGATAATGATACGGA  771

Query  61   CTGTGAATGATTCAATAATAGAGATTATTTGCCCATATATGTTTATTTGTAGAGGTATTG  120
Sbjct  770  CTGTGAATGATTCAATAATAGAGATTATTTGCCCATATATGTTTATTTGTAGAGGTATTG  711

Query  121  TATCTATCGAAATTGGGATAAGATCCAAGGATTTTAGTTCGGATACATTTGTGTTGTGAA  180
Sbjct  710  TATCTATCGAAATTGGGATAAGATCCAAGGATTTTAGTTCGGATACATTTGTGTTGTGAA  651

Query  181  AGAGTGGAGTGAATGAGAAAGATAGTGAATTTTTGTTTGAACTGAATCGCTGATGaaaaaa  240
Sbjct  650  AGAGTGGAGTGAATGAGAAAGATAGTGAATTTTTGTTTGAACTGAATCGCTGATGAAAAAA  591

Query  241  aaaaGGAGGATAAATATTTAGGAAGTAAAATTACCTTTTTATTGGGGATAGAGGGACTTG  300
Sbjct  590  AAAAGGAGGATAAATATTTAGGAAGTAAAATTACCTTTTTATTGGGGATAGAGGGACTTG  531

Query  301  AACCCTCACGATTTCTAAAGTCGACGGATTTTCCTCTTACTATAAAATTTCATTGTTGTCG  360
Sbjct  530  AACCCTCACGATTTCTAAAGTCGACGGATTTTCCTCTTACTATAAAATTTCATTGTTGTCG  471

Query  361  GTATTGACATGTAGAATGGGACTCTCTCTTTATTCTCGTCCGATTAATCAGTTTTTCAAA  420
Sbjct  470  GTATTGACATGTAGAATGGGACTCTCTCTTTATTCTCGTCCGATTAATCAGTTTTTCAAA  411

Query  421  AGATCTATCAAACTCTGGAATGAATGATTTAATAATTGAATTGAATATTCAATTCTTCTT  480
Sbjct  410  AGATCTATCAAACTCTGGAATGAATGATTTAATAATTGAATTGAATATTCAATTCTTCTT  351

Query  481  TCAACTTCCATCGGACTGGATTCACAATAACTCTAATTCTGAATTTTTCATATTATAATT  540
Sbjct  350  TCAACTTCCATCGGACTGGATTCACAATAACTCTAATTCTGAATTTTTCATATTATAATT  291

Query  541  ATCCATATAATATAATGGATTCGAGTCATGATTAATCGTTTGATTTGATATGTCAGTATG  600
Sbjct  290  ATCCATATAATATAATGGATTCGAGTCATGATTAATCGTTTGATTTGATATGTCAGTATG  231

Query  601  TATACGTATGTATTAGGTATATAGGAACATCTTTTCTGTAATTTTGATAGACGGATTCCA  660
Sbjct  230  TATACGTATGTATTAGGTATATAGGAACATCTTTTCTGTAATTTTGATAGACGGATTCCA  171

Query  661  ACTACCGACGAAACGTAGTCAACTTCATTCGTTAGAACAGCTTCCATTGAGTCTCTGCAC  720
Sbjct  170  ACTACCGACGAAACGTAGTCAACTTCATTCGTTAGAACAGCTTCCATTGAGTCTCTGCAC  111

Query  721  CTATCCtttttttttt  736
Sbjct  110  CTATCCTTTTTTTTTT  95
```

Sequences producing significant alignments:

Select: All None Selected:0

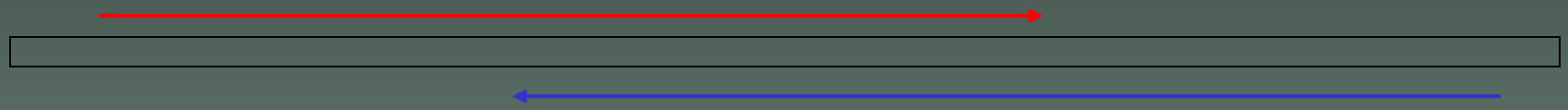
Alignments Download GenBank Graphics Distance tree of results

Curcuma colorata tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spa							1
Curcuma purpurascens tRNA-Leu (trnL) gene, partial sequence; trnL-trnF interger							1
Curcuma xanthorrhiza tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic							1
Curcuma soloensis tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic s							1
Curcuma longa tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic space							1
Curcuma heyneana tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic s							1
Curcuma aromatica tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic s							1
Curcuma aurantiaca tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast							1
Curcuma zedoaria tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast	1326	1326	99%	0.0	100%	DQ471974.1	
Curcuma ochrorhiza tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast	1326	1326	99%	0.0	100%	DQ471972.1	
Curcuma elata tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast	1326	1326	99%	0.0	100%	DQ471969.1	
Curcuma australasica tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast	1323	1323	100%	0.0	99%	DQ666419.1	
Curcuma alismatifolia tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast	1323	1323	100%	0.0	99%	DQ471962.1	
Curcuma aeruginosa tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast	1321	1321	100%	0.0	99%	DQ471966.1	
Stahlianthus involucratus tRNA-Leu (trnL) gene, partial sequence; trnL-trnF intergenic spacer, complete sequence; and tRNA-Phe (trnF) gene, partial sequence; chloroplast	1321	1321	99%	0.0	99%	AY424799.1	

Contig

- sequence reconstructed from overlaping segments (assembling)
- forward sequence (5´ → 3´)
- reverse sequence (5´ → 3´)
  - convert to „reverse complement“ sequence (i.e., 3´ → 5´ sequence with complementary bases)
  - i.e., ACTGAAT → ATTCAGT
- combination of forward and reverse sequences (must overlap!) into one
- CAP3 (contig assembly program), e.g. http://doua.prabi.fr/software/cap3

Contig



forward

ACTTGCAGCTGGGTGCCAAGGTTC

AATAATGTCTCTCGGGGAACCTTGGCAC

reverse CACGGTTCCAAGGGGCTCTCTGTAATAA

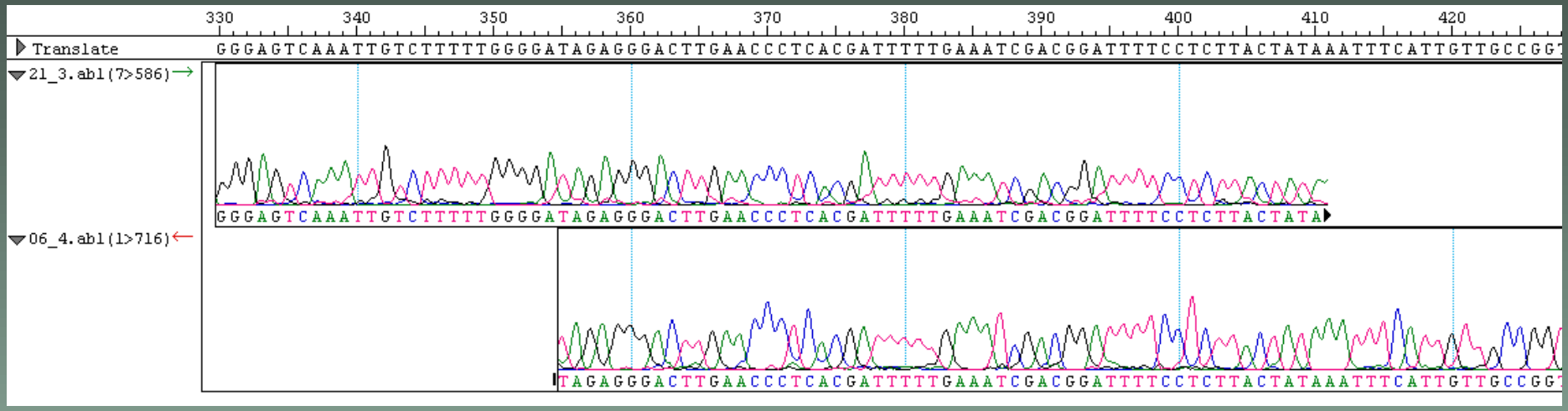
reverse-complement GTGCCAAGGTTCCCCGAGAGACATTATT

ACTTGCAGCTGGGTGCCAAGGTTCCCCGAGAGACATTATT

reverse-complement

- http://www.bioinformatics.org/sms/rev_comp.html
- http://www.famd.me.uk/AGL/RC.zip

Contig in SeqMan software



- sequence in FASTA format

```
>sequence1
GGAGTCAAATTGTCTTTTTGGGGATAGAGGGACTTGA
ACCCTCACGATTTTTGAAATCGACGGATTTTCCTCTT
ACTATAAATTTCATTGTTGCCGG
```

Contig in Geneious software

Pairwise sequence alignment

- arrangement of homologous bases at the same position
- lining up sequences to achieve maximal levels of identity

- formalized problem – looking for alignment with best score
  - **+** base identity
  - **–** base difference, gap opening/insertion, gap elongation/extension

  - gap opening – large negative score, gap extension – small penalty

```
AAGCTAGTCAGTTAGC
|||||  ||  |||||||
AAGCTGGT-AGTTAGC
```

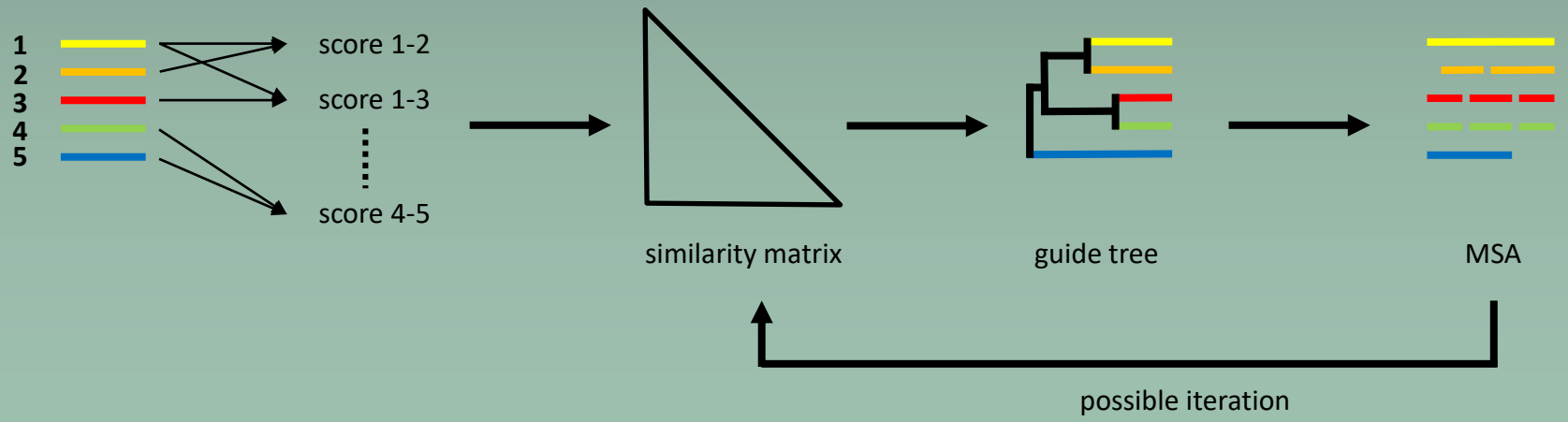
mismatch

gap

score = Σ(identities, mismatches) - Σ(gap penalties)

Multiple sequence alignment (MSA)

- calculate pairwise scores (all-against-all)
- this similarity matrix is used to generate a *guide tree*
- progressive adding of sequences to alignment in the order specified by the tree
- phylogenetic information is incorporated to guide the alignment process



Automated multiple alignment methods – software

- *progressive* (ClustalW, T-COFFEE…) – pairwise alignment of two most similar sequences, successive addition of less similar sequences
- *iterative* – repeated re-alignment of subgroups of sequences, alignment of subgroups to the global alignment, repeated re-alignment produces better score (PRRN, DIALIGN, MUSCLE…)
- progressive with iterative refinement
  - MAFFT (Multiple Alignment using Fast Fourier Transform)

Web servers for alignment

- https://www.genome.jp/tools-bin/mafft
  - ClustalW, MAFFT, PRRN
- MAFFT
  - http://mafft.cbrc.jp/alignment/server
- T-coffee
  - http://tcoffee.vital-it.ch
- Kalign
  - https://msa.sbc.su.se/cgi-bin/msa.cgi
- many aligners at EMBL webpage
  - https://www.ebi.ac.uk/Tools/msa/

Alignment

save in FASTA format with gaps

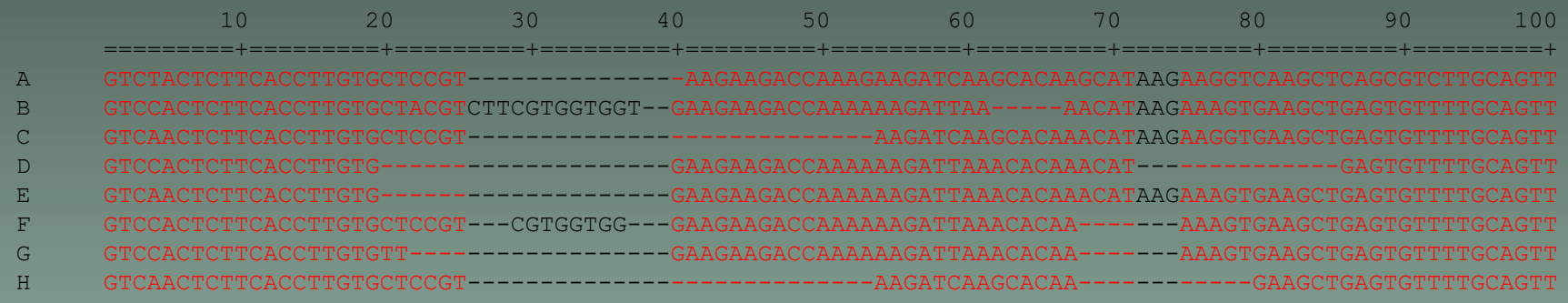
```
>sequence1
GGAGT—CAAATT-GTCTTTTTGGGG----ATAGAGG
GACTTG---AACCC
```

Alignment trimming

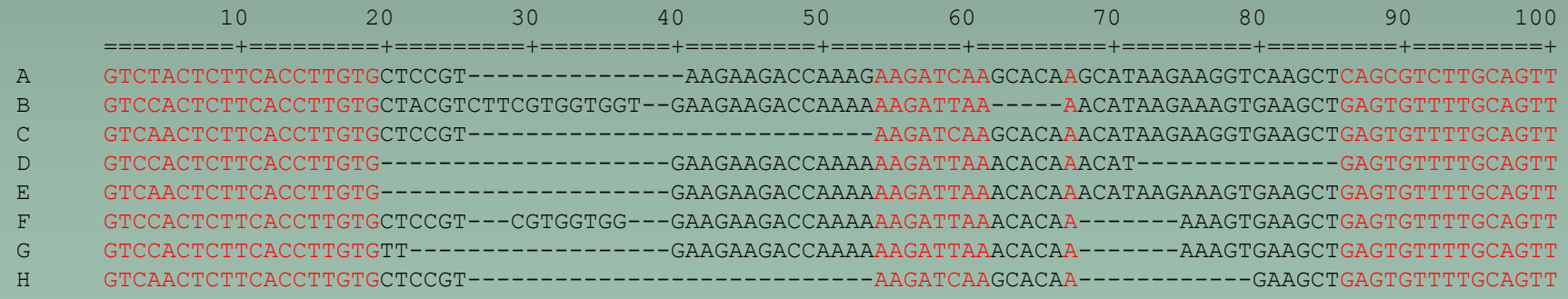
- selecting blocks of conserved regions in the alignment
- automated elimination of poorly aligned positions (may not be homologous, saturated etc.)

- **Gblocks** (http://molevol.cmima.csic.es/castresana/Gblocks)
- **trimAl** (http://trimal.cgenomics.org/) – also available as webserver (http://phylemon.bioinfo.cipf.es/utilities.html)

Alignment trimming



-gappyout – gappy regions (black) removed



-nogaps – colums with at least one gap (black) removed

Basic sequence formats

FASTA

```
>A
GTCTACTCTTCACCTTGTGAAGATCAAACAGCGTCTTGCAGTT
>B
GTCCACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
>C
GTCAACTCTTCACCTTGTGAAGATCAAAGAGTGTTTTGCAGTT
>D
GTCCACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
>E
GTCAACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
>F
GTCCACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
>G
GTCCACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
>H
GTCAACTCTTCACCTTGTGAAGATCAAAGAGTGTTTTGCAGTT
```

NEXUS

```
#NEXUS

BEGIN DATA;
 DIMENSIONS NTAX=8 NCHAR=43;
 FORMAT DATATYPE=DNA INTERLEAVE=yes GAP=-;

MATRIX
A     GTCTACTCTT CACCTTGTGA AGATCAAACA GCGTCTTGCA GTT
B     GTCCACTCTT CACCTTGTGA AGATTAAAGA GTGTTTTGCA GTT
C     GTCAACTCTT CACCTTGTGA AGATCAAAGA GTGTTTTGCA GTT
D     GTCCACTCTT CACCTTGTGA AGATTAAAGA GTGTTTTGCA GTT
E     GTCAACTCTT CACCTTGTGA AGATTAAAGA GTGTTTTGCA GTT
F     GTCCACTCTT CACCTTGTGA AGATTAAAGA GTGTTTTGCA GTT
G     GTCCACTCTT CACCTTGTGA AGATTAAAGA GTGTTTTGCA GTT
H     GTCAACTCTT CACCTTGTGA AGATCAAAGA GTGTTTTGCA GTT
;
END;
```

PHYLIP

‘nr seqs’ ‘nr chars’

```
8 43
A  GTCTACTCTTCACCTTGTGAAGATCAAACAGCGTCTTGCAGTT
B  GTCCACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
C  GTCAACTCTTCACCTTGTGAAGATCAAAGAGTGTTTTGCAGTT
D  GTCCACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
E  GTCAACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
F  GTCCACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
G  GTCCACTCTTCACCTTGTGAAGATTAAAGAGTGTTTTGCAGTT
H  GTCAACTCTTCACCTTGTGAAGATCAAAGAGTGTTTTGCAGTT
```

Format conversion

- web services
  - EMBOSS Seqret (also command line program, both Win and Unix) https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/
  - Sequence conversion http://sequenceconversion.bugaco.com/converter/biology/sequences
  - Format Converter v2.0.5 http://hcv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html
- PGDSpider (http://www.cmpg.unibe.ch/software/PGDSpider/)
- CONVERT, Formatomatic, Create
  - conversion to formats for POPGEN, Arlequin, Structure, BAPS, Phylip, MSA, SPAGeDi, FSTAT etc.
- readAl – part of trimAl distribution (commandline programe)

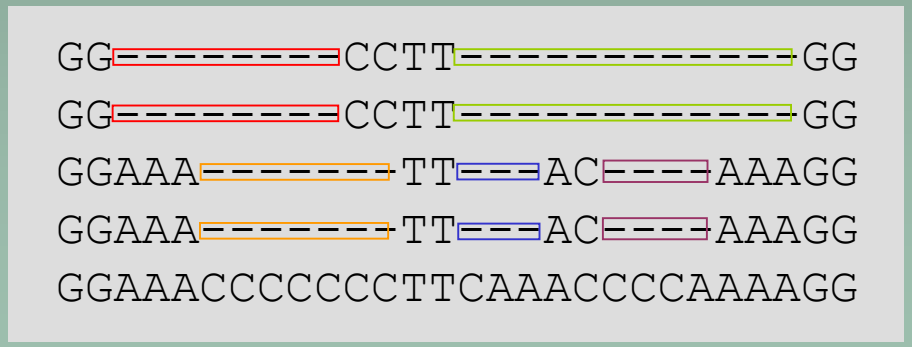
Indel coding

- gaps (indels=insertions/deletions) can give additional phylogenetic information
- different coding approaches
  - do not treat/discard (missing data)
  - other character (5th state)
  - simple indel coding (SIC) – Simmons & Ochoterena 2000
  - more complicated approaches (i.e., complex indel coding, modified complex indel coding…)

Simple indel coding (SIC)

Simmons & Ochoterena 2000

- indels beginning and ending at the same position are treated as one character
- overlapping indels – independent characters
- if a long indel includes a shorter one completely within than the long indel is treated as missing when the shorter indel is coding



Gap coding – literature

- Simmons, M.P., Ochoterena, H., 2000. *Gaps as characters in sequence based phylogenetic analyses*. Syst. Biol. 49, 369–381.
- Simmons, M.P., Ochoterena, H., Carr, T.G., 2001. *Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses*. Syst. Biol. 50, 454–462.
- Müller, K., 2006. *Incorporating information from length-mutational events into phylogenetic analysis*. Mol. Phylogenet. Evol. 38, 667–676.
- Ogden, T.H., Rosenberg, M.S., 2007. *How should gaps be treated in parsimony? A comparison of approaches using simulation*. Molec. Phylog. Evol. 42, 817-826.
- Graham, S.W., Reeves, P.A., Burns, A.C.E., Olmstead, R.G., 2000. *Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference*. Int. J. Plant Sci. 161, S83–S86.
- Young, N., Healy, J., 2003. *GapCoder automates the use of indel characters in phylogenetic analysis*. BMC Bioinformatics 4, 6.

Automatical gap coding – SeqState

http://systevol.nees.uni-bonn.de/software/SeqState

- input – FASTA with indels, NEXUS sequential
- output – NEXUS interleaved with gaps coded as 0/1/?

```
>seqA
GG--------CCTT------------GG
>seqB
GG--------CCTT------------GG
>seqC
GGAAA-------TT---AC----AAAGG
>seqD
GGAAA-------TT---AC----AAAGG
>seqE
GGAAACCCCCCCTTCAAACCCCAAAAGG
```

```
MATRIX
seqA GG--------CCTT------------GG
seqB GG--------CCTT------------GG
seqC GGAAA-------TT---AC----AAAGG
seqD GGAAA-------TT---AC----AAAGG
seqE GGAAACCCCCCCTTCAAACCCCAAAAGG
[5 indels coded]
seqA 10?1?
seqB 10?1?
seqC 01101
seqD 01101
seqE 00000
```

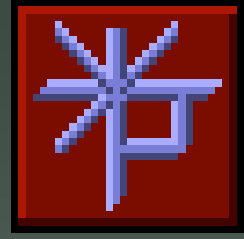
Online fasta sequence toolbox

- https://users-birc.au.dk/palle/php/fabox/index.php

- works with header data (extract, replace…)
- alignment merging, cutting
- extracts variable sites only
- formats for TCS, MrBayes, Arlequin, Excel…
- detects number of different sequences (DNA to haplotype collapser)

TCS (Templeton, Crandall & Sing)

http://darwin.uvigo.es/software/tcs.html



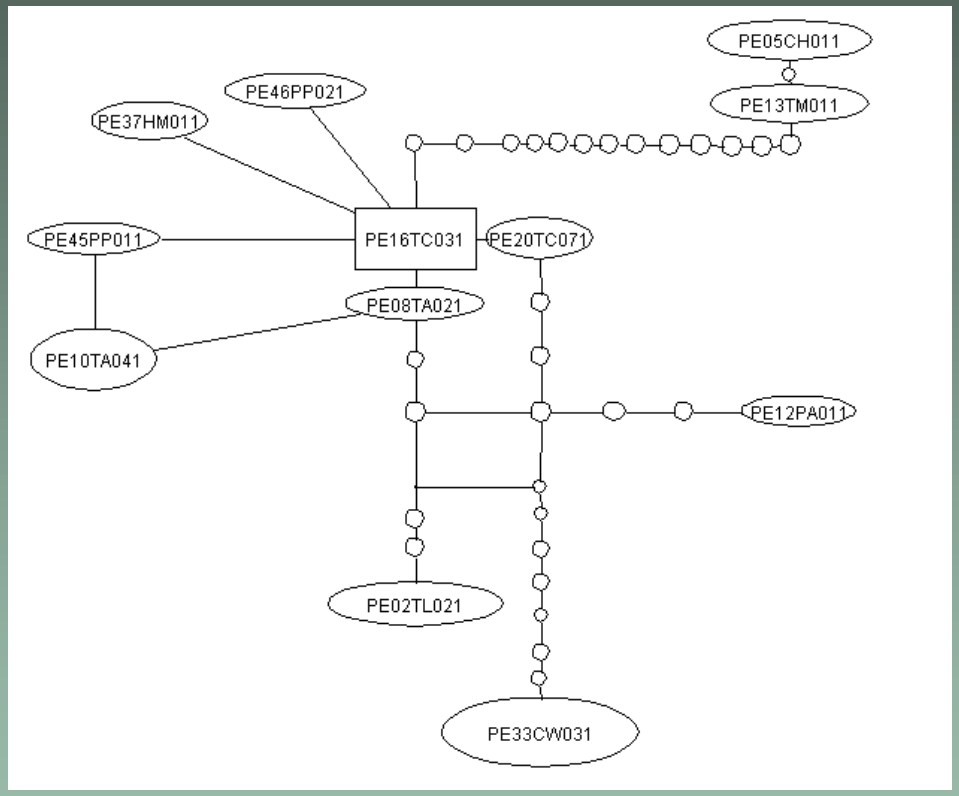
- estimates gene genealogy using statistical parsimony
- intraspecific, population level
- over traditional phylogeny reconstruction it can be considered
  - recombination
  - presence of ancestral haplotypes in populations
  - low number of variable characters
  - other than strictly dichotomous divergence

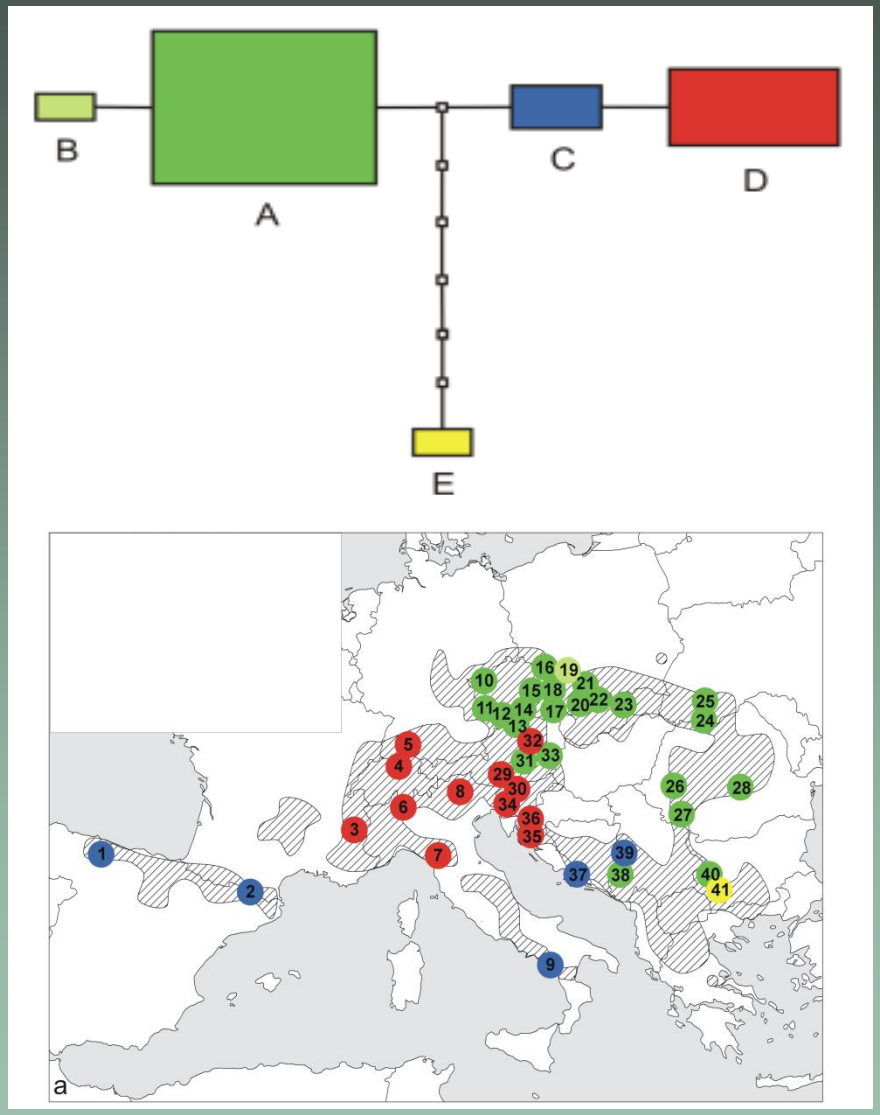
- Clement M, Posada D & Crandall KA (2000) *TCS: a computer program to estimate gene genealogies*. Molecular Ecology 9: 1657-1659.
- Templeton AR, Crandall KA & Sing CF (1992) *A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation*. Genetics 132: 619-633.
- Posada D, Crandall KA (2001) *Intraspecific gene genealogies: Trees grafting into networks*. Trends Ecol Evol 16: 37-45.

TCS (Templeton, Crandall & Sing)

- colapses sequences to haplotypes
- calculates haplotype frequences
- frequences are used to estimate probability that haplotype is an outgroup (correlates with haplotype age)
- haplotypes are connected when the parsimony probability is higher than 0.95
- resulting graph includes also missing haplotypes (extinct or non-sampled) – on the connections among detected haplotypes

TCS (Templeton, Crandall & Sing)





Pericallis

Rosa pendulina