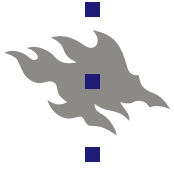
DIRECT OPTIMIZATION

Filip Högnabba
Botanical Museum
Finnish Museum of Natural History

Finnish Museum of Natural History
Botanical Museum
University of Helsinki



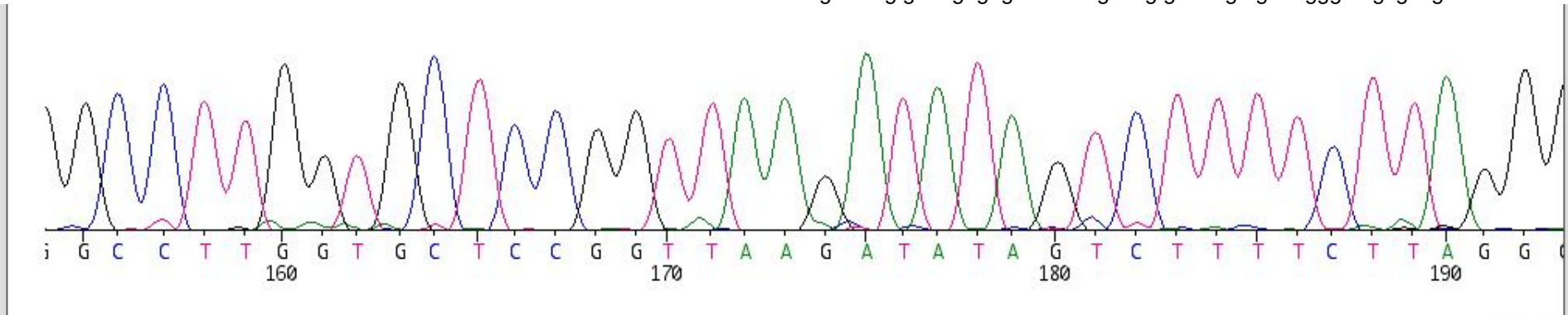
UNIVERSITY OF HELSINKI

DNA-sequences

- OBSERVATIONS of the nucleotide order in a studied sequence

/product="large subunit ribosomal RNA"

```
  1 ggattgcctc agtaacggcg agtgaagcgg caacagctca aatttgaaat ctggttcttt
 61 cgggggcccg agttgtaatt tgtagaggat gtttcgggtg cggcgccggt ctaagttcct
121 tggaacagga cgtcatagag ggtgagaatc ccgtatgcga ctggcgatca accccatgtg
181 aaaccccttc gacgagtcga gttgtttggg aatgcagctc aaaatgggtg gtaaatttca
241 tctaaagcta aataccggcc agagaccgat agcgcacaag tagagtgatc gaaagatgaa
301 aagcactttg gaaagagagt caaacagtac gtgaaattgt tgaaagggaa gcgcttgcaa
```



From observations to phylogeneti hypotheses

DNA sequences

Phylogenetic hypotheses

From observations to phylogeneti hypotheses

DNA sequences

1) alignment
2) phylogenetic analyses

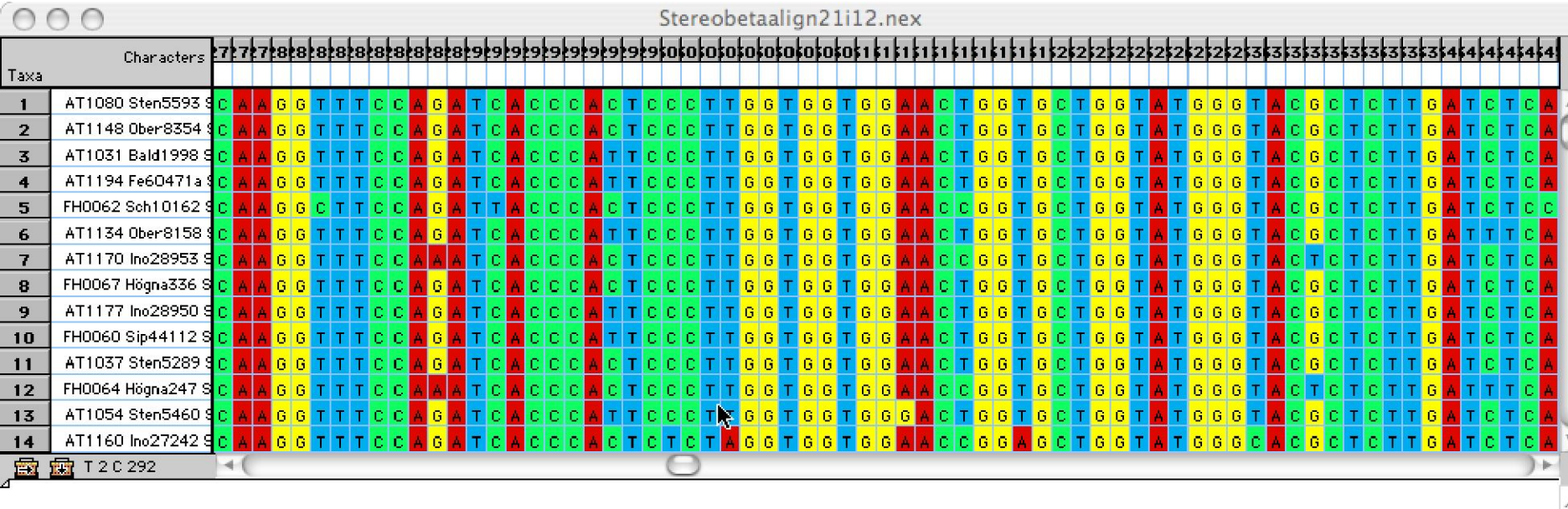
Phylogenetic hypotheses

Alignment

- homology assumptions on nucleotide level
- insertion of gaps to indicate nucleotide insertions or deletions (indels) - gaps are NOT OBSERVATIONS
- not always self-evident, particularly for non-coding sequences

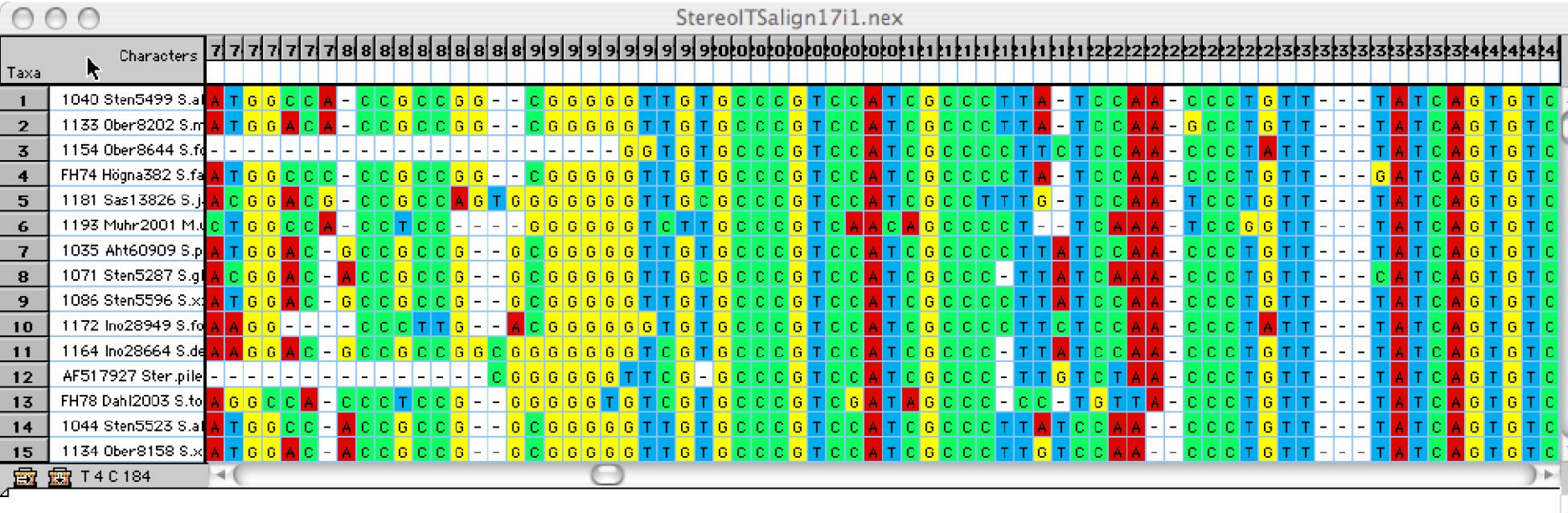
Protein-coding DNA

- no length variation
- unambiguous homology assumptions possible



Non-coding DNA

- length variation present
- unambiguous homology assumptions not possible



Alignment

- critically important in phylogenetic studies
- a dataset aligned according to different criteria or indel treatments may support contradictionary phylogenies
- no structural or developmental complexity to test nucleotide homology
- nucleotide homology can be evaluated only in reference to a topology
- static homology assumptions - may not be optimal for the final phylogenetic hypothesis

Number of alignments

f (n,m) $1 \leq n \leq 10$; $1 \leq m \leq 5$

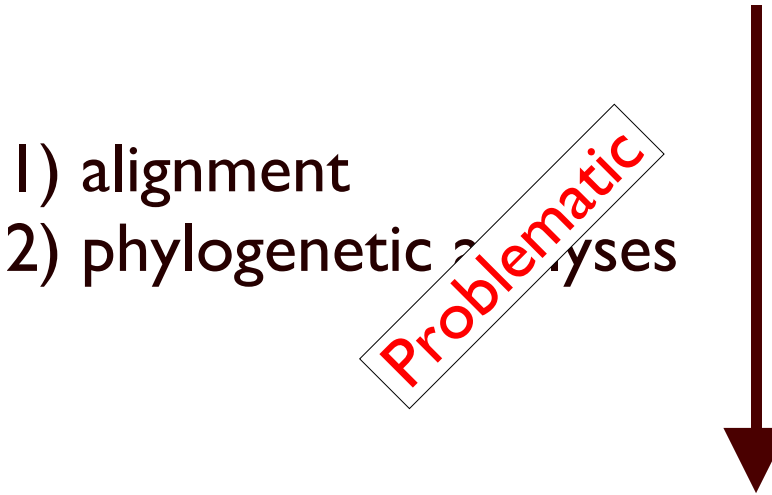
n\m	2	3	4	5
1	3	13	75	541
2	13	409	23917	2244361
3	63	16081	10681263	14638756721
4	321	699121	5552351121	11762959485121
5	1683	32193253	3147728203035	1.05 * 10^18
10	8097453	9850349744182729	3.32 * 10^26	1.35 * 10^38

- IMPOSSIBLE to find the OPTIMAL alignment even for a few short sequences

Slowinski (1998)

From observations to phylogeneti hypotheses

DNA sequences



Phylogenetic hypotheses

From observations to phylogeneti hypotheses

DNA sequences

1) alignment
2) phylogenetic analyses

Direct optimization

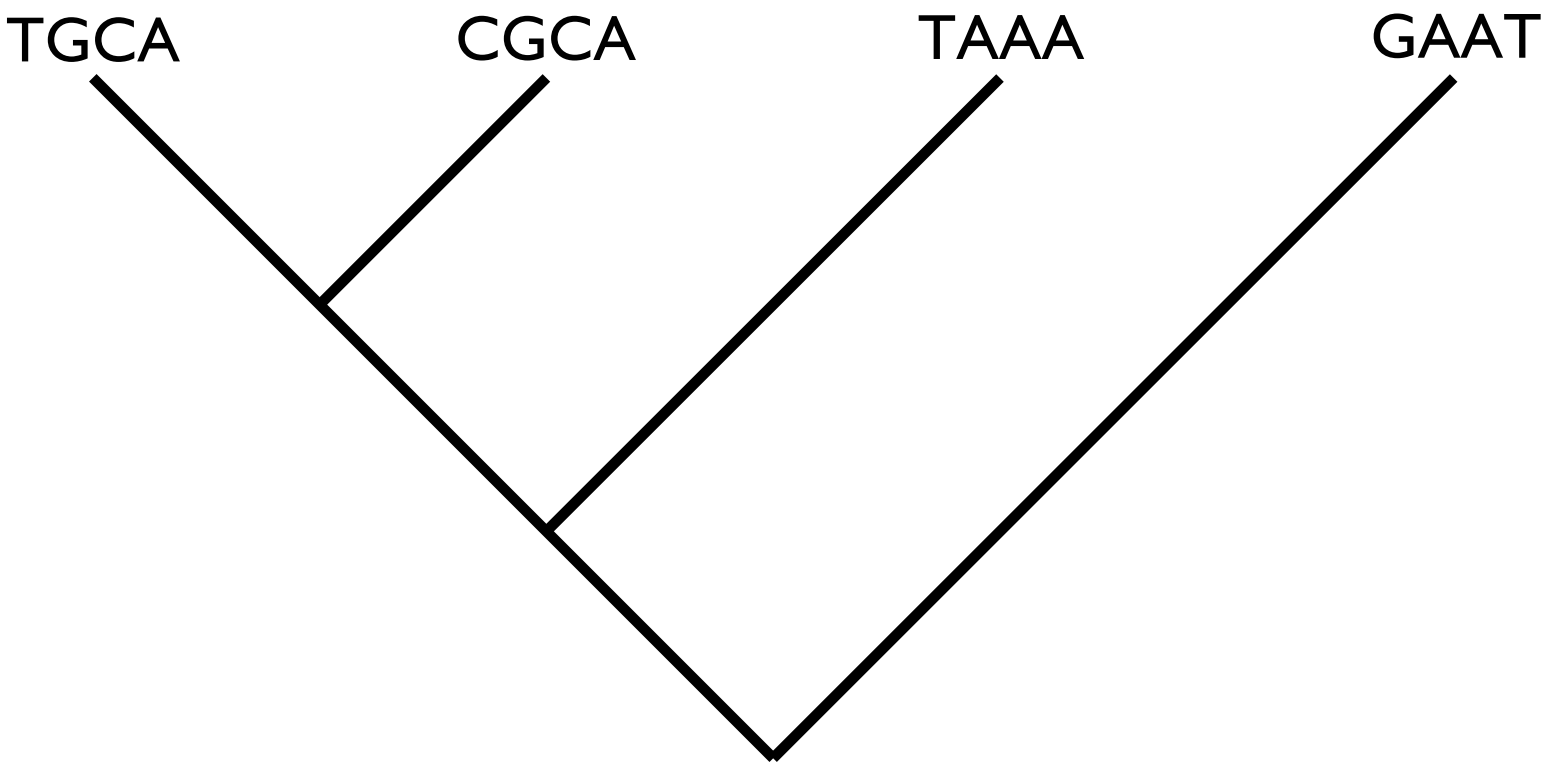
Wheeler (1996)

Phylogenetic hypotheses

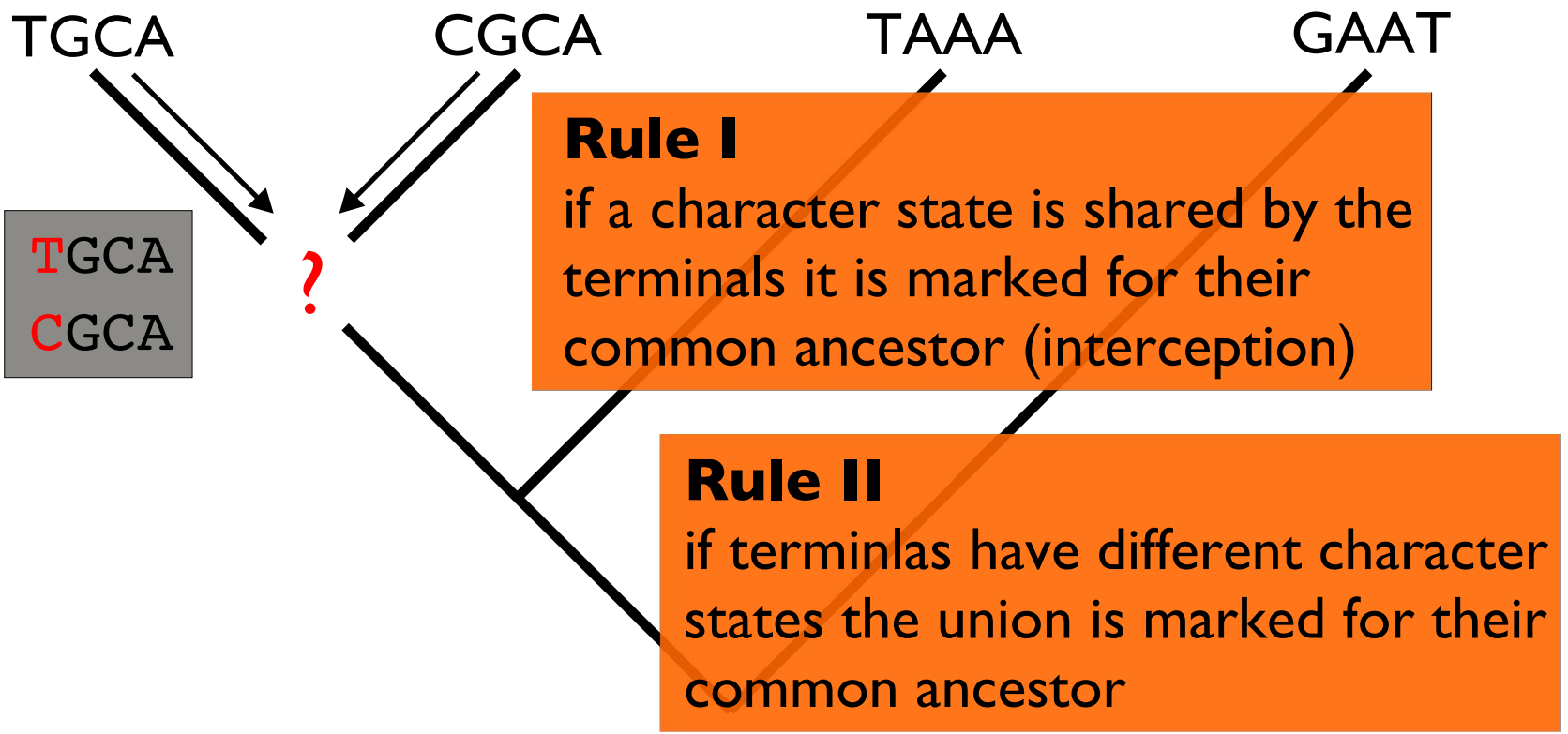
Direct optimization

- algorithm for optimization of sequence data on a tree by Sankoff (1975)
- homology assumptions and phylogenetic analyses made simultaneously
- unique homology assumptions for different topologies - dynamic homology assumptions
- indels treated equivalent to any other transformation

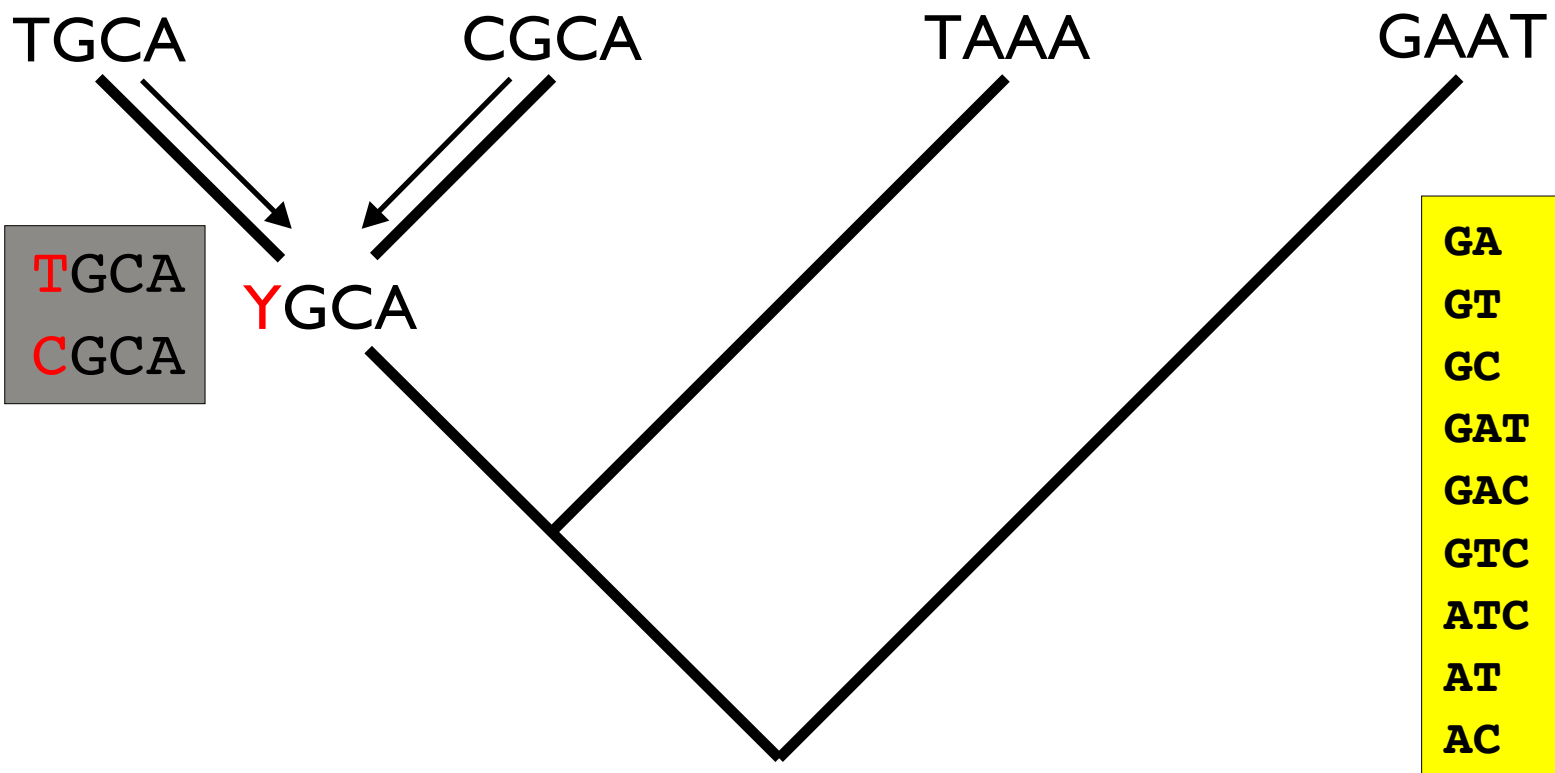
Direct optimization



Direct optimization

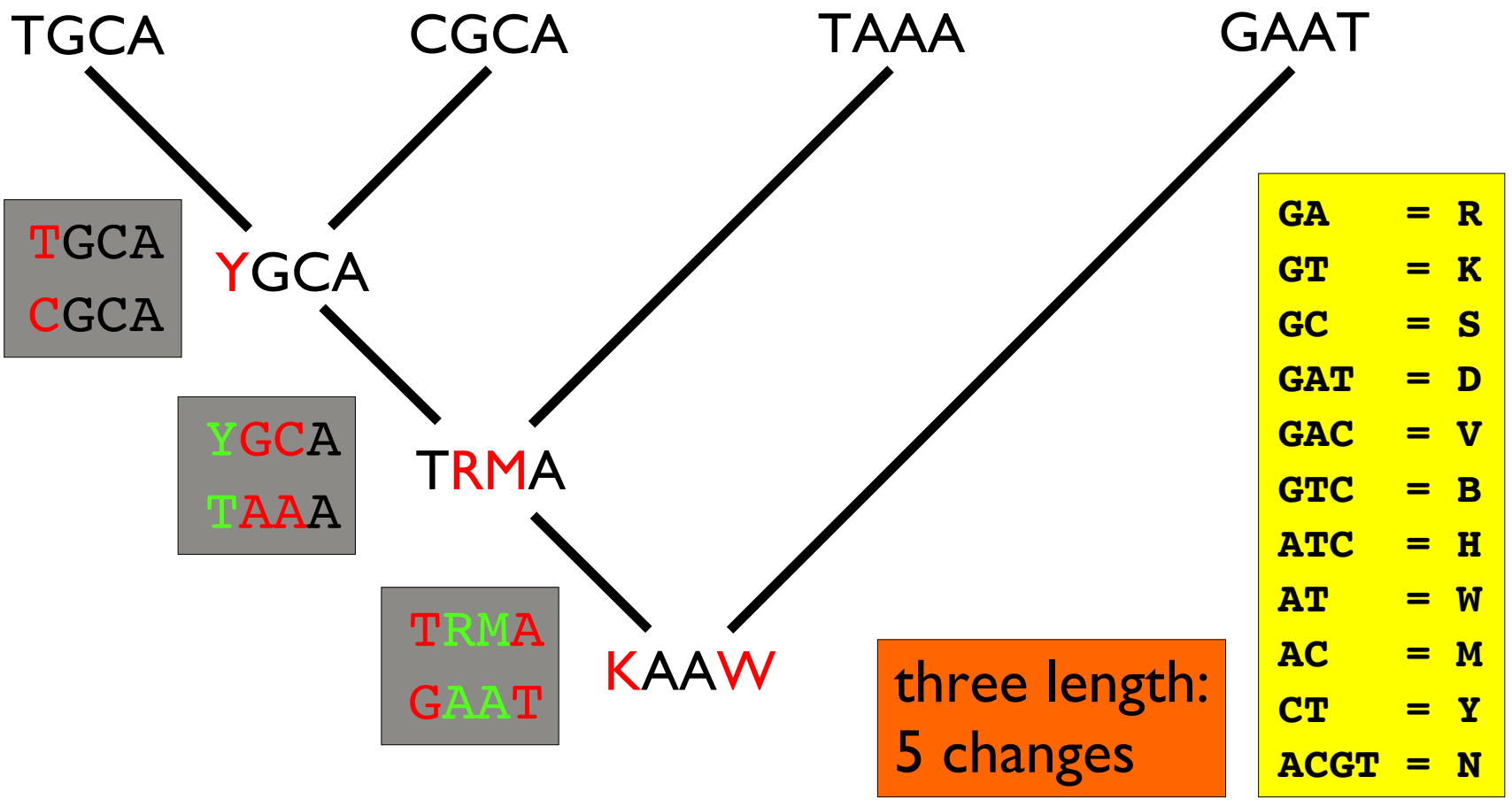


Direct optimization

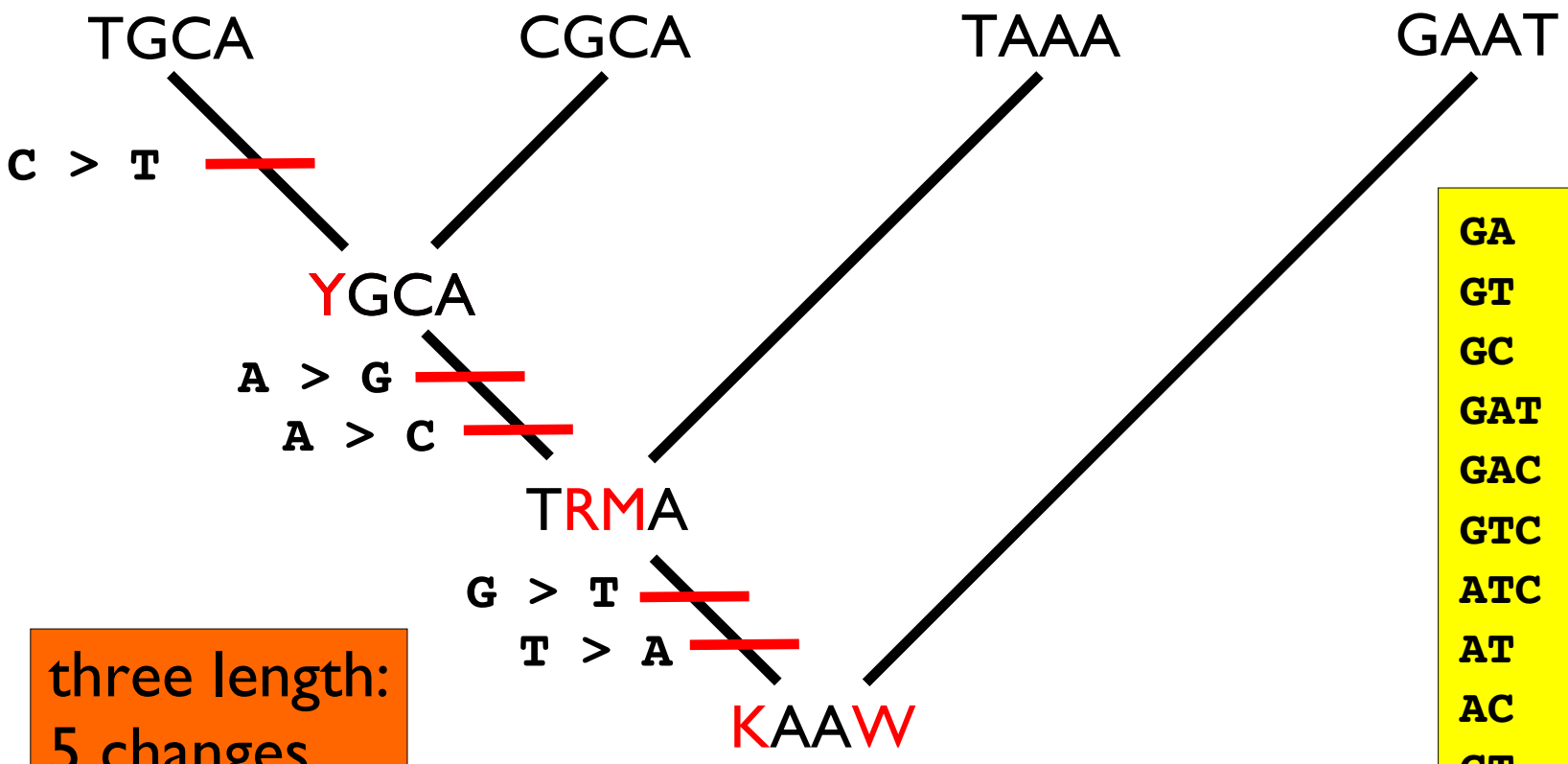


GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization



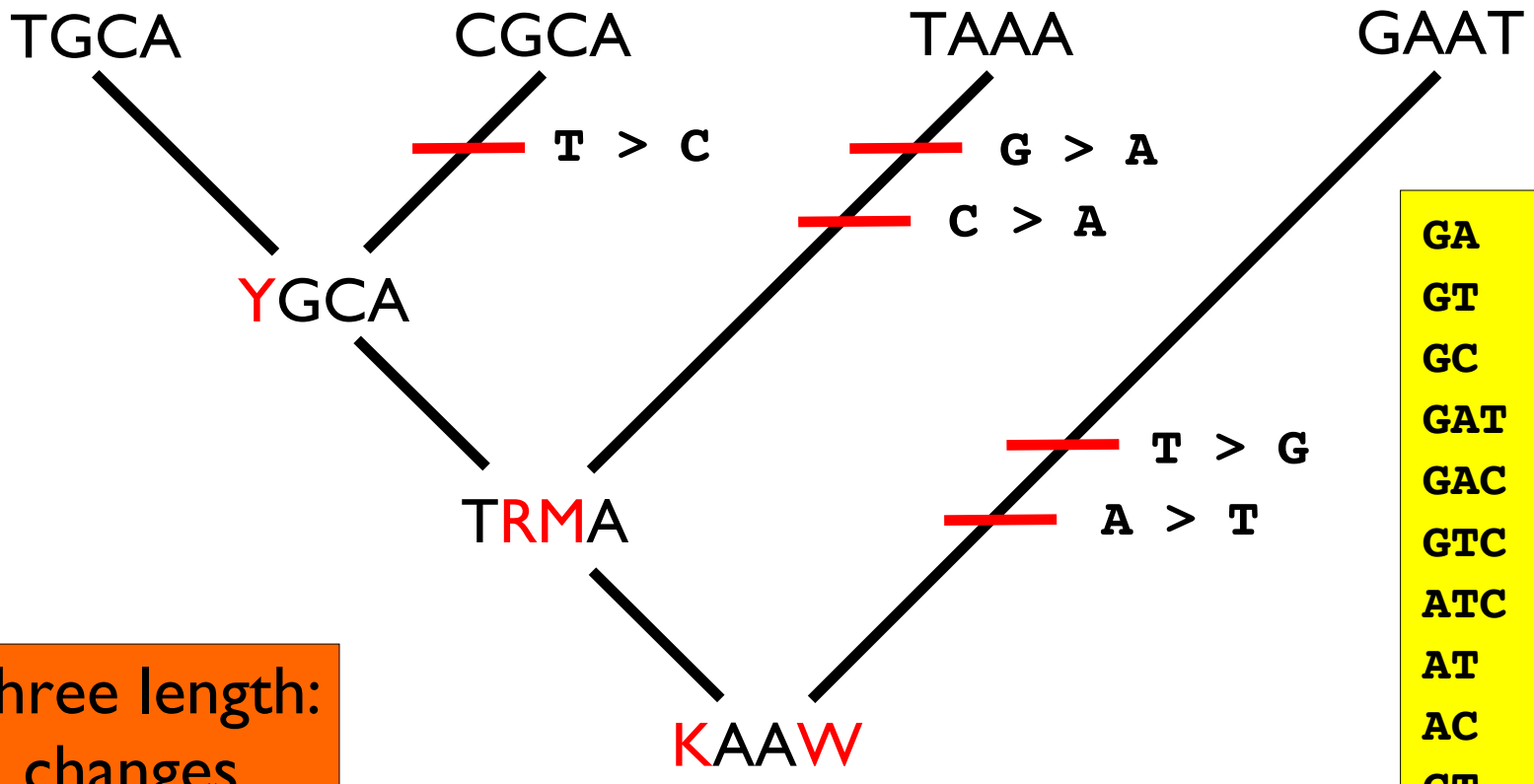
Direct optimization



three length:
5 changes

GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

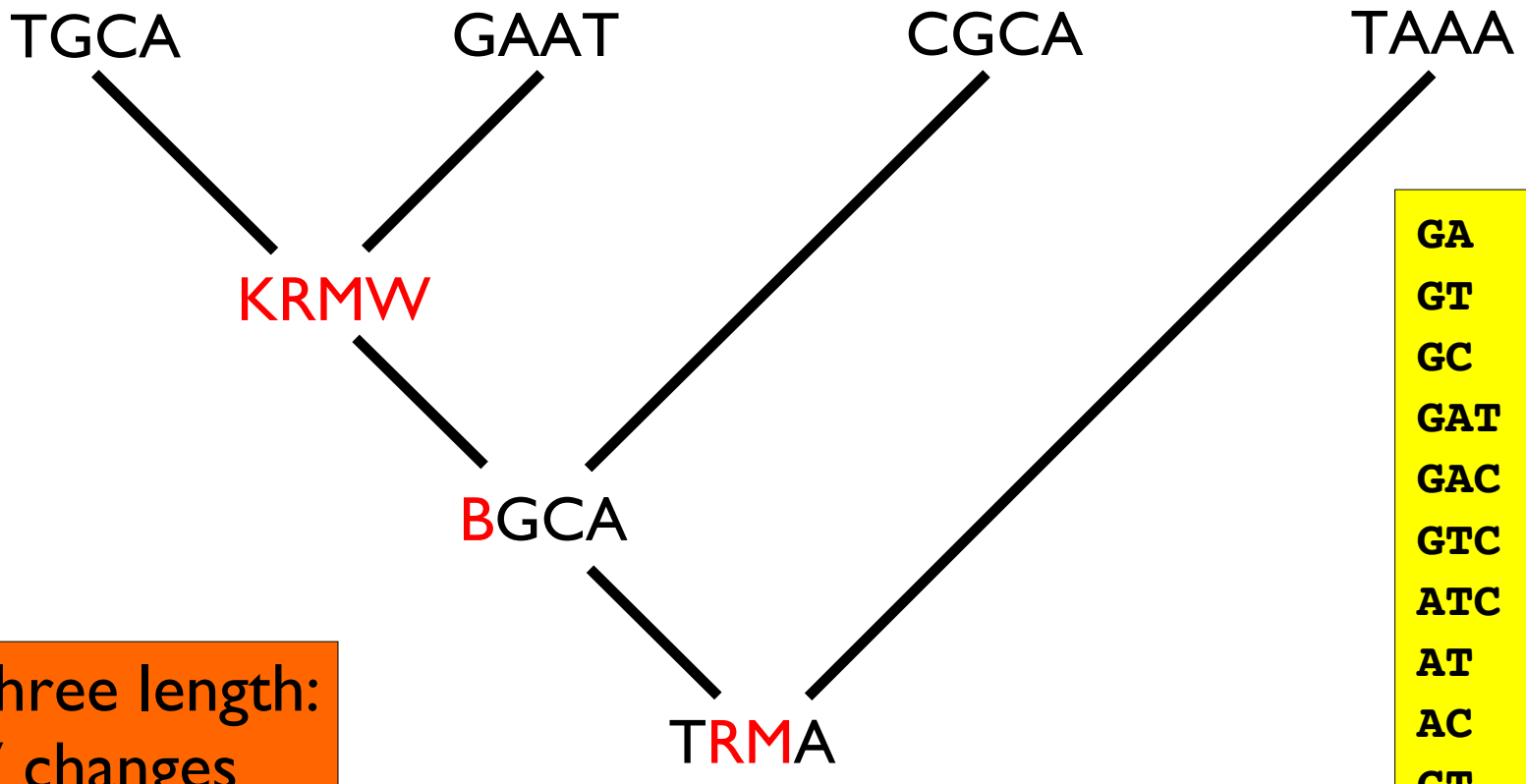
Direct optimization



three length:
5 changes

GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

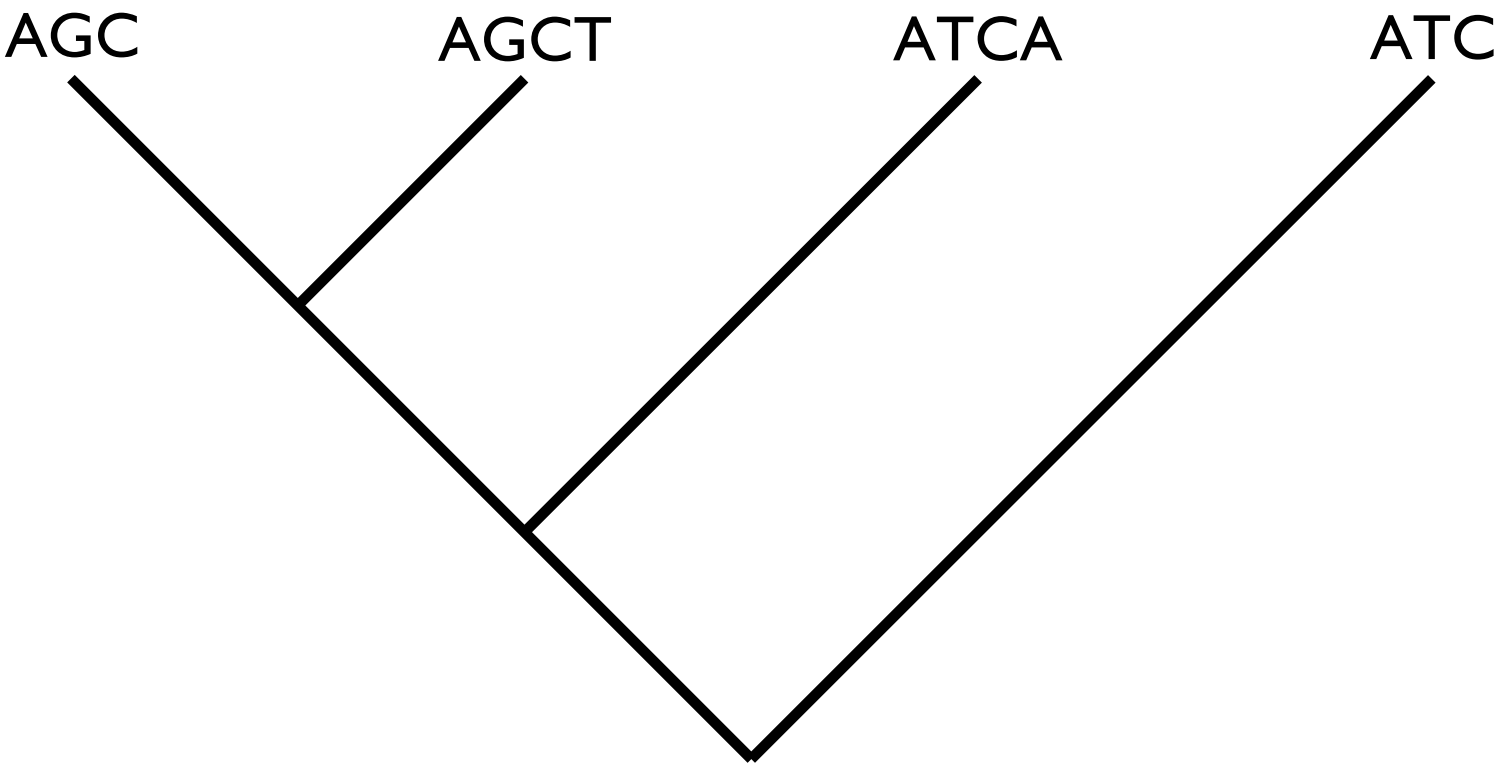
Direct optimization



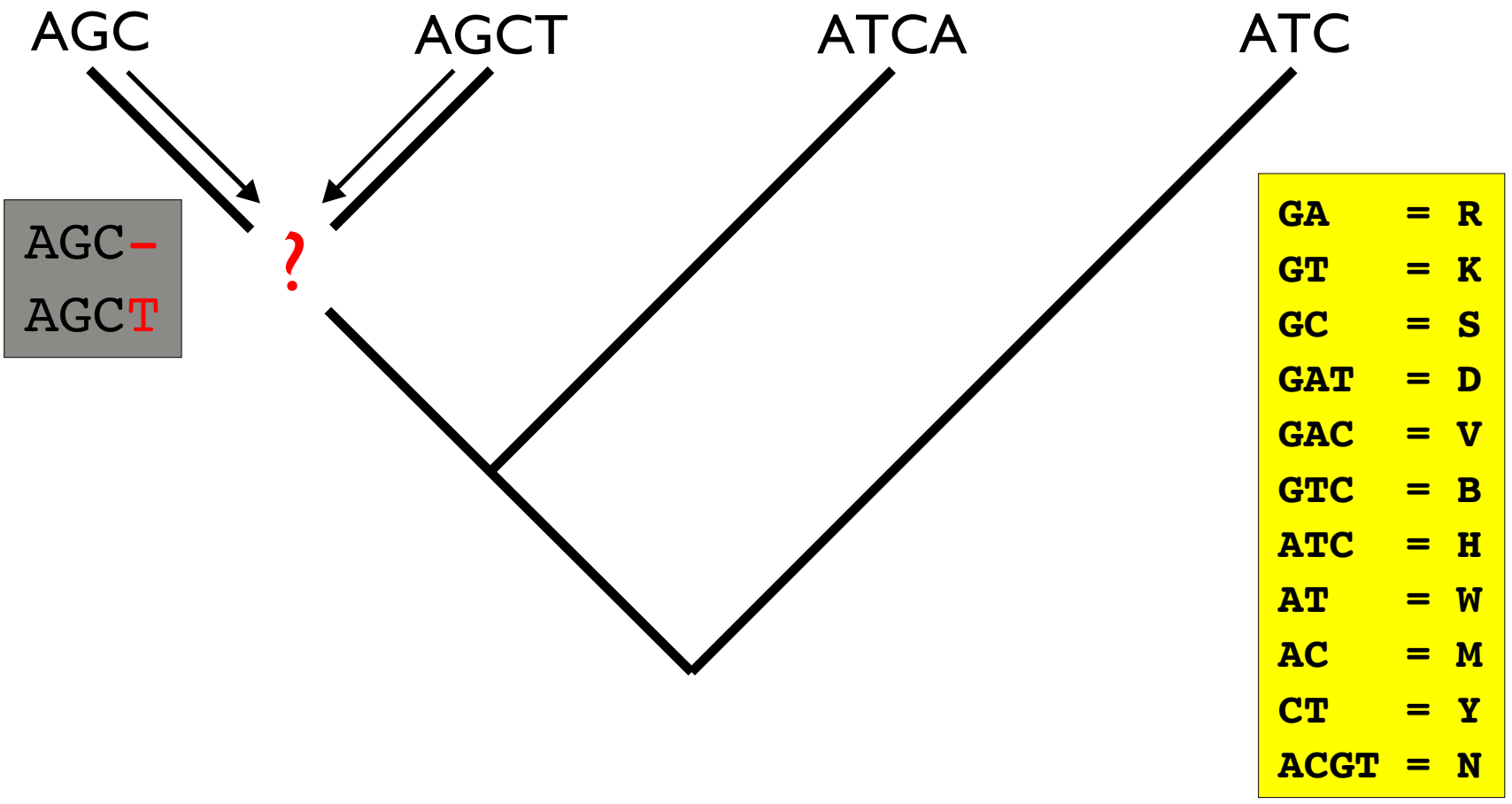
three length:
7 changes

GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

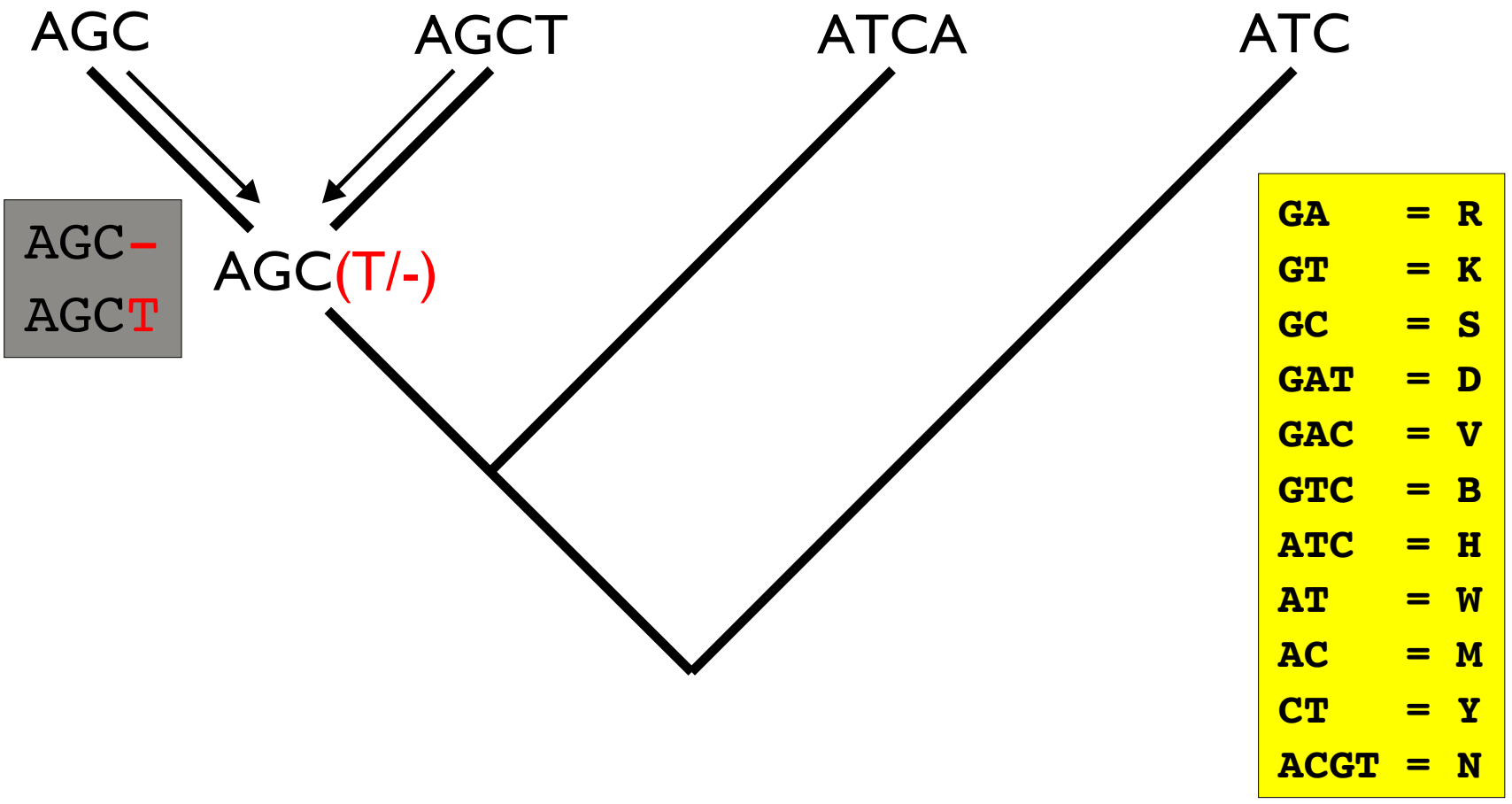
Direct optimization



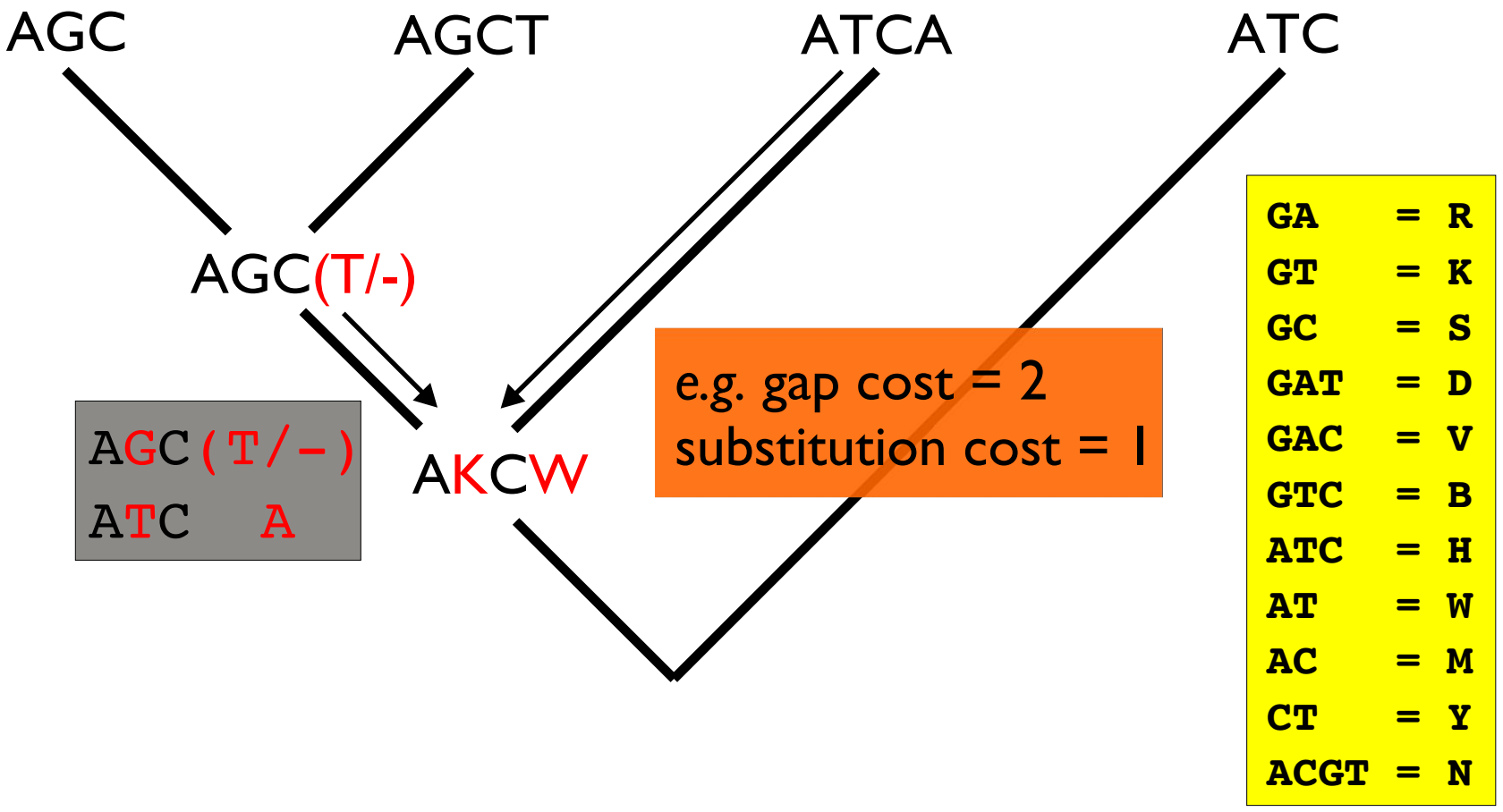
Direct optimization



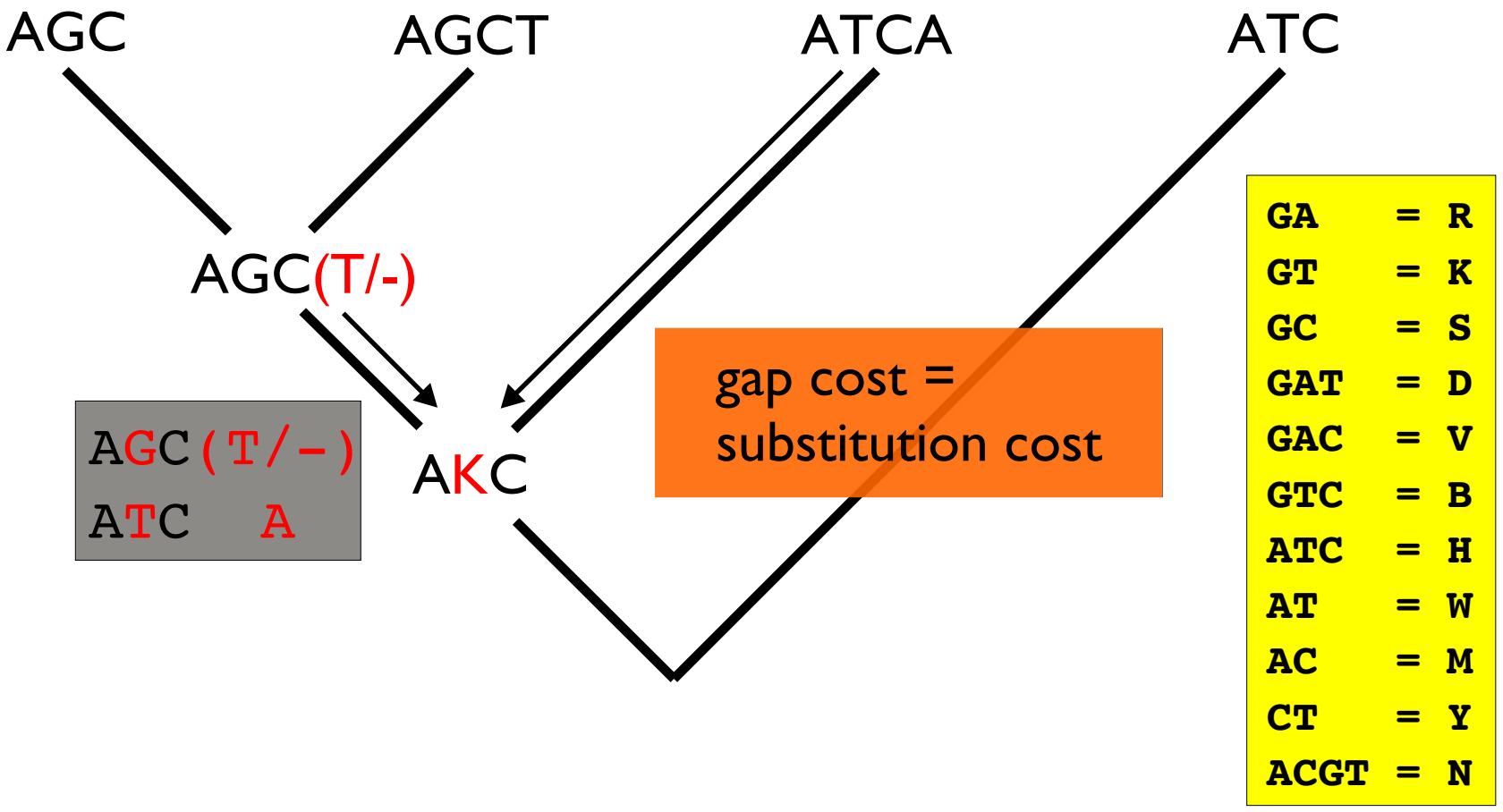
Direct optimization



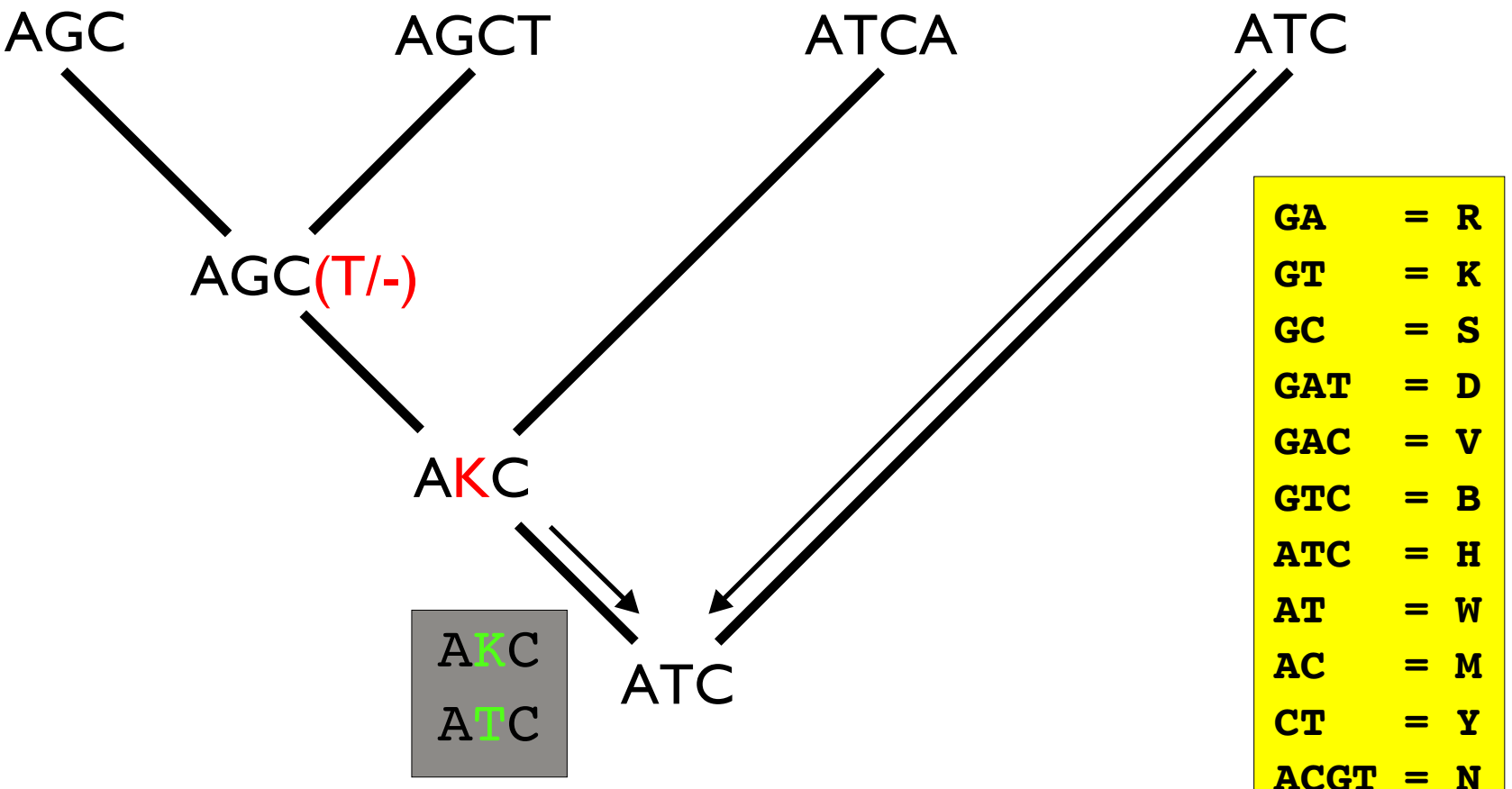
Direct optimization



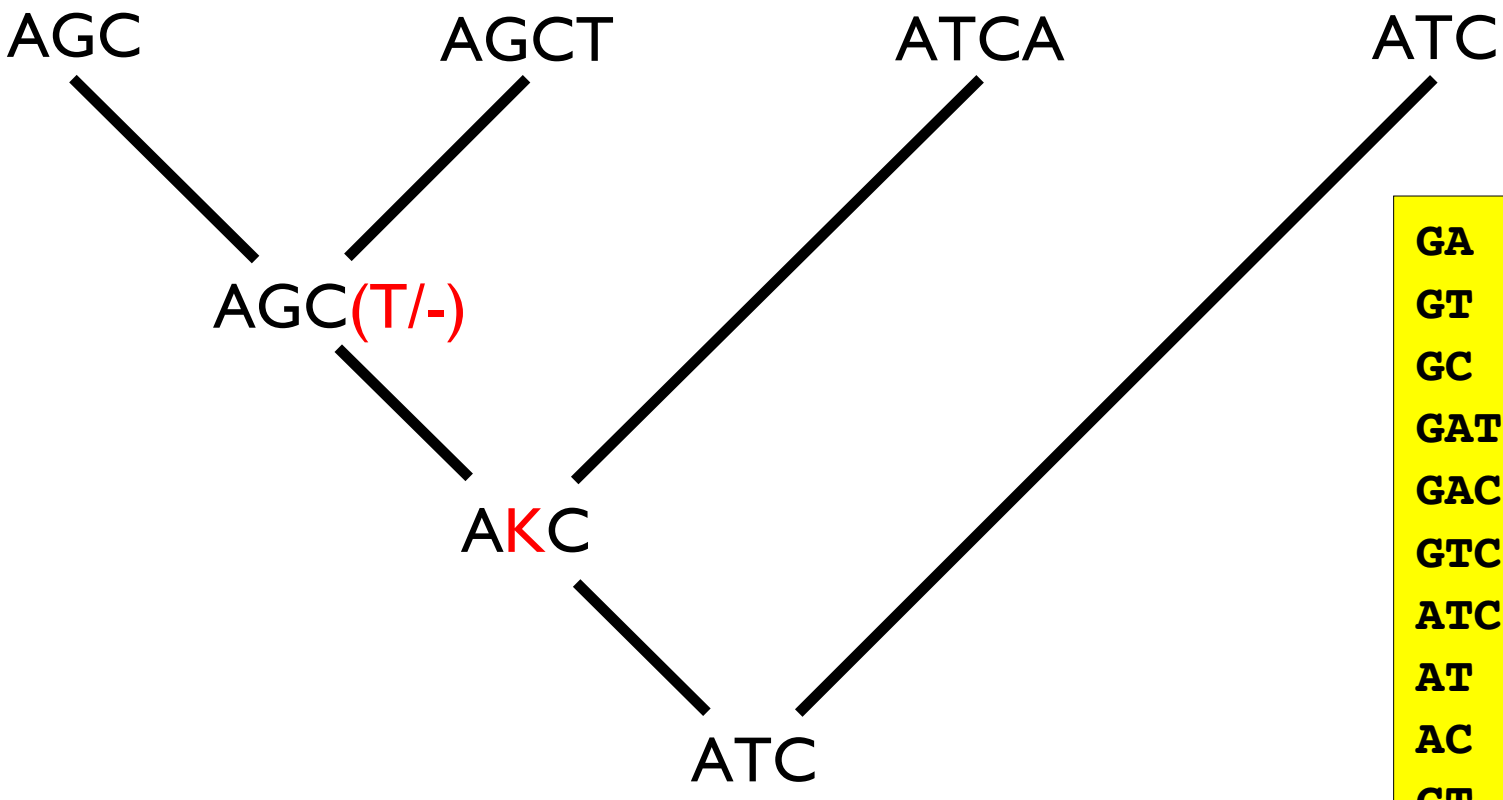
Direct optimization



Direct optimization

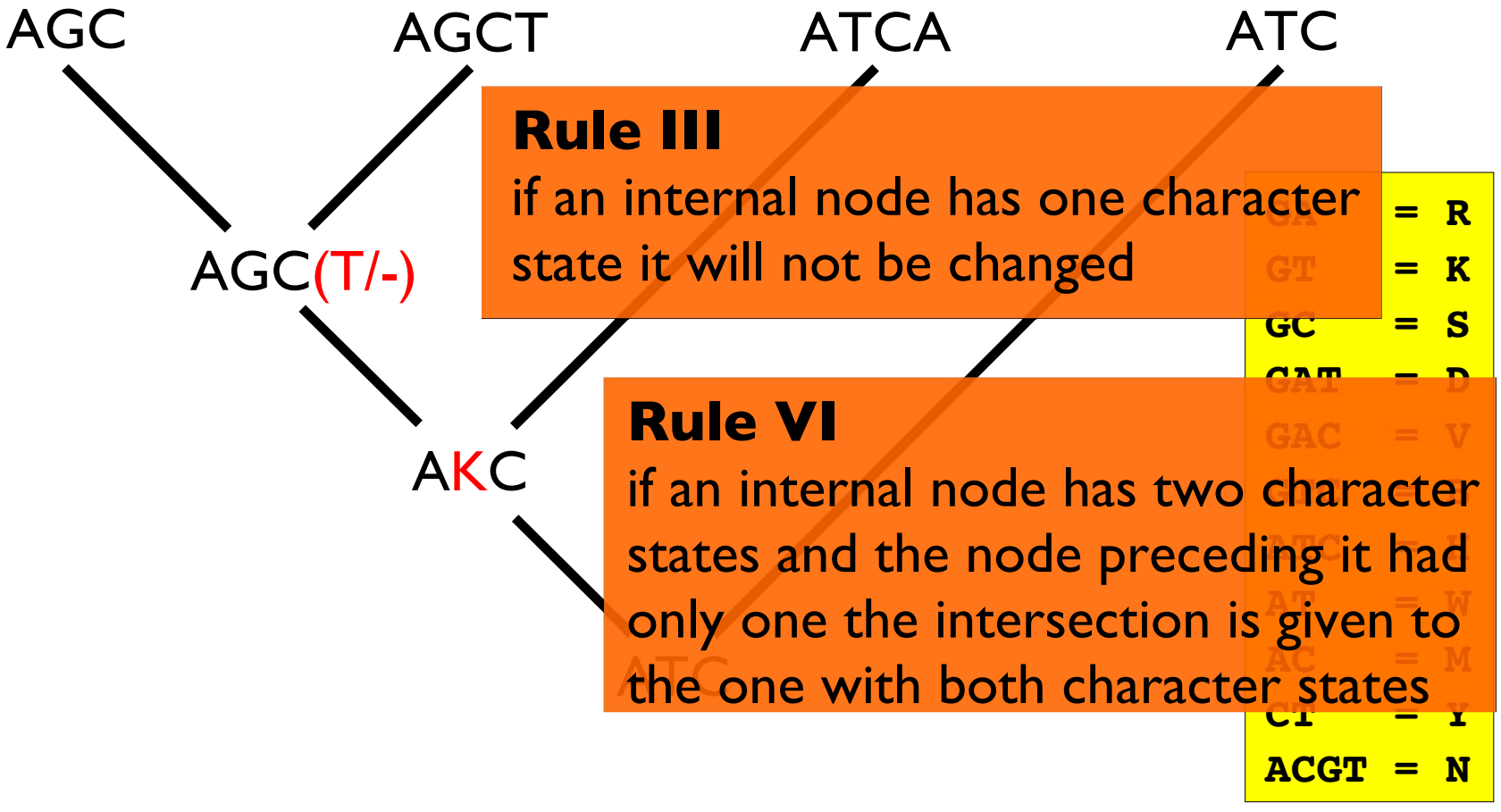


Direct optimization

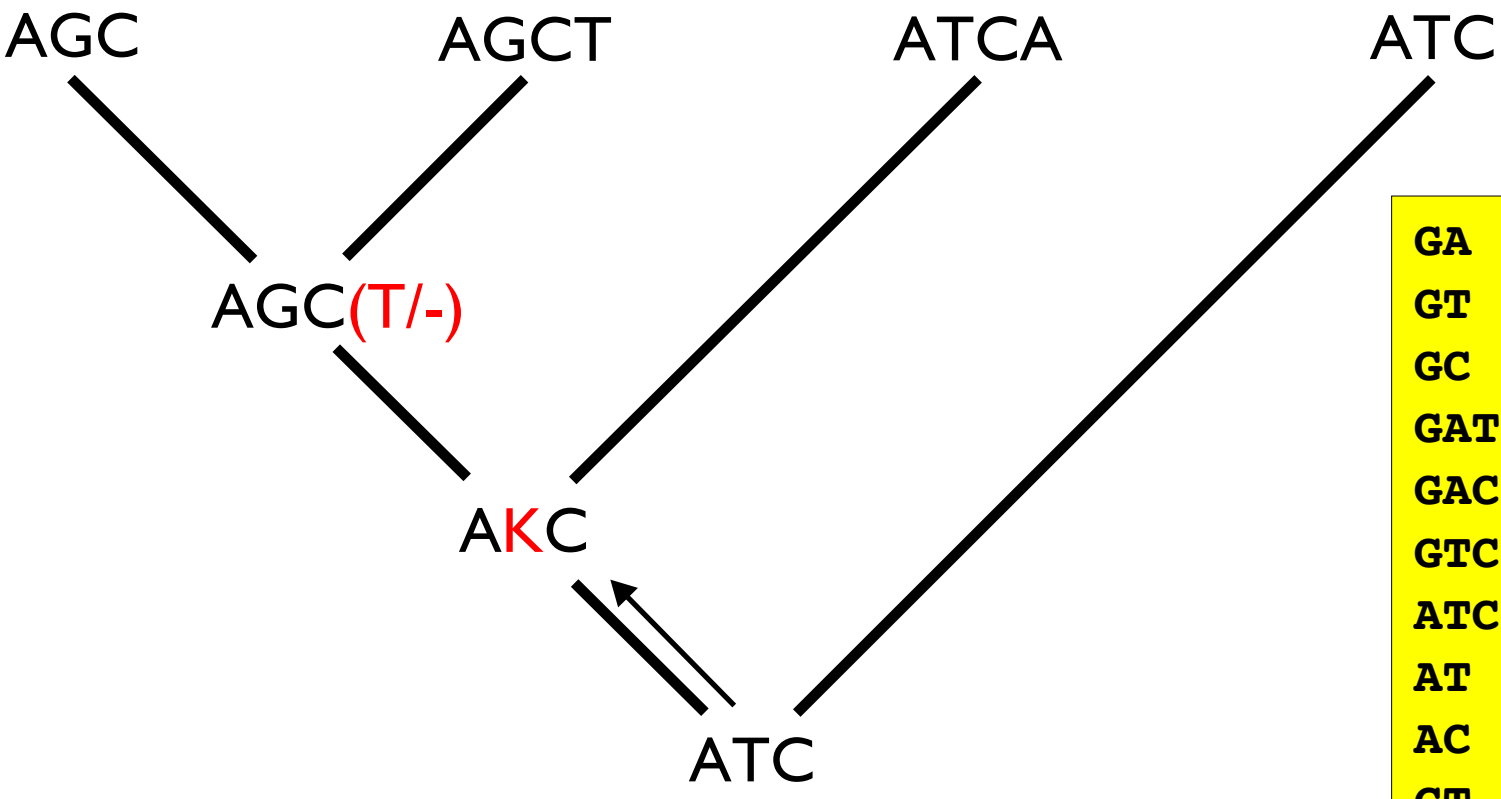


GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization

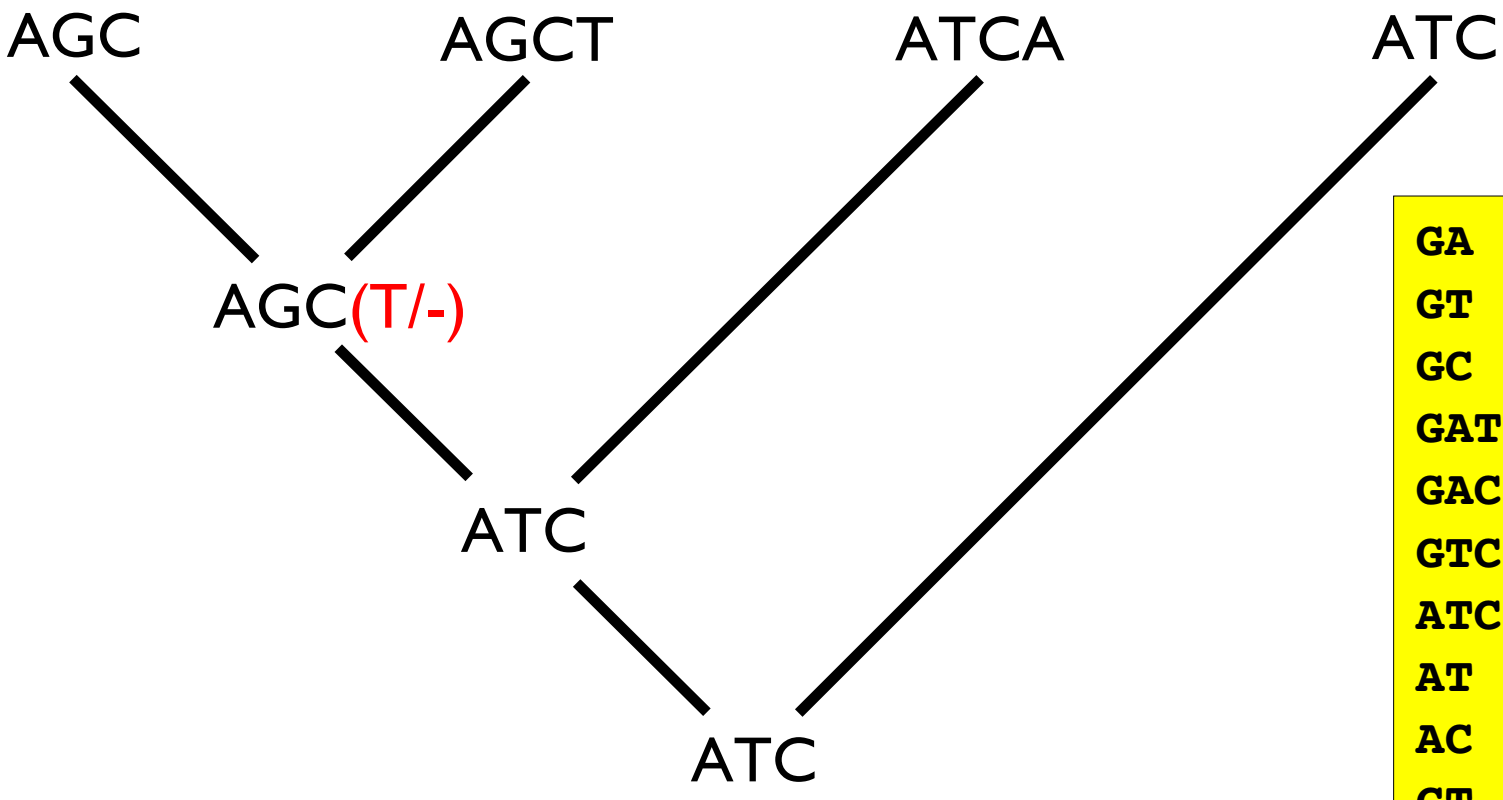


Direct optimization



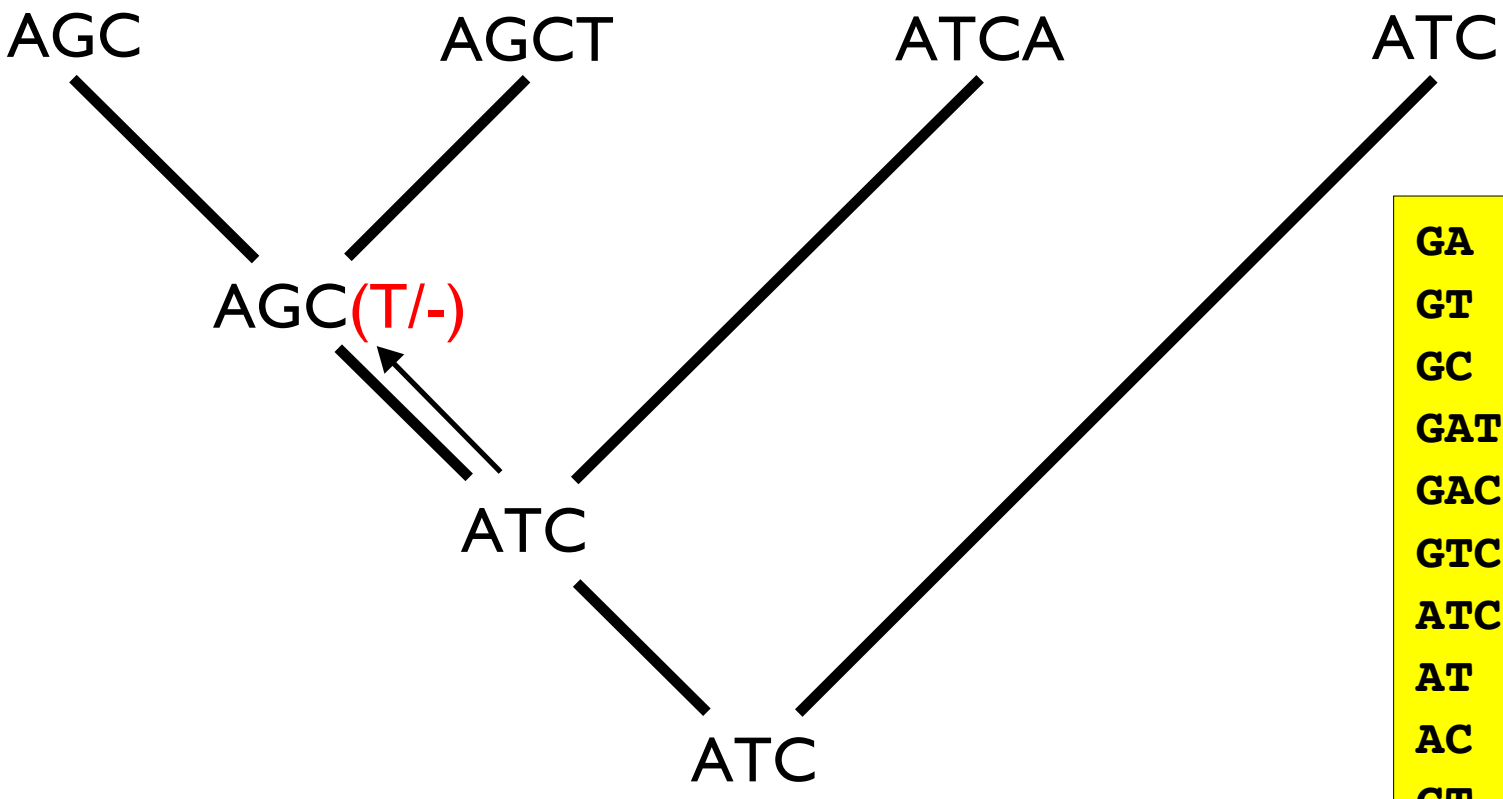
GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization



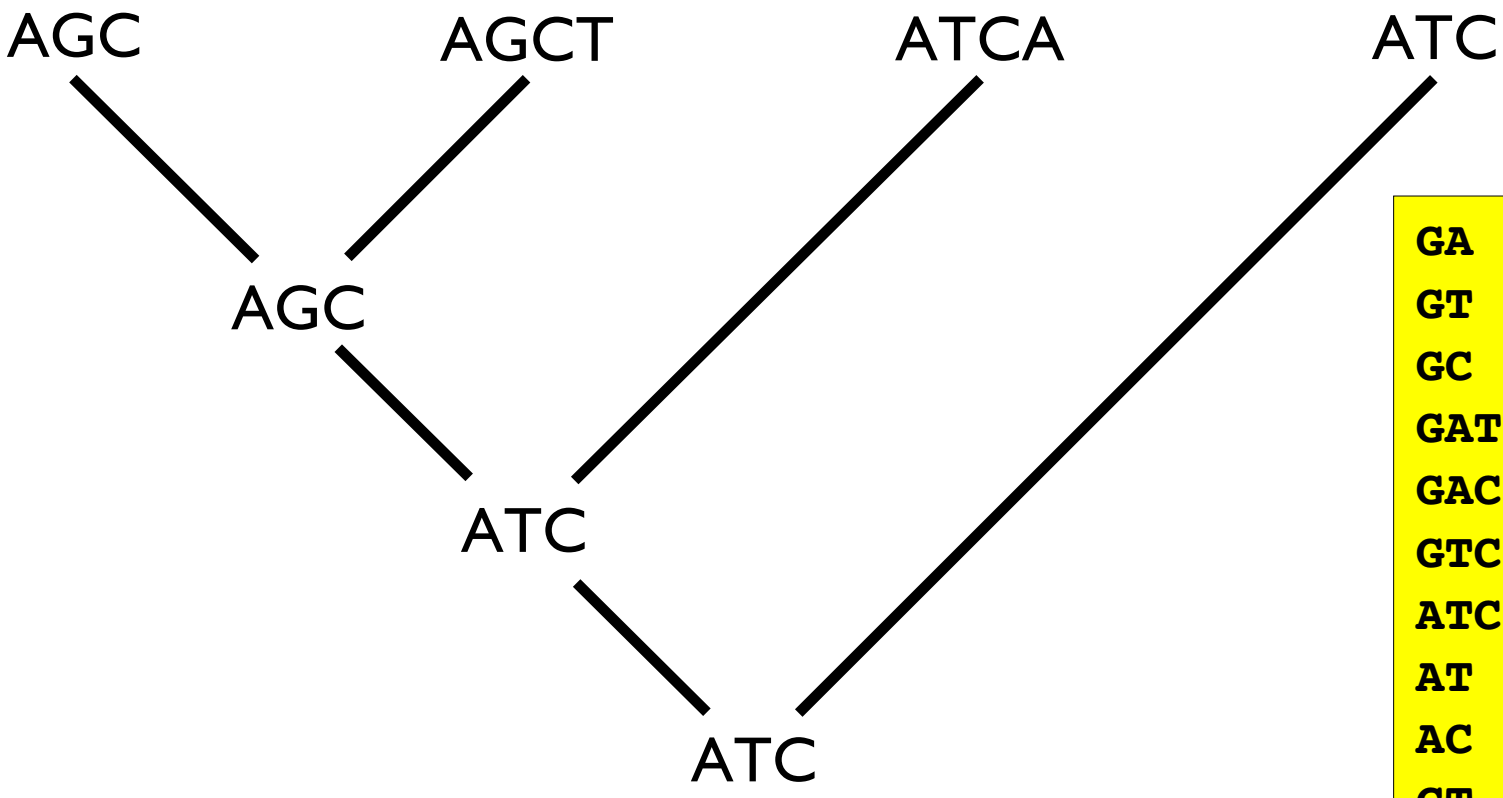
GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization



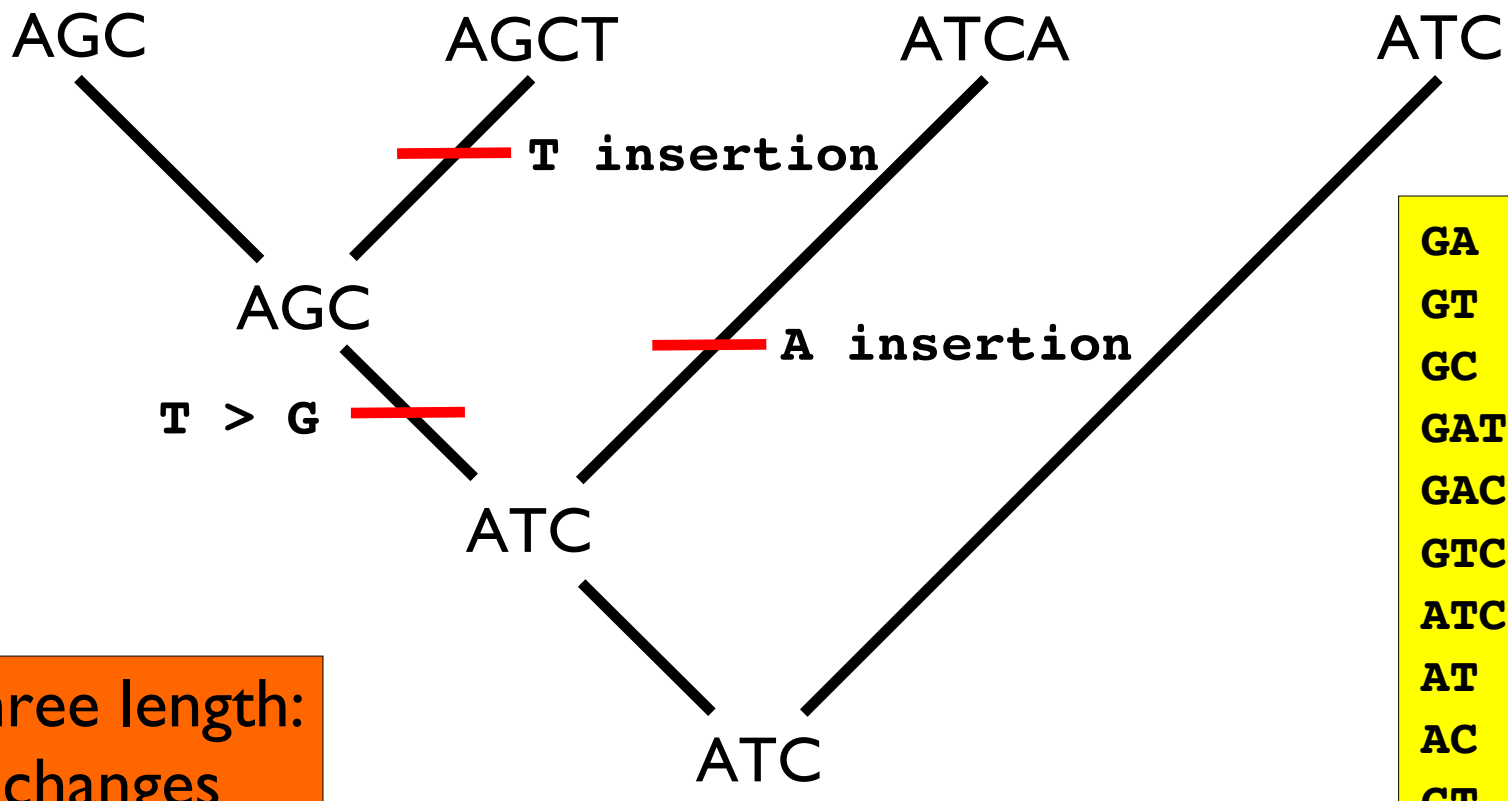
GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization



GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization



three length:
3 changes

GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization in practice

- direct optimization is implemented in the program POY
- Wheeler, W., Gladstein, D., De Laet, J. (2003)
- http://research.amnh.org/scicomp/projects/poy.php

POY

- implements classic heuristic tree search strategies, e.g. branch swapping, treedrifting, treefusing, and ratcheting
- heuristic procedures to optimize sequences on trees, Direct optimization and Fixed states is implemented and tightly integrated with the tree search heuristics

POY

- computionally demanding
- variuous techniques needed to reduce calculation time

How to reduce calculation time?

- parallel computing - essential even for moderate sized data-sets
- cut sequences into shorter fragments in CONSERVED regions
- perform analyses in pieces - use intermediate results

Lets try !