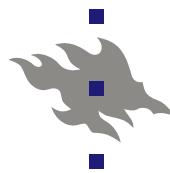


DIRECT OPTIMIZATION

Filip Högnabba
Botanical Museum
Finnish Museum of Natural History



Finnish Museum of Natural History
Botanical Museum
University of Helsinki



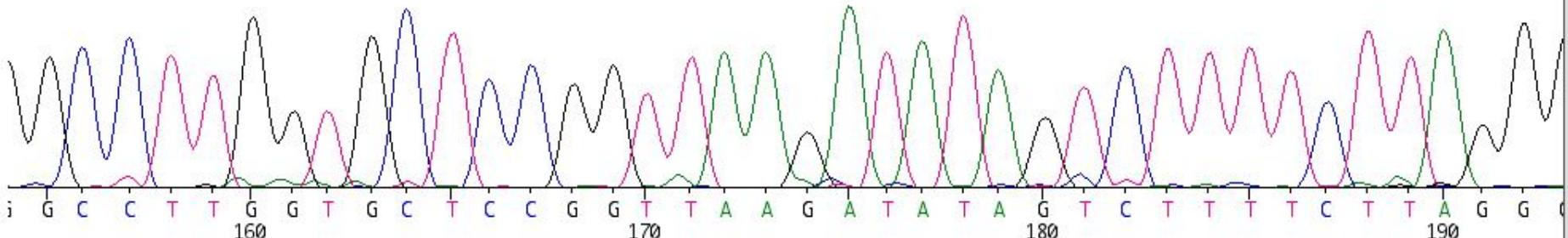
UNIVERSITY OF HELSINKI

DNA-sequences

- **OBSERVATIONS** of the nucleotide order in a studied sequence

/product="large subunit ribosomal RNA"

1 ggattgcctc agtaacggcg agtgaagcgg caacagctca aatttgaat ctggttcttt
61 cgggggccccg agttgttaatt ttttagaggat gtttcgggtg cggccgggt ctaagtccct
121 tgAACACAGGA CGTCATAGAG GGTGAGAATC CCCTATGCAG CTGGCGATCA ACCCCATGTG
181 aaACCCCTTC GACGAGTCGA GTTGTGGG AATGCAAGTC AAAATGGGTG GTAAATTCA
241 tCTAAAGCTA AATACCGGCC AGAGACCGAT AGCGCACAAG TAGAGTGTAC GAAAGATGAA
301 aAGCAGTTG GAAAGAGAGT CAAACAGTAC GTGAAATTGT TGAAGGGAA GCGCTTGCAA



From observations to phylogenetic hypotheses

DNA sequences

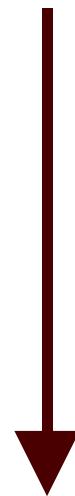


Phylogenetic hypotheses

From observations to phylogenetic 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

Stereobetaalign21i12.nex

Taxa	Characters
1	AT1080 Sten5593 S C A A G G T T T C C A G A T C A C C C A C T C C C C T T G G T G G T G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T C T C A
2	AT1148 Ober8354 S C A A G G T T T C C A G A T C A C C C A C T C C C C T T G G T G G T G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T C T C A
3	AT1031 Bald1998 S C A A G G T T T C C A G A T C A C C C A T T C C C C T T G G T G G T G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T C T C A
4	AT1194 Fe60471a S C A A G G T T T C C A G A T C A C C C A T T C C C C T T G G T G G T G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T C T C A
5	FH0062 Sch10162 S C A A G G C T T C C A G A T T A C C C A C T C C C C T T G G T G G T G G A A C C G G T G C T G G T A T G G G T A C G C T C T T G A T C T C C
6	AT1134 Ober8158 S C A A G G T T T C C A G A T C A C C C A T T C C C C T T G G T G G T G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T T T C A
7	AT1170 Ino28953 S C A A G G T T T C C A A A T C A C C C A C T C C C C T T G G T G G T G G A A C C G G G T G C T G G T A T G G G T A C T C T C T T G A T C T C A
8	FH0067 Högna336 S C A A G G T T T C C A G A T C A C C C A C T C C C C T T G G T G G T G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T C T C A
9	AT1177 Ino28950 S C A A G G T T T C C A G A T C A C C C A T T C C C C T T G G T G G T G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T C T C A
10	FH0060 Sip44112 S C A A G G T T T C C A G A T C A C C C A T T C C C C T T G G T G G T G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T C T C A
11	AT1037 Sten5289 S C A A G G T T T C C A G A T C A C C C A C T C C C C T T G G T G G T G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T C T C A
12	FH0064 Högna247 S C A A G G T T T C C A A A T C A C C C A C T C C C C T T G G T G G T G G A A C C G G G T G C T G G T A T G G G T A C T C T C T T G A T T T C A
13	AT1054 Sten5460 S C A A G G T T T C C A G A T C A C C C A C T C C C C T A G G T G G T G G G A A C T G G T G C T G G T A T G G G T A C G C T C T T G A T C T C A
14	AT1160 Ino27242 S C A A G G T T T C C A G A T C A C C C A C T C C C C T A G G T G G T G G G A A C C G G G A G C T G G T A T G G G C A C G C T C T T G A T C T C A

Non-coding DNA

- length variation present
- unambiguous homology assumptions not possible

Taxa	Characters	7	7	7	7	7	7	8	8	8	8	8	8	9	9	9	9	9	9	9	9	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4											
1	1040 Sten5499 S.a	A	T	G	G	C	C	A	-	C	C	G	C	G	G	-	C	G	G	G	G	T	T	G	T	G	C	C	G	T	C	C	C	C	T	T	A	-	T	C	A	G	T	G	T	C																	
2	1133 Ober8202 S.m	A	T	G	G	A	C	A	-	C	C	G	C	G	G	-	C	G	G	G	G	T	T	G	T	G	C	C	G	T	C	C	C	T	T	A	-	T	C	A	G	T	G	T	C																		
3	1154 Ober8644 S.f	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	T	T	G	G	C	C	G	T	C	C	C	T	T	C	C	A	-	C	C	T	A	T	C	A	G	T	G	T	C															
4	FH74 Högna382 S.fa	A	T	G	G	C	C	C	-	C	C	G	C	C	G	G	-	C	G	G	G	G	T	T	G	T	G	C	C	G	T	C	C	C	T	A	-	T	C	C	A	-	G	A	T	C	A	G	T	G	T	C											
5	1181 Sas13826 S.j	A	C	G	G	A	C	G	-	C	C	G	C	C	A	G	T	G	G	G	G	G	T	T	G	D	G	C	C	G	T	C	C	T	T	G	-	T	C	C	A	-	T	C	C	T	G	T	C														
6	1193 Muhr2001 M.u	C	T	G	G	C	C	A	-	C	C	T	C	C	C	-	-	G	G	G	G	G	T	C	T	T	G	C	C	G	T	O	A	A	C	A	G	C	C	C	T	-	-	T	C	A	G	T	G	T	C												
7	1035 Aht60909 S.p	A	T	G	G	A	C	-	G	C	C	G	C	C	G	-	G	C	G	G	G	G	T	T	G	T	G	C	C	G	T	O	C	C	T	A	-	T	C	C	T	G	T	T	-	-	T	A	T	C	A	G	T	G	T	C							
8	1071 Sten5287 S.g	A	C	G	G	A	C	-	A	C	C	G	C	C	G	-	G	C	G	G	G	G	T	T	G	C	G	C	C	G	T	O	C	C	T	T	-	T	T	A	T	C	A	G	T	G	T	C															
9	1086 Sten5596 S.x	A	T	G	G	A	C	-	G	C	C	G	C	C	G	-	G	C	G	G	G	G	T	T	G	T	G	C	C	G	T	O	C	C	T	T	-	T	A	T	C	A	G	T	G	T	C																
10	1172 Ino28949 S.f	A	A	G	G	-	-	-	C	C	C	T	T	G	-	-	A	C	G	G	G	G	T	G	T	G	C	C	G	T	O	C	C	T	T	C	C	A	-	C	C	T	A	T	T	-	-	T	A	T	C	A	G	T	G	T	C						
11	1164 Ino28664 S.de	A	A	G	G	A	C	-	G	C	C	G	G	C	G	G	G	G	G	G	G	T	C	G	T	B	C	C	G	T	O	C	C	T	T	C	C	A	-	C	C	T	A	T	T	-	-	T	A	T	C	A	G	T	G	T	C						
12	AF517927 Ster.pile	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	T	T	G	-	G	C	C	G	T	O	C	C	T	T	T	G	T	C	T	A	A	-	C	C	T	G	T	T	-	-	T	A	T	C	A	G	T	G	T	C
13	FH78 Dahl2003 S.to	A	G	G	C	C	A	-	C	C	C	T	C	C	G	-	G	G	G	G	G	G	T	T	G	T	G	C	C	G	T	O	G	A	T	A	G	C	C	-	C	-	T	G	T	T	-	-	T	A	T	C	A	G	T	G	T	C					
14	1044 Sten5523 S.a	A	T	G	G	C	C	A	-	C	C	G	C	C	G	-	G	C	G	G	G	G	T	T	G	T	G	C	C	G	T	O	C	C	T	T	A	T	C	C	A	-	C	C	T	G	T	T	-	-	T	A	T	C	A	G	T	G	T	C			
15	1134 Ober8158 S.x	A	T	G	G	A	C	-	A	C	C	G	C	C	G	-	G	C	G	G	G	G	T	T	G	T	G	C	C	G	T	O	C	C	T	T	G	T	C	C	A	-	C	C	T	G	T	T	-	-	T	A	T	C	A	G	T	G	T	C			

Alignment

- critically important in phylogenetic studies
- a dataset aligned according to different criteria or indel treatments may support contradictory 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	117629959485121
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

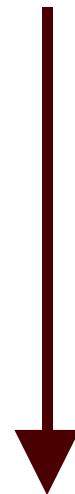
Slowinski (1998)

From observations to phylogenetic hypotheses

DNA sequences

- 1) alignment
- 2) phylogenetic analyses

Problematic

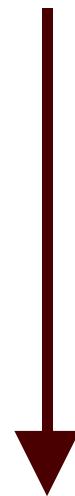


Phylogenetic hypotheses

From observations to phylogenetic hypotheses

DNA sequences

- 1) alignment
- 2) phylogenetic analyses



Direct optimization

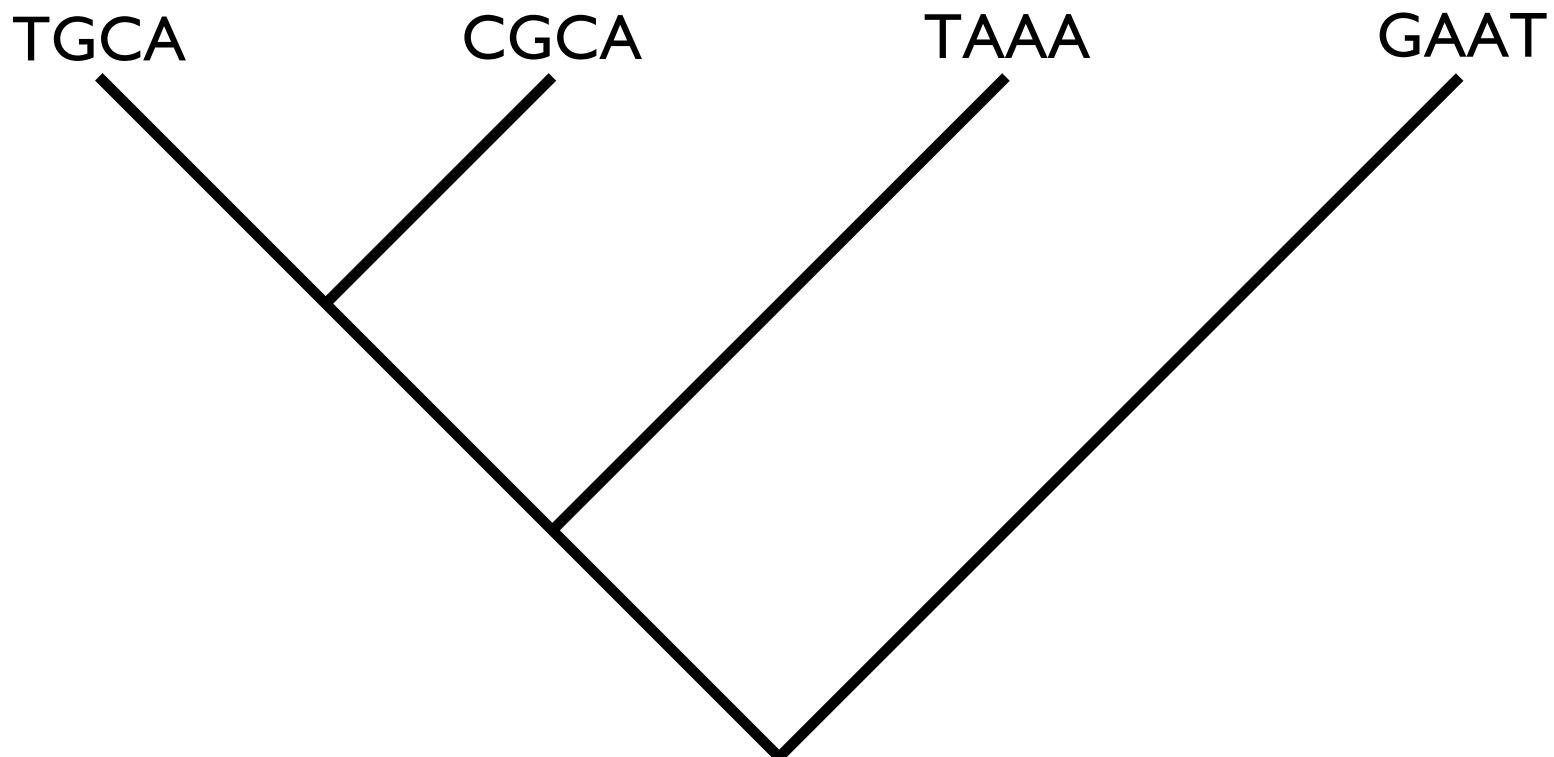
Wheeler (1996)

Phylogenetic hypotheses

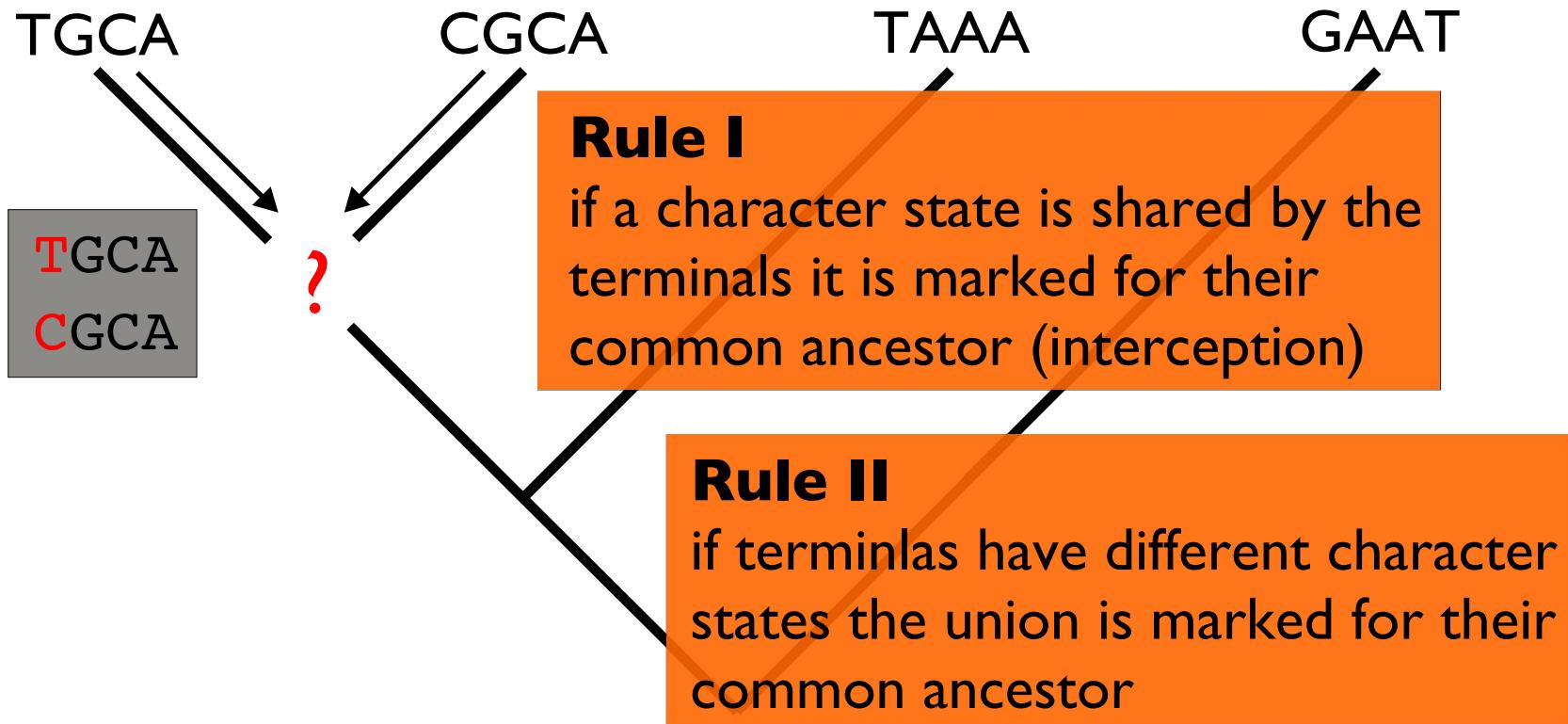
Direct optimization

- 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

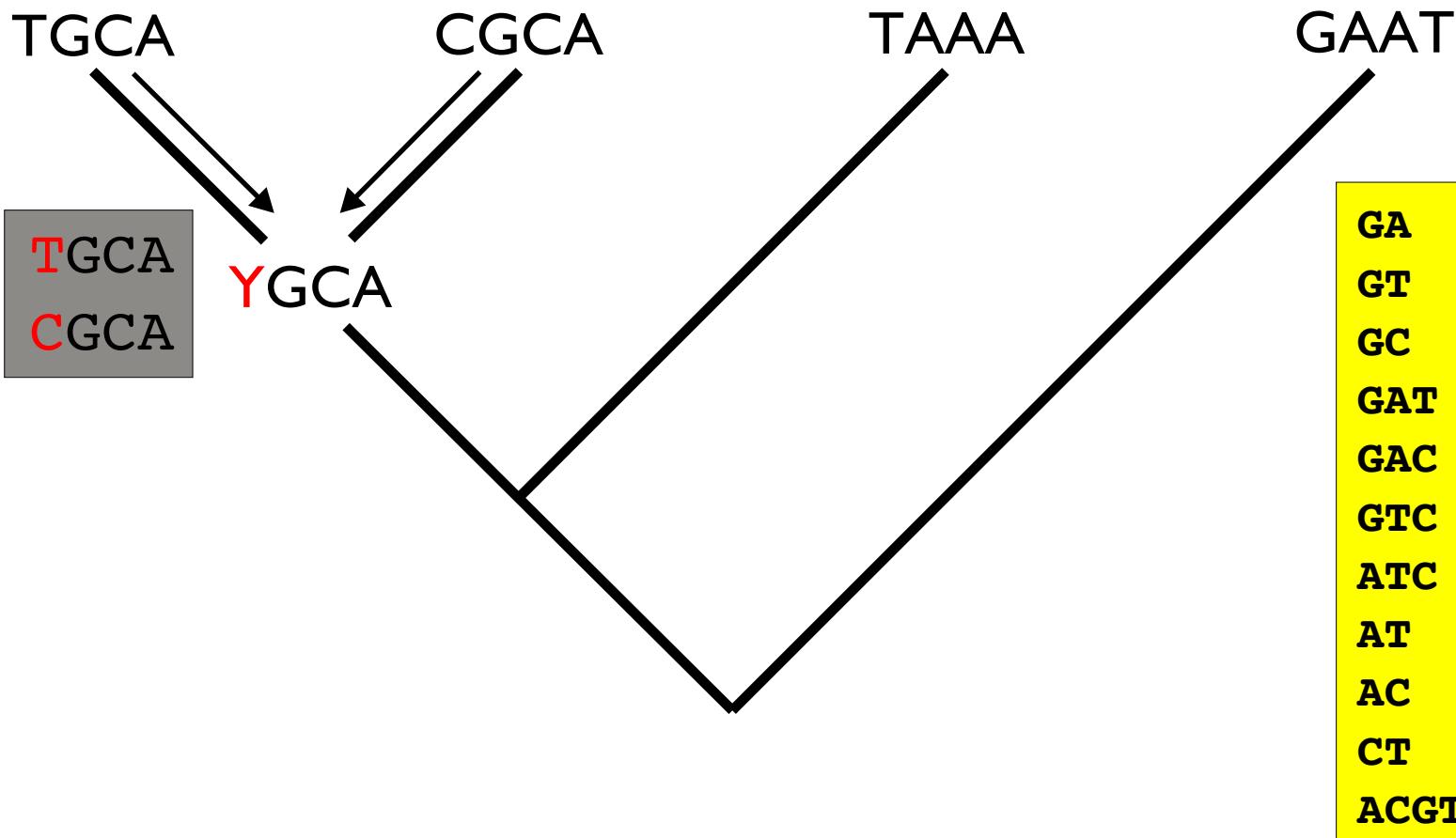
Direct optimization



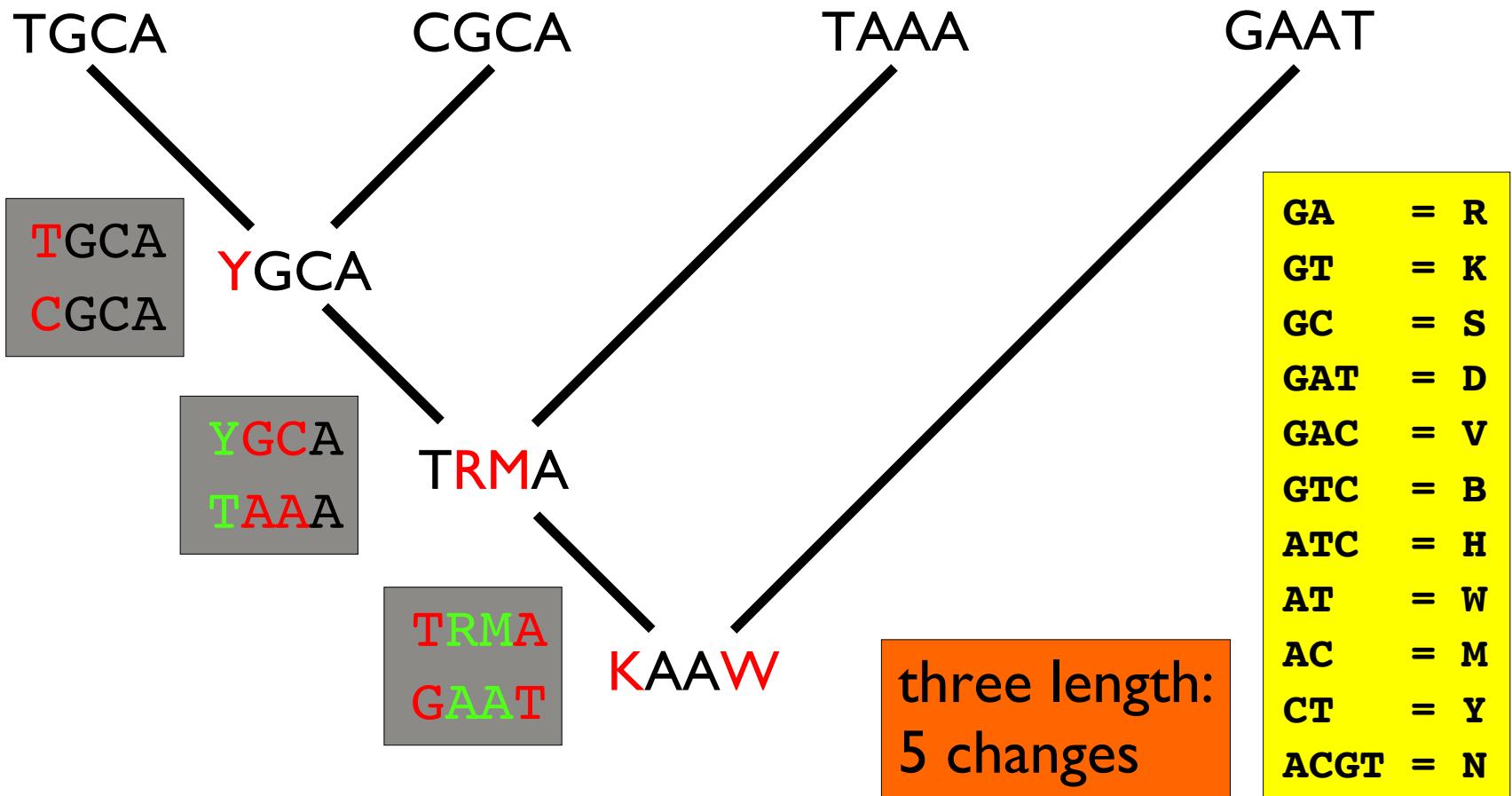
Direct optimization



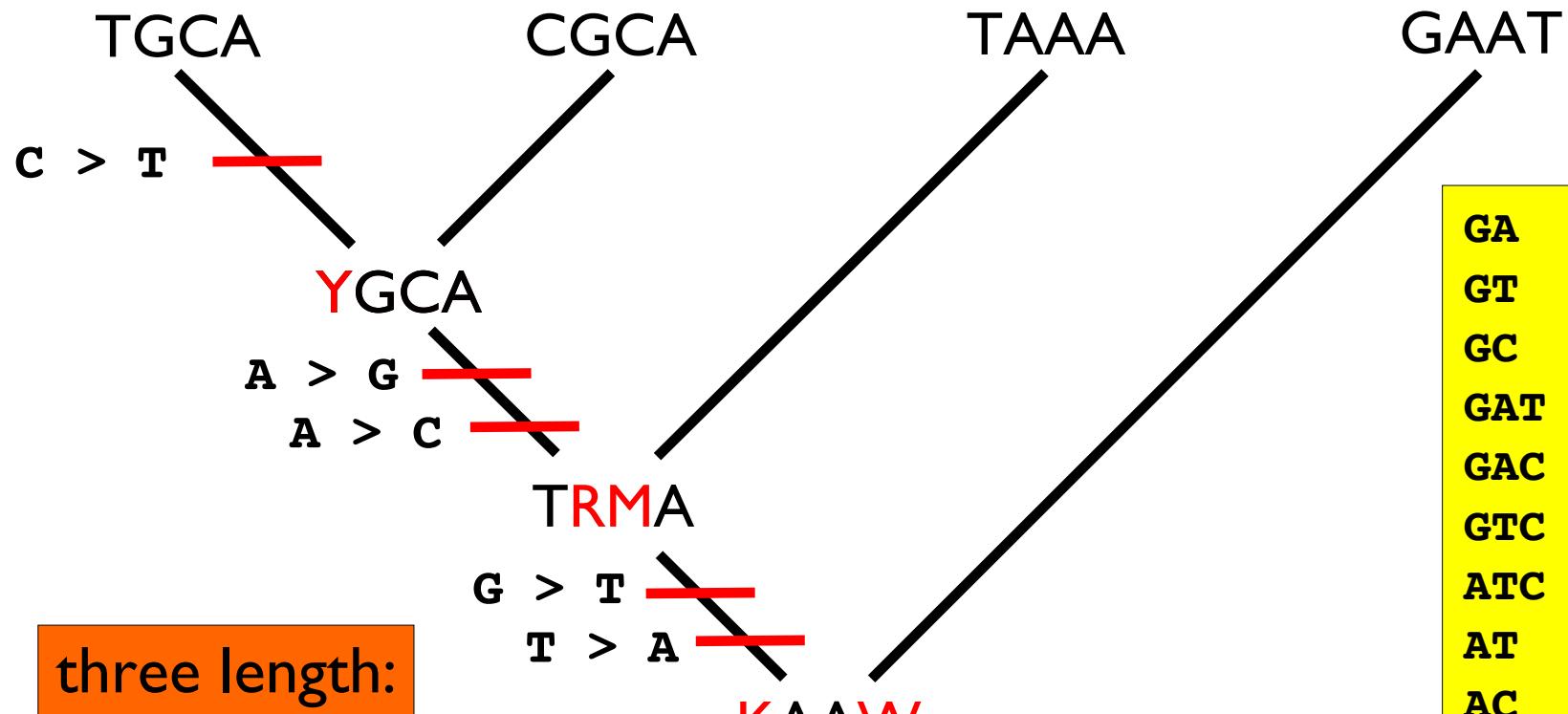
Direct optimization



Direct optimization

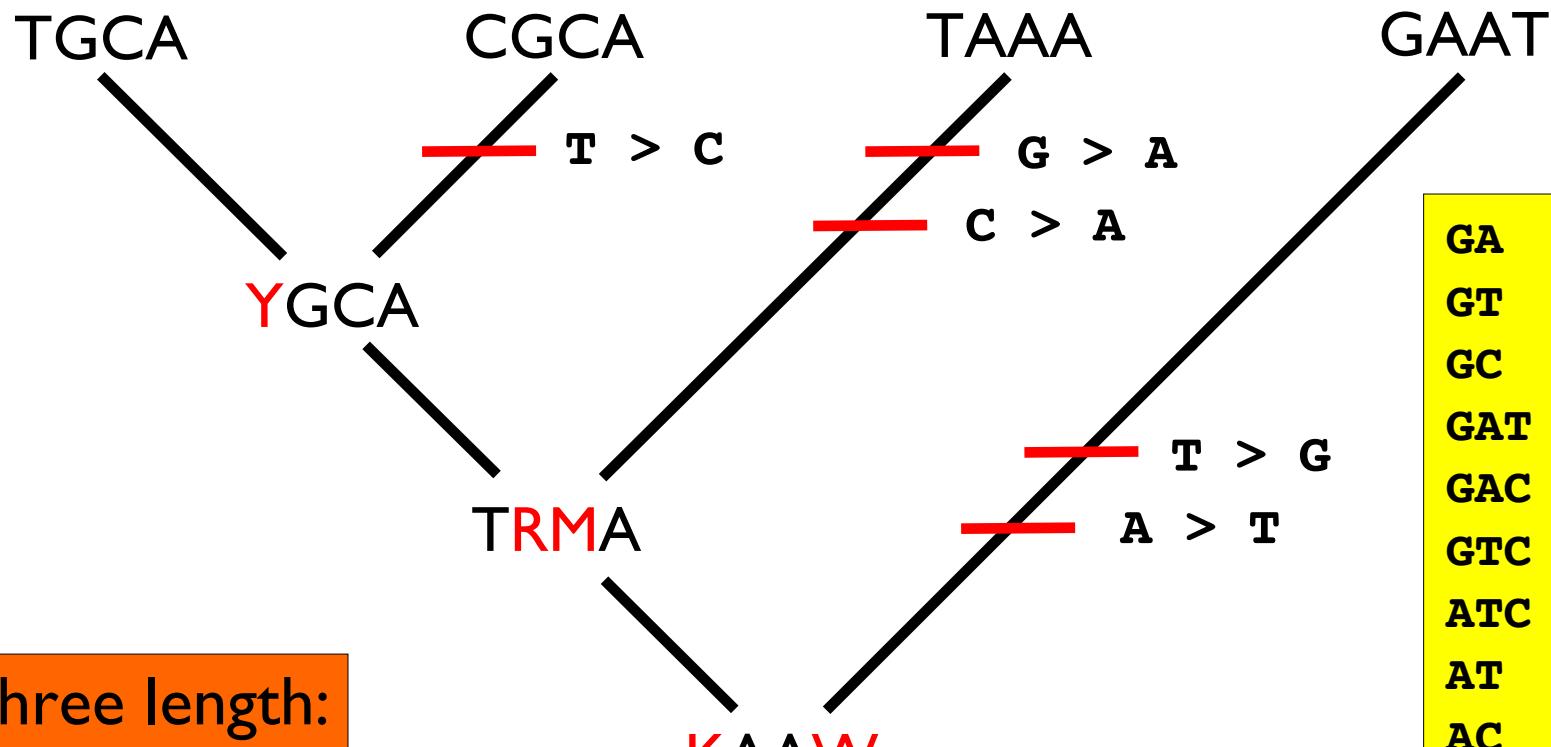


Direct optimization



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

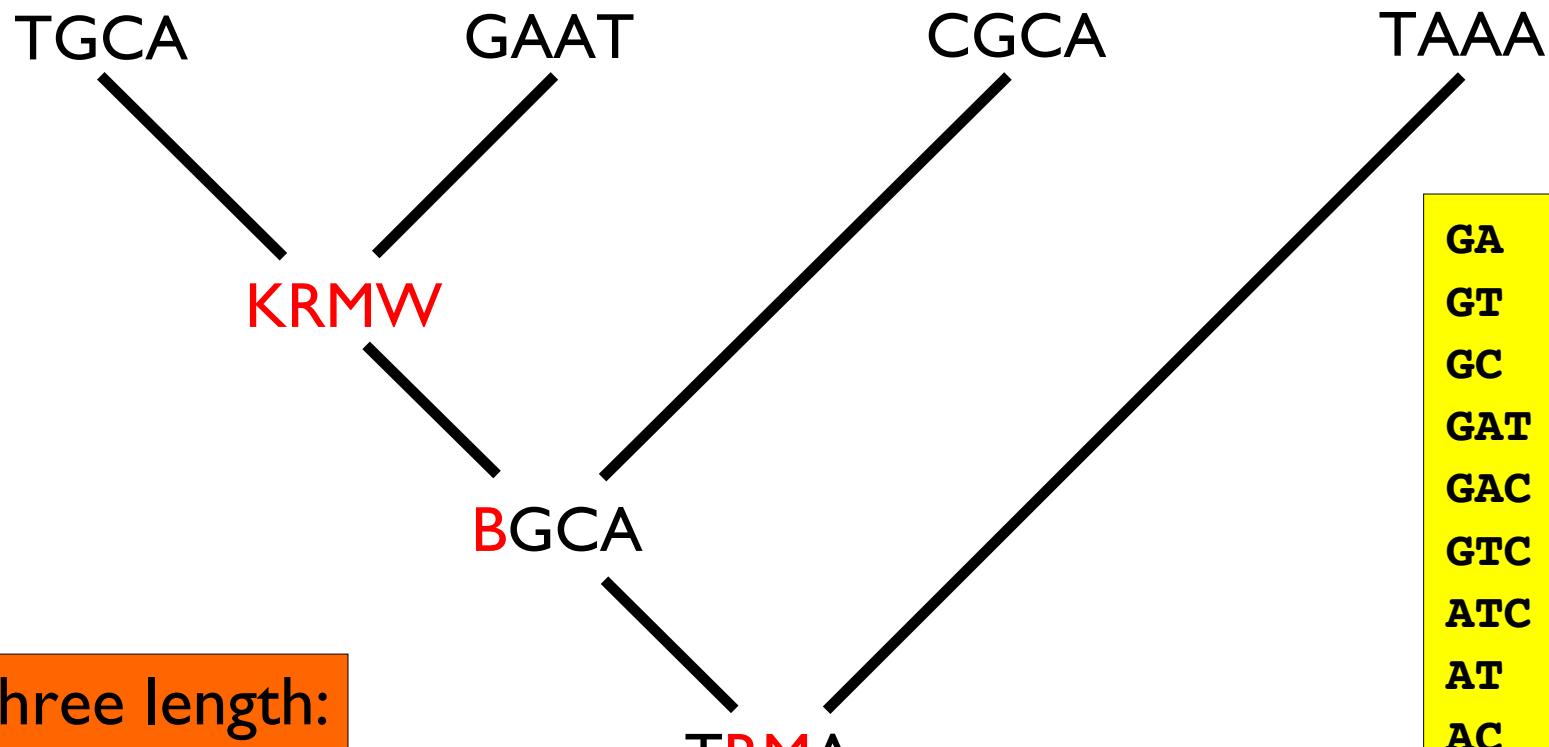
Direct optimization



three length:
5 changes

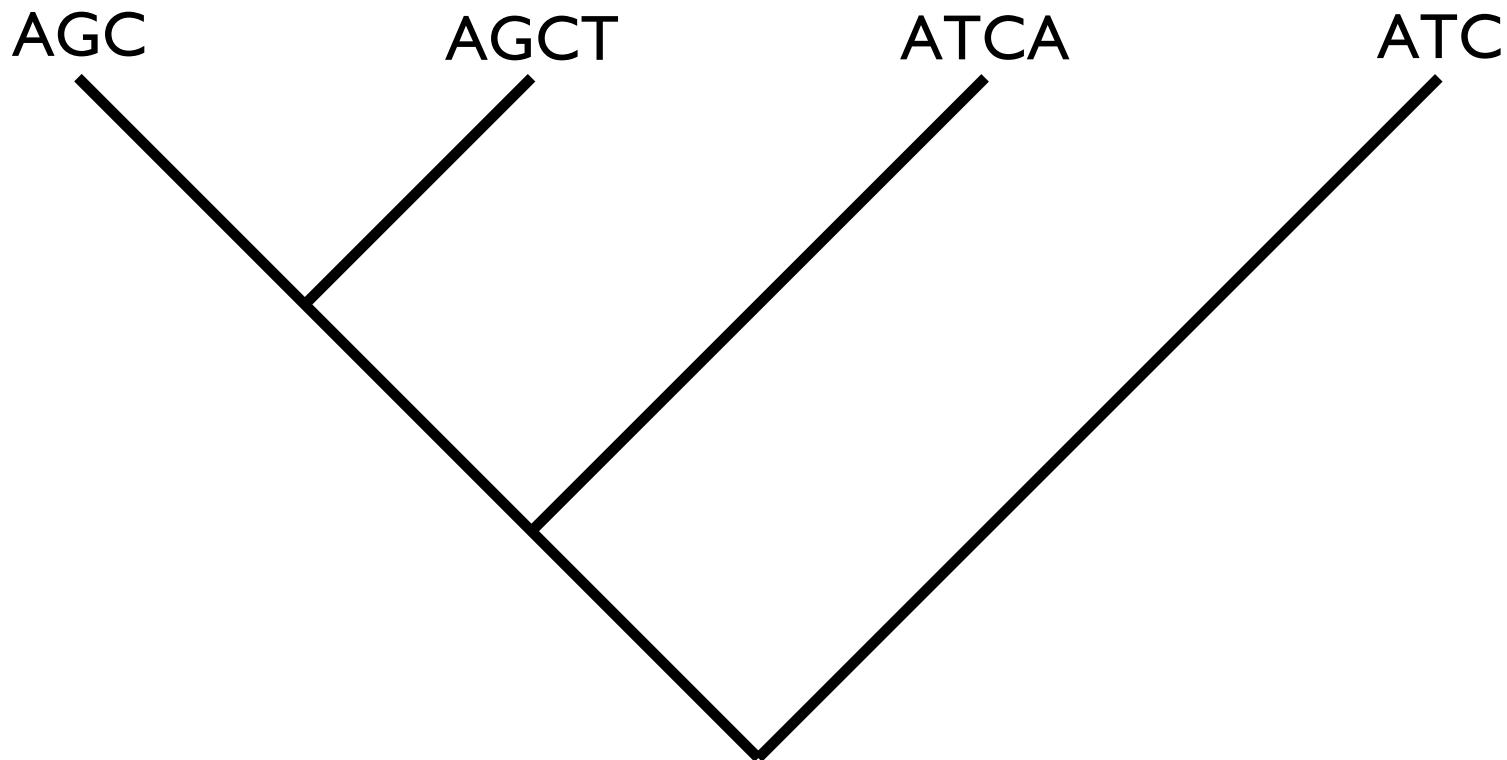
GA	= R
GT	= K
GC	= S
GAT	= D
GAC	= V
GTC	= B
ATC	= H
AT	= W
AC	= M
CT	= Y
ACGT	= N

Direct optimization

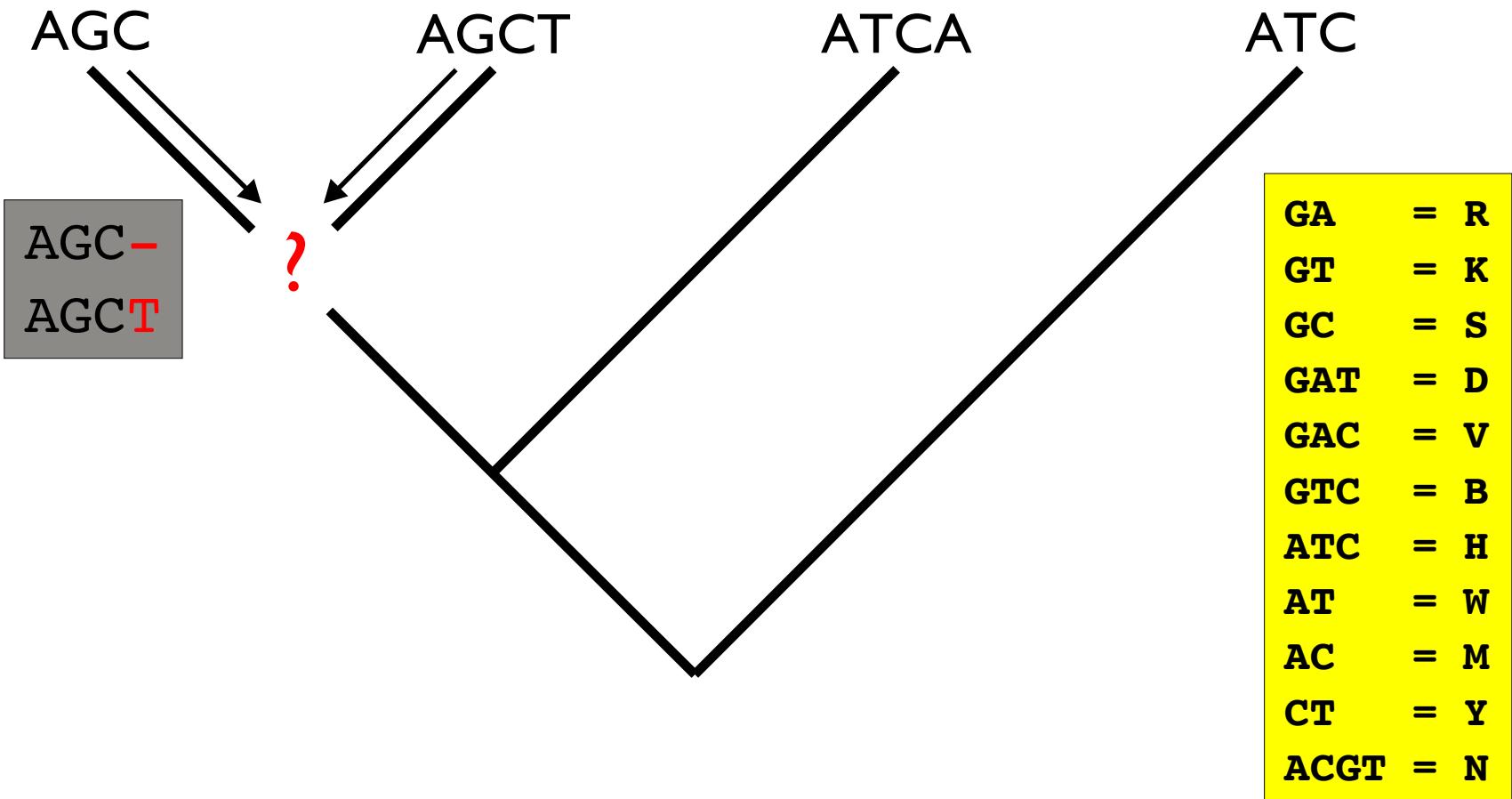


GA	= R
GT	= K
GC	= S
GAT	= D
GAC	= V
GTC	= B
ATC	= H
AT	= W
AC	= M
CT	= Y
ACGT	= N

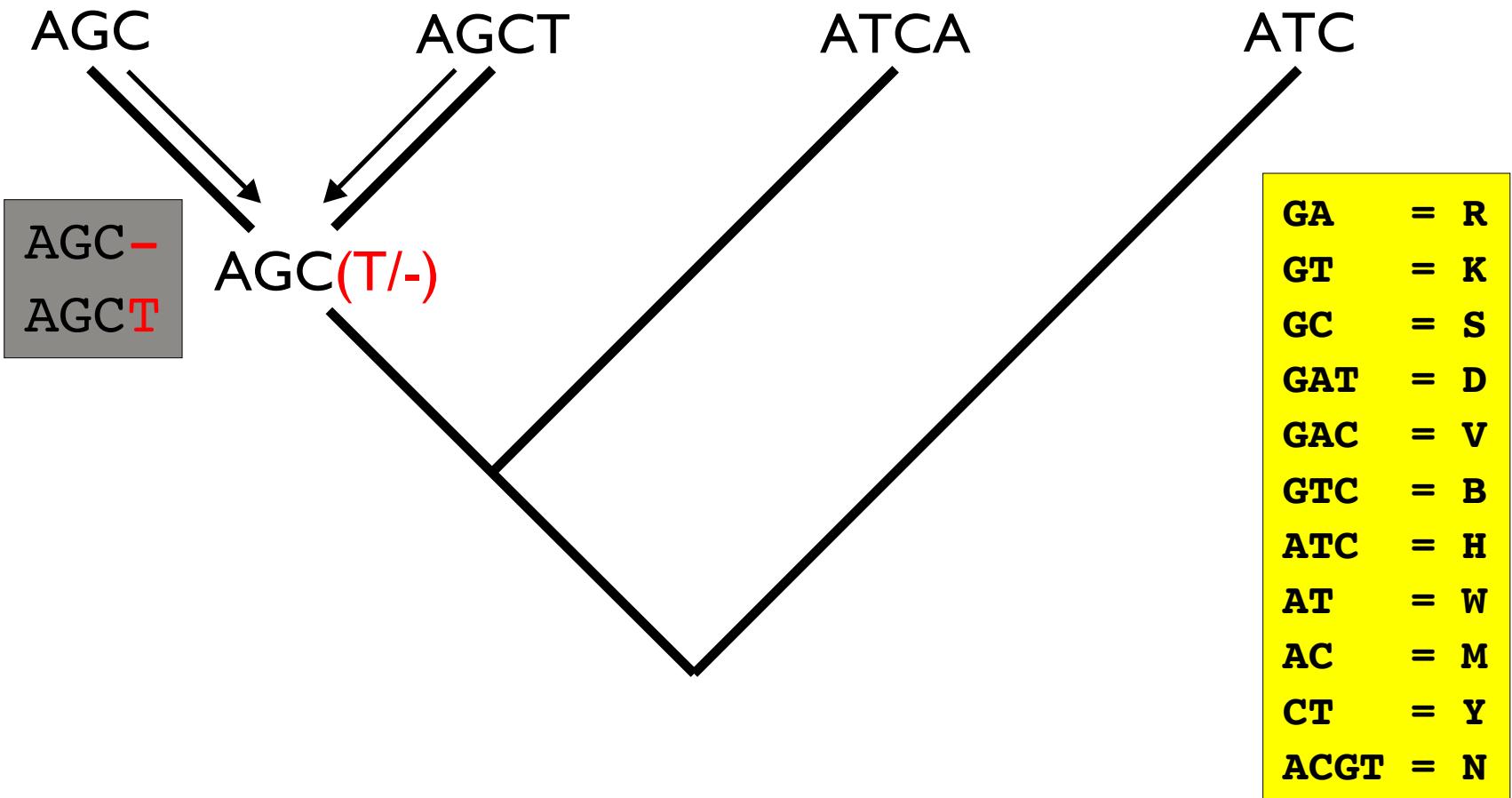
Direct optimization



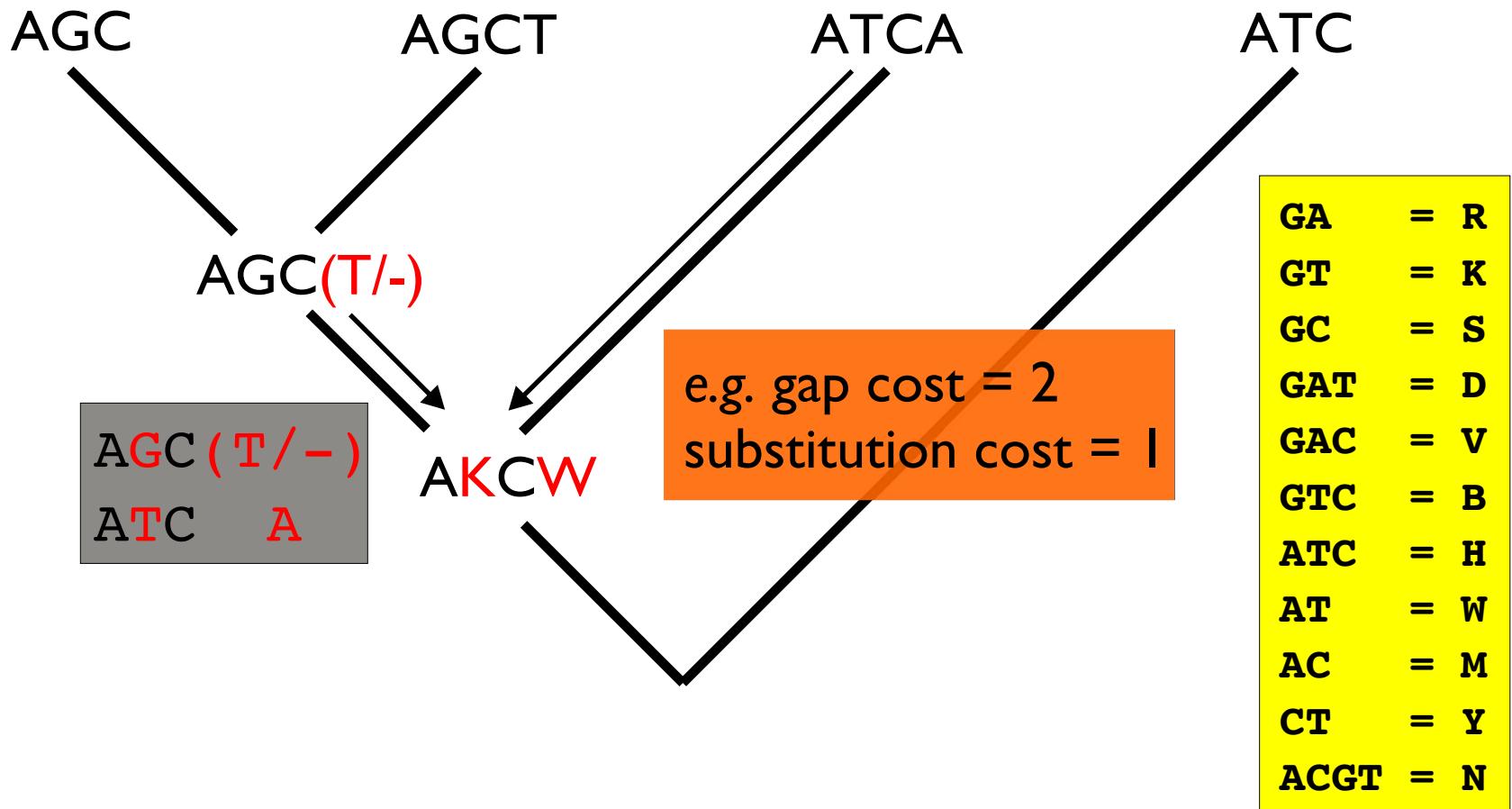
Direct optimization



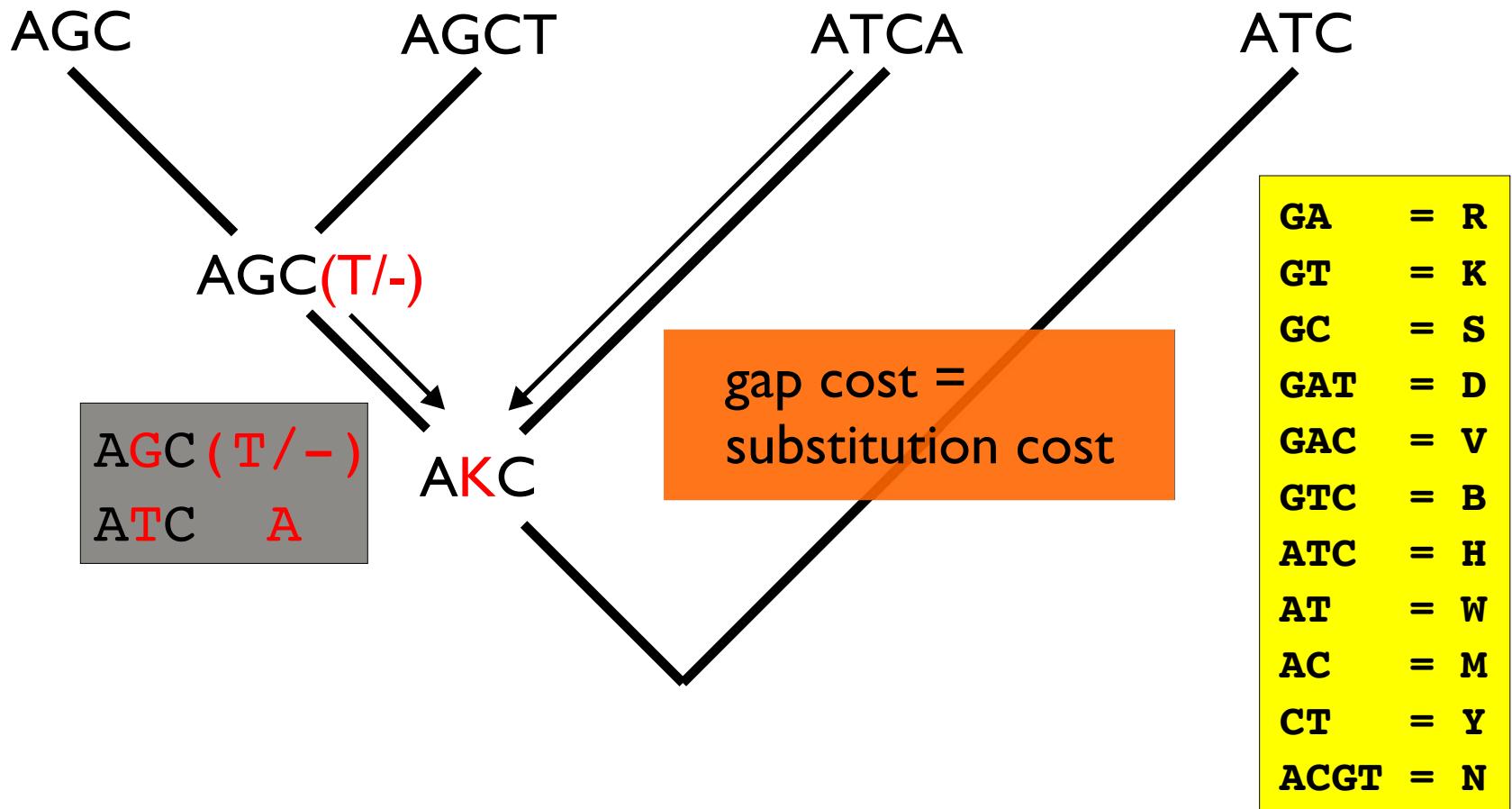
Direct optimization



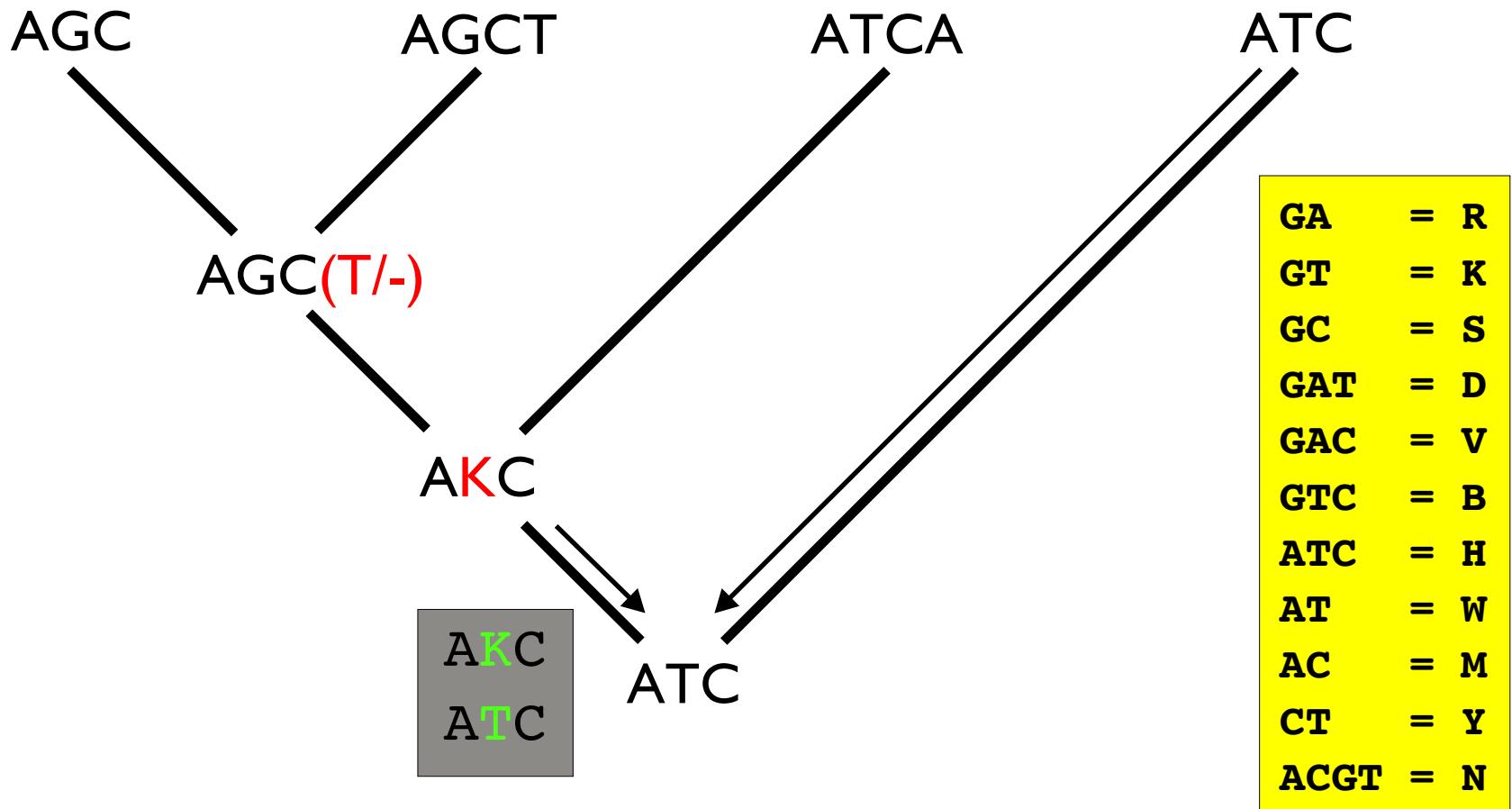
Direct optimization



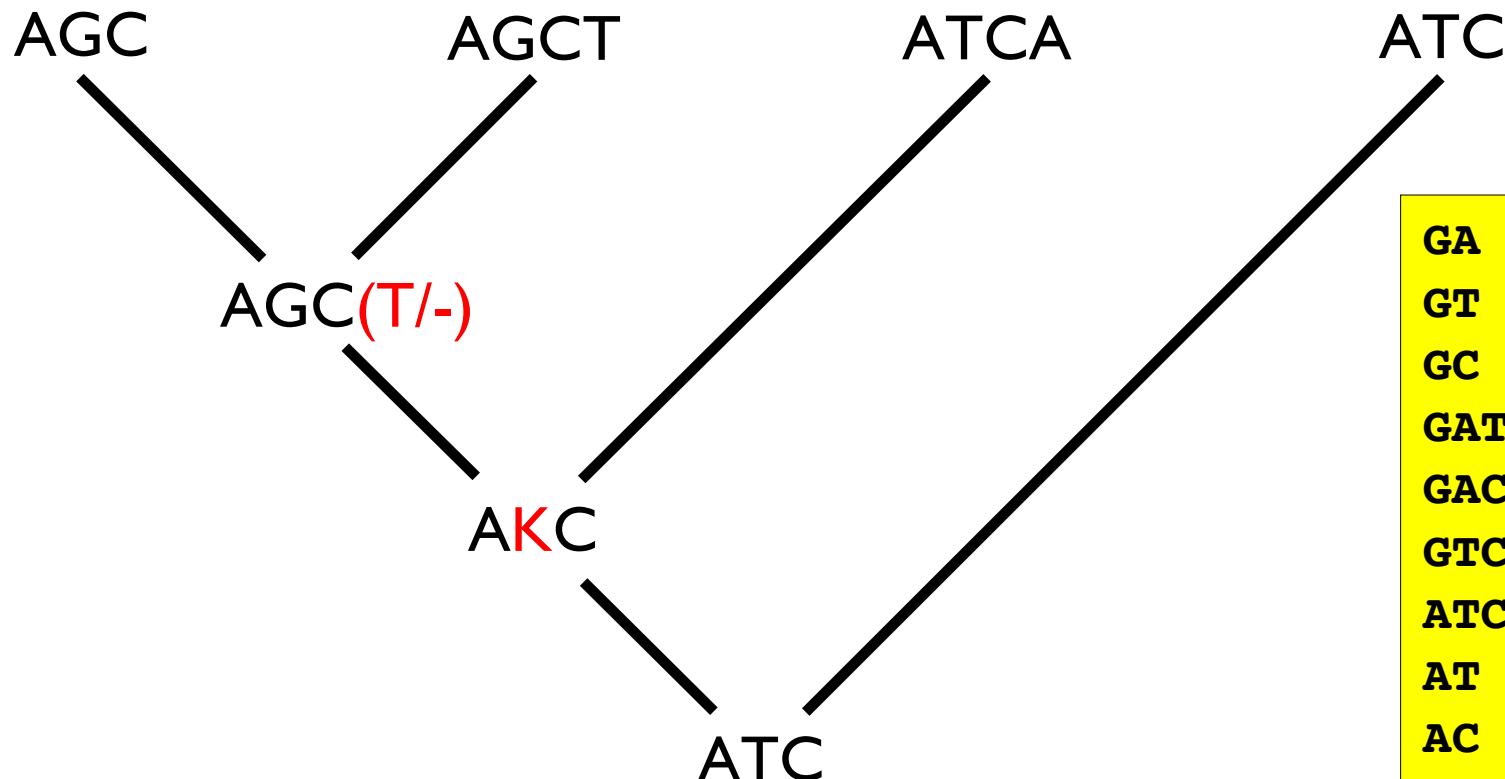
Direct optimization



Direct optimization

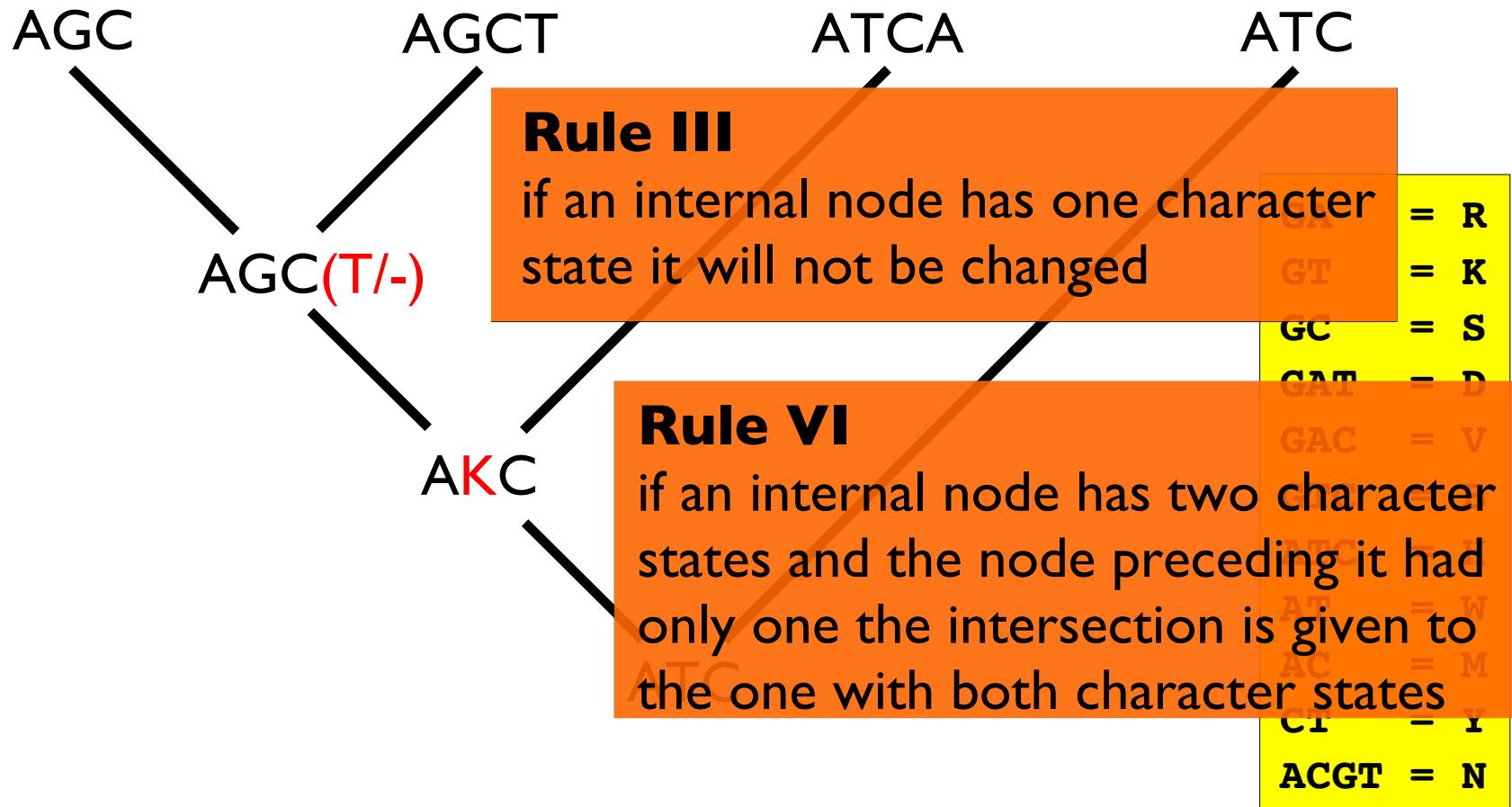


Direct optimization

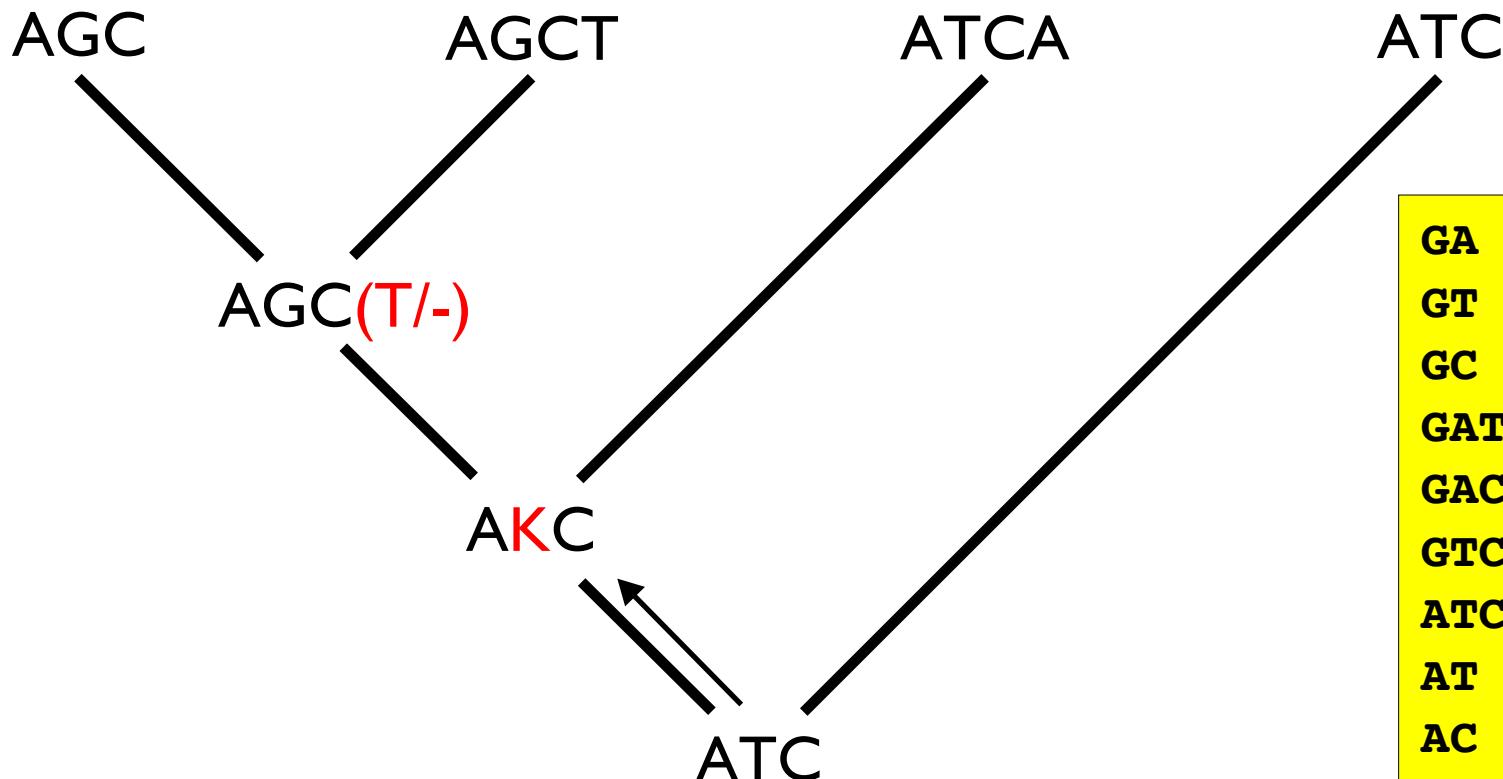


GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization

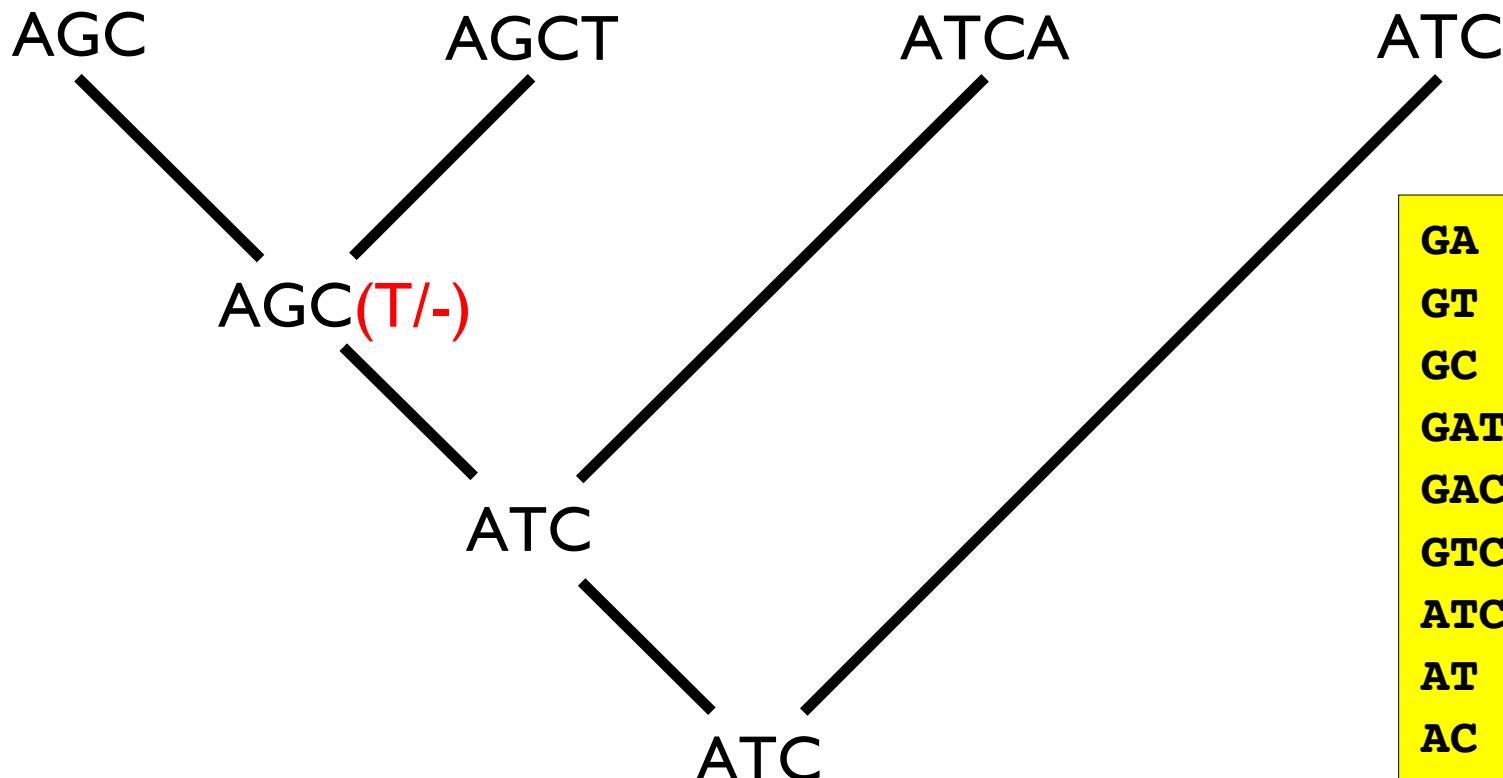


Direct optimization



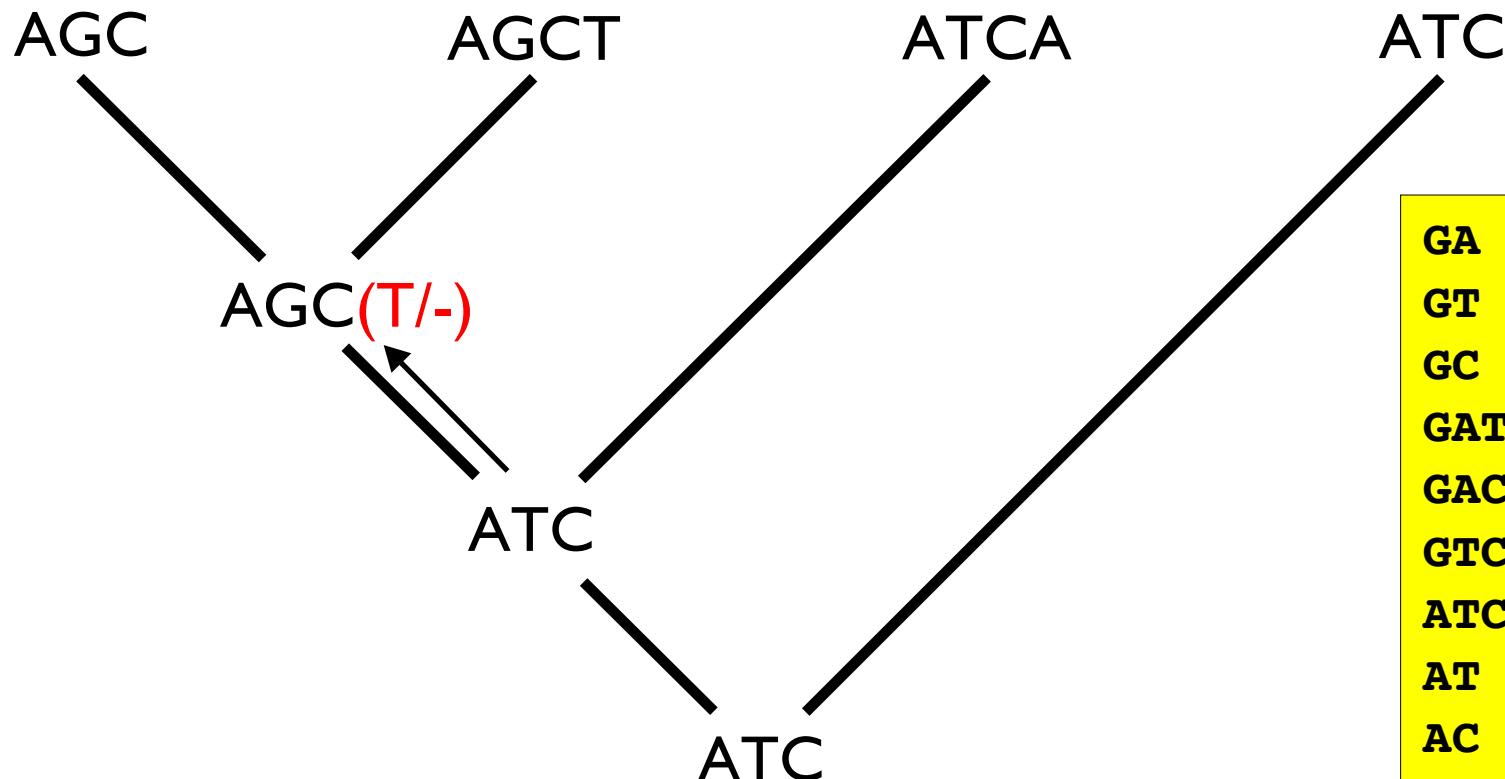
GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization



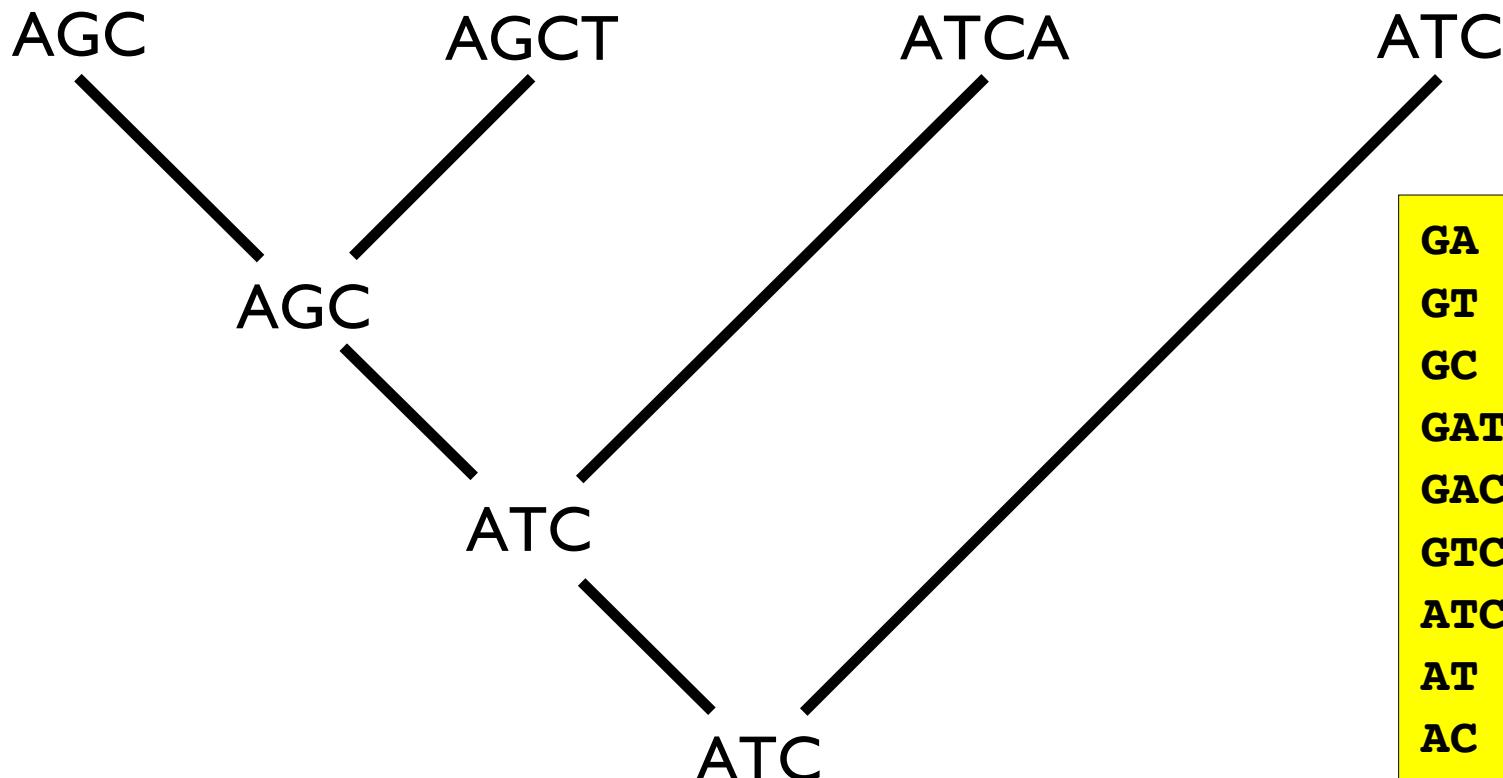
GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

Direct optimization



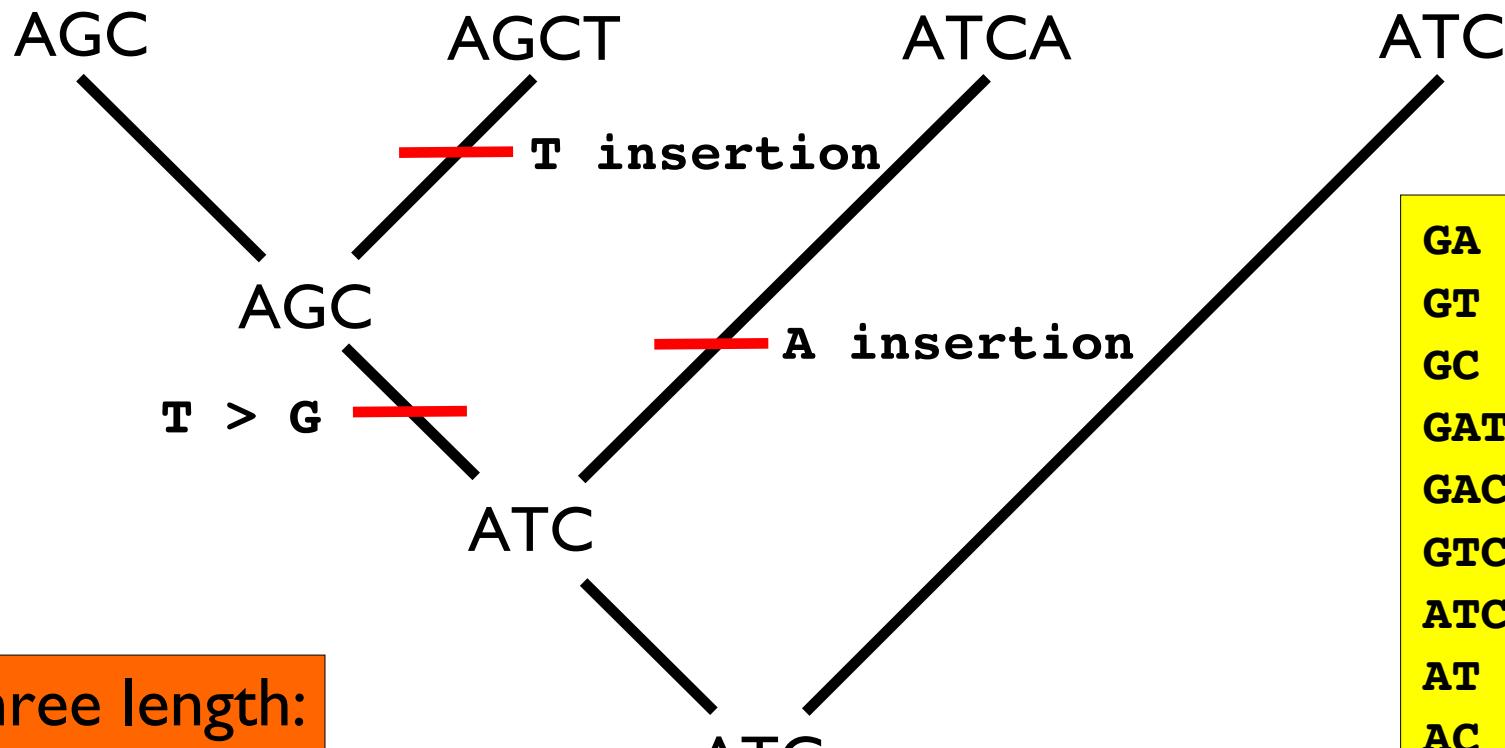
GA	=	R
GT	=	K
GC	=	S
GAT	=	D
GAC	=	V
GTC	=	B
ATC	=	H
AT	=	W
AC	=	M
CT	=	Y
ACGT	=	N

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



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 !
