Introduction to Computational Biology; An Evolutionary Approach: Exact Matching

Bernhard Haubold & Thomas Wiehe

ICB © 2006 Birkhäuser Verlag

Outline

1 Exact Matching in Linear Time

2 Set-Matching with Keyword Trees

3 Suffix Trees

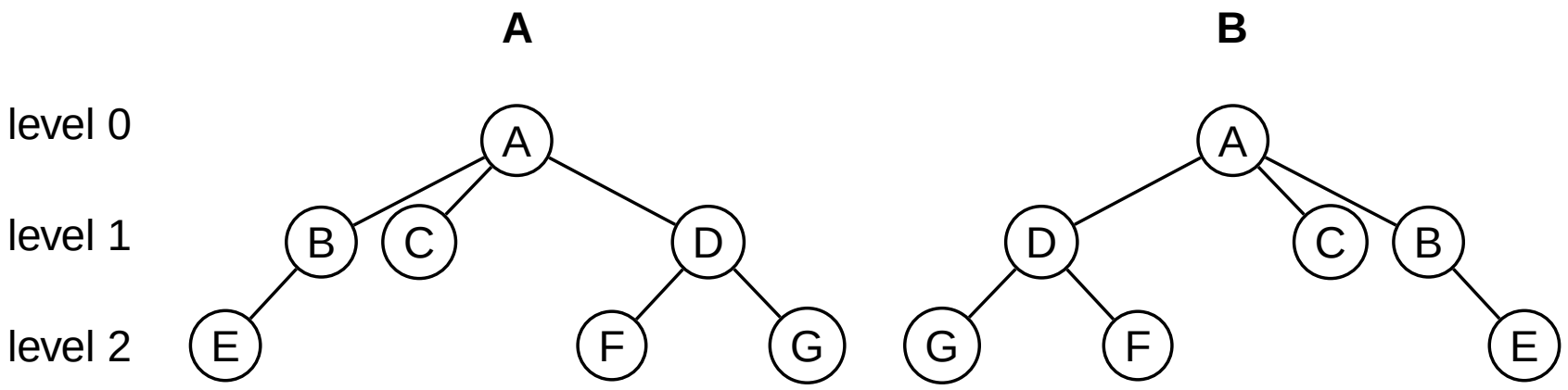
Naive Exact Matching

	Phase 1	Phase 2	Phase 3
Text	CCACAGACACAT	CCACAGACACAT	CCACAGACACAT
Pattern	ACA	ACA	ACA
	^	^	***

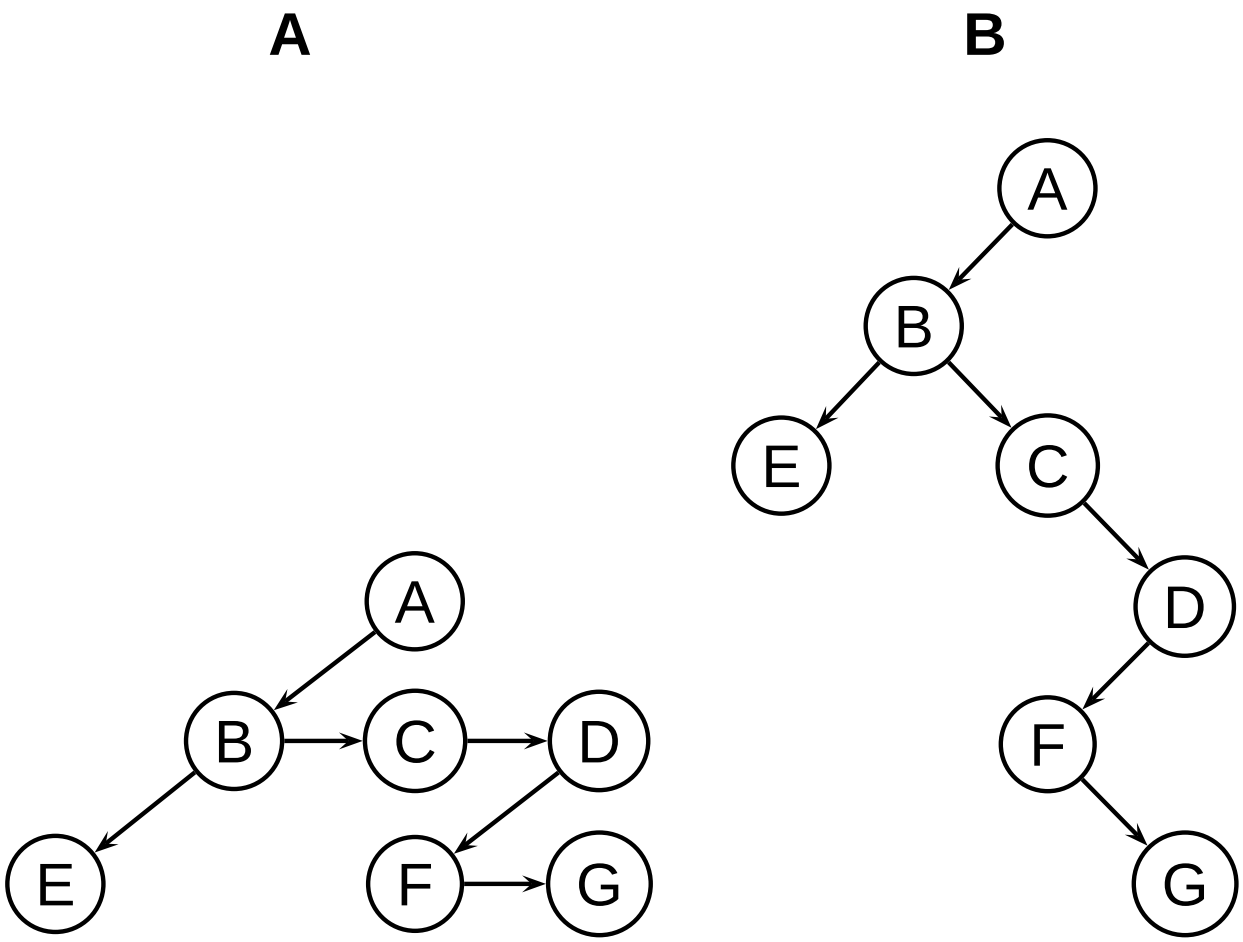
Z-Algorithm

index	123456789
Z_i	901003?
String	ACA$CACAG
	$\sqcup$
	β

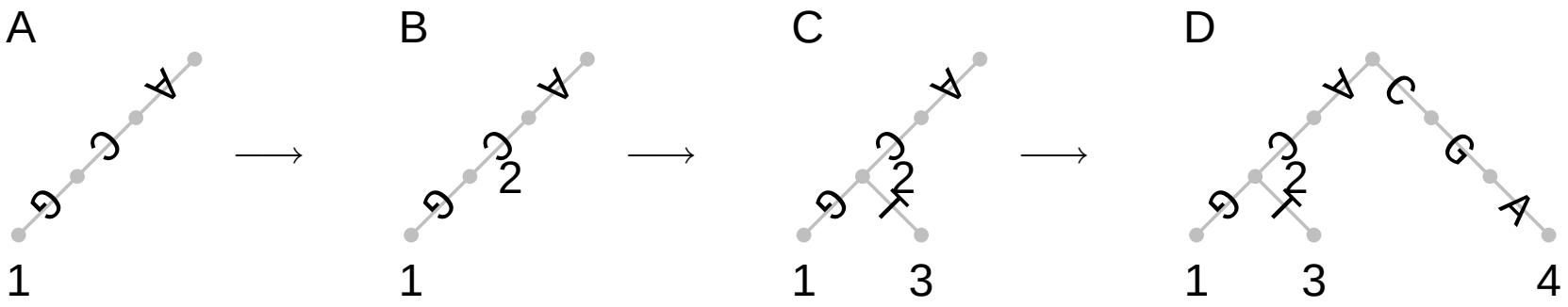
Trees



Binary Trees

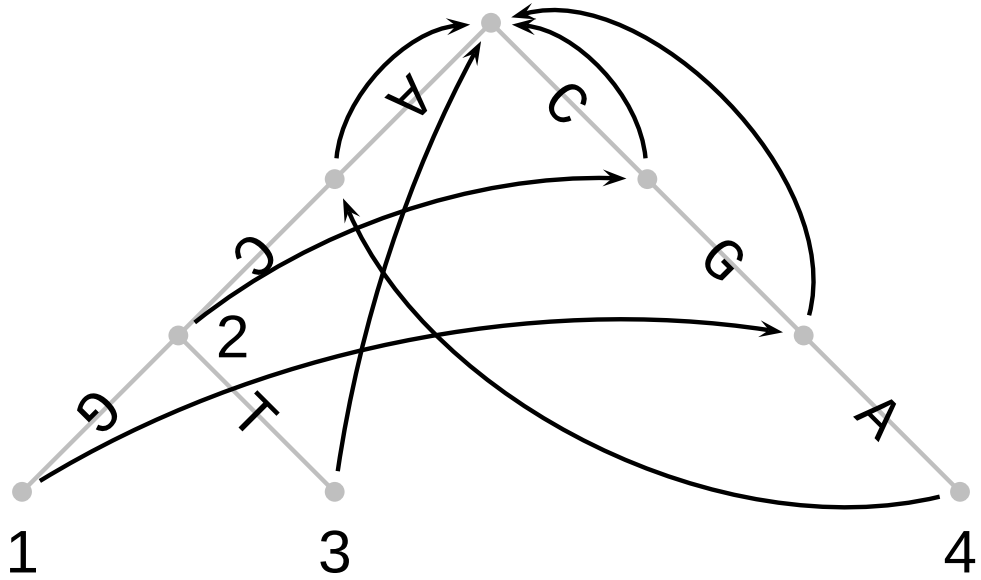


Keyword Tree



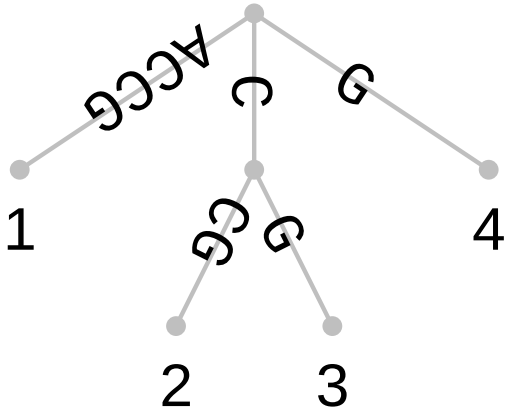
$$P_1 = \mathtt{ACG},\ P_2 = \mathtt{AC},\ P_3 = \mathtt{ACT},\ P_4 = \mathtt{CGA}$$

Failure Links



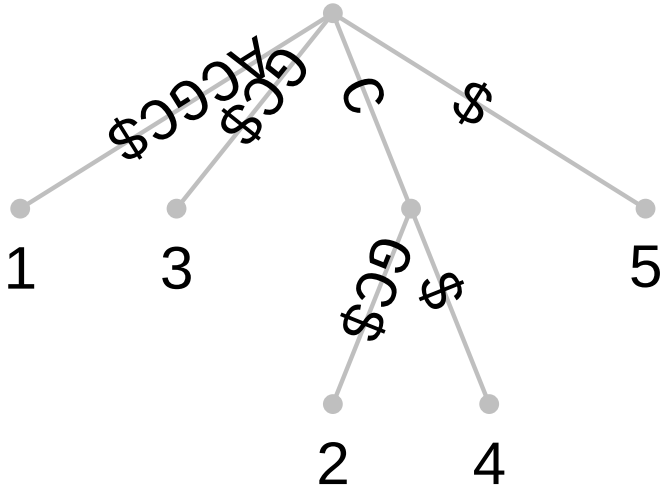
$P_1 = \mathtt{ACG}$, $P_2 = \mathtt{AC}$, $P_3 = \mathtt{ACT}$, $P_4 = \mathtt{CGA}$

Suffix Tree



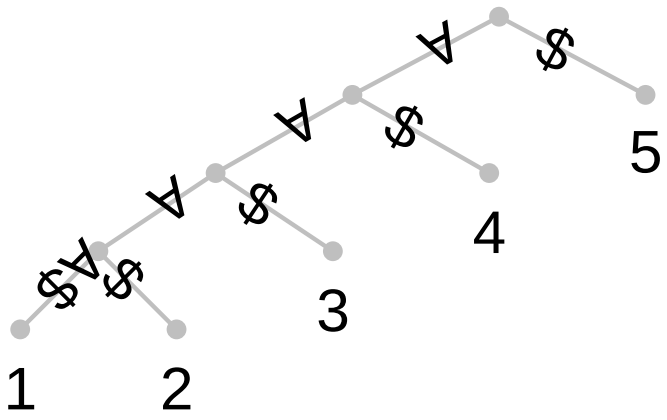
1	2	3	4
A	C	C	G

Sentinel Character



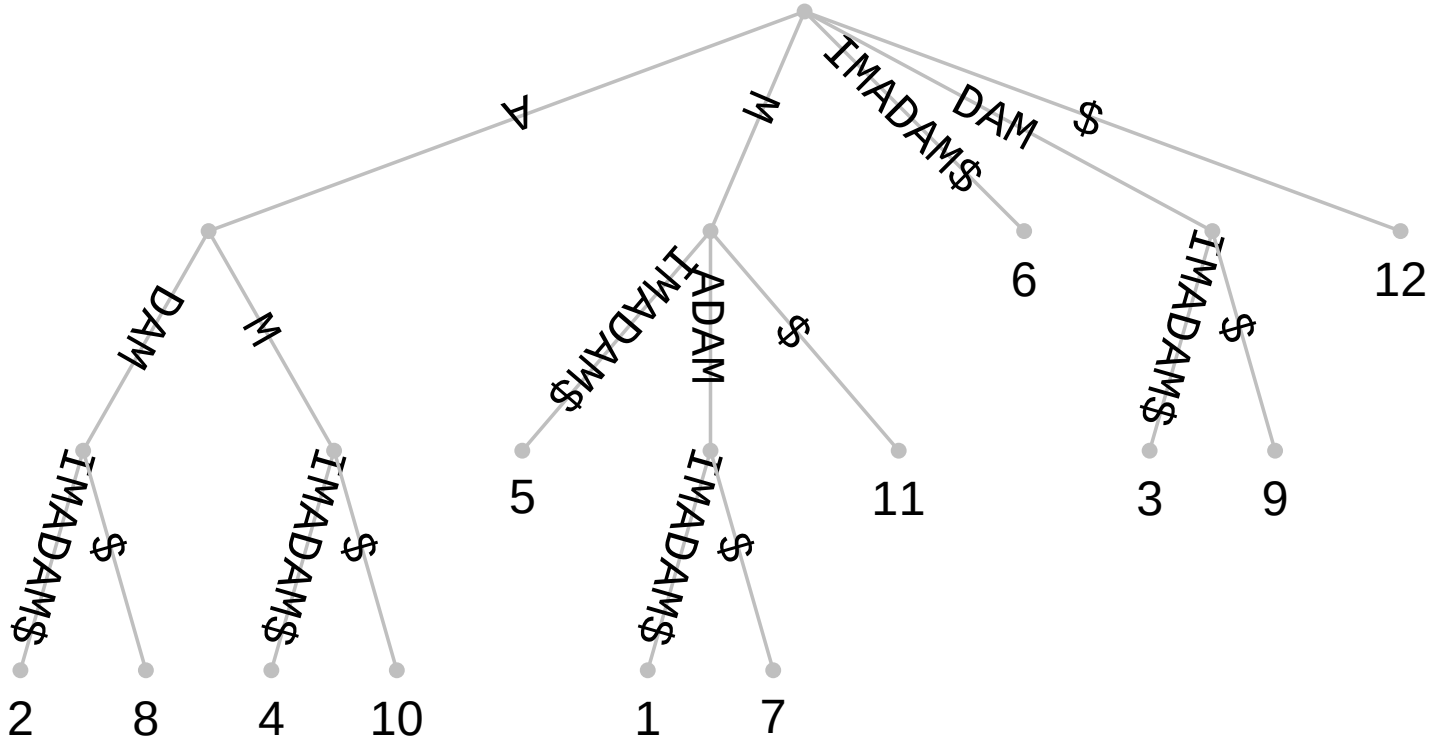
1	2	3	4	5
A	C	G	C	$

Monomorphic Text



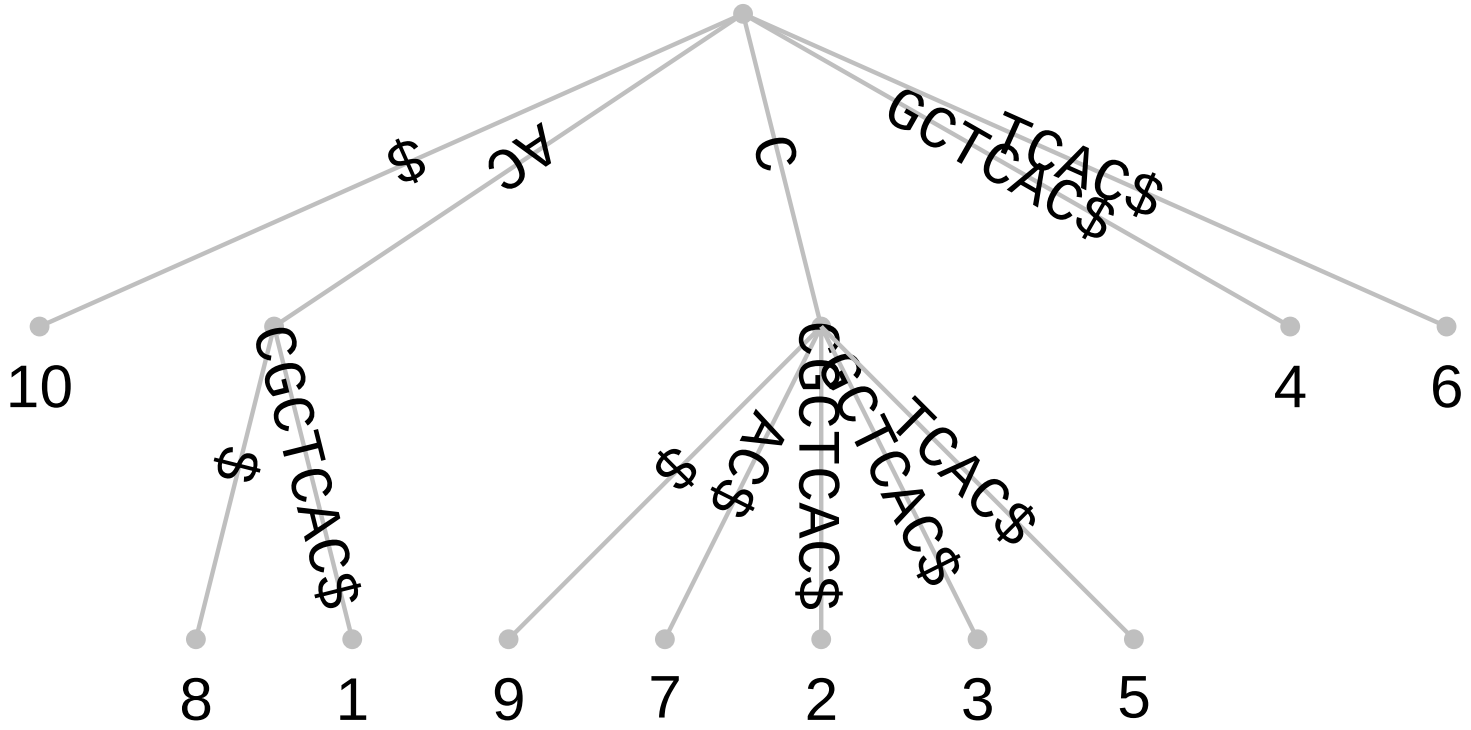
1	2	3	4	5
A	A	A	A	$

English Text



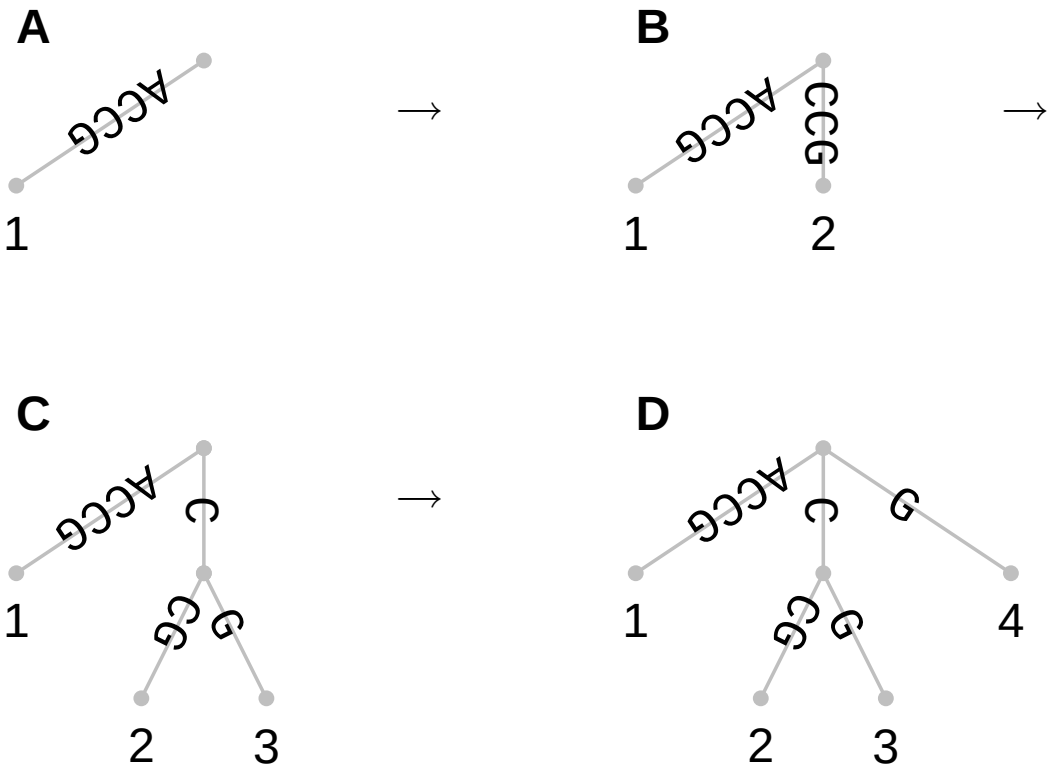
1	2	3	4	5	6	7	8	9	10	11	12
M	A	D	A	M	I	M	A	D	A	M	$

Degree



	1	2	3	4	5	6	7	8	9	10
T	A	C	C	G	C	T	C	A	C	$

Naive Suffix Tree Construction

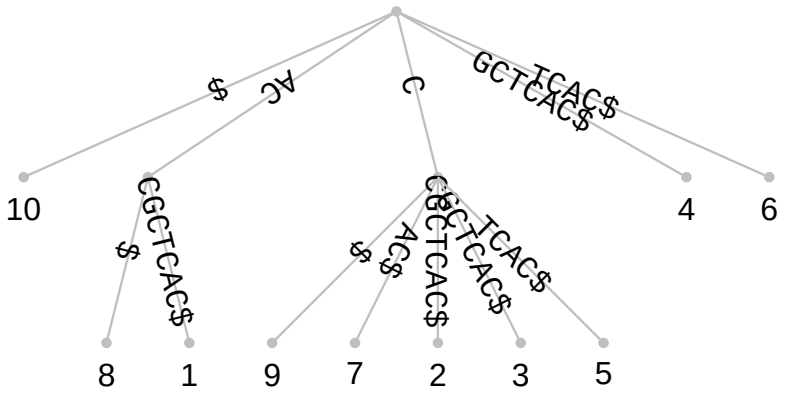


1	2	3	4
A	C	C	G

Suffix Array

	1	2	3	4	5	6	7	8	9	10
T	A	C	C	G	C	T	C	A	C	$

index	suffix start position	suffix
1	8	AC$
2	1	ACCGCTCAC$
3	9	C$
4	7	CAC$
5	2	CCGCTCAC$
6	3	CGCTCAC$
7	5	CTCAC$
8	4	GCTCAC$
9	6	TCAC$



Repetitive Sequences

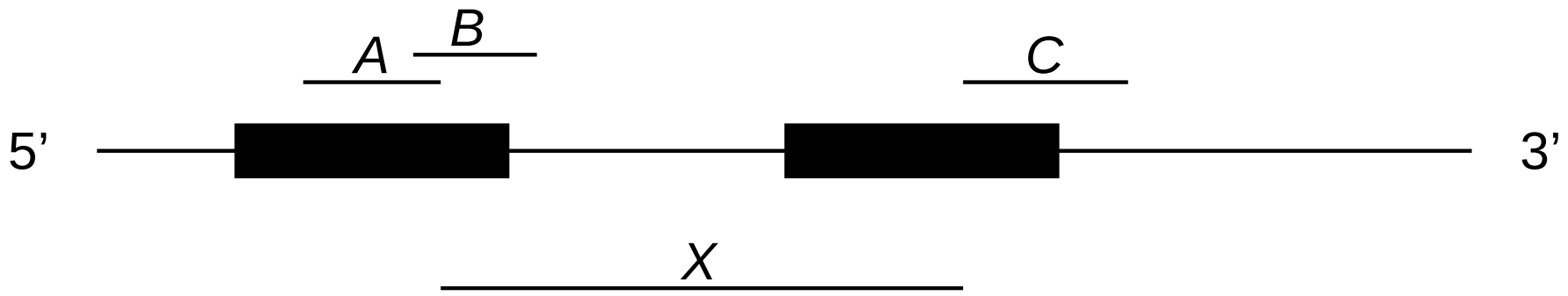
1. interspersed repeats
2. processed pseudogenes
3. microsatellites, e.g. $(\mathtt{A})_n$ or $(\mathtt{GC})_n$
4. segmental duplications
5. blocks of tandemly repeated sequences, e.g. centromers, telomers, and ribosomal gene clusters.

Interspersed Repeats

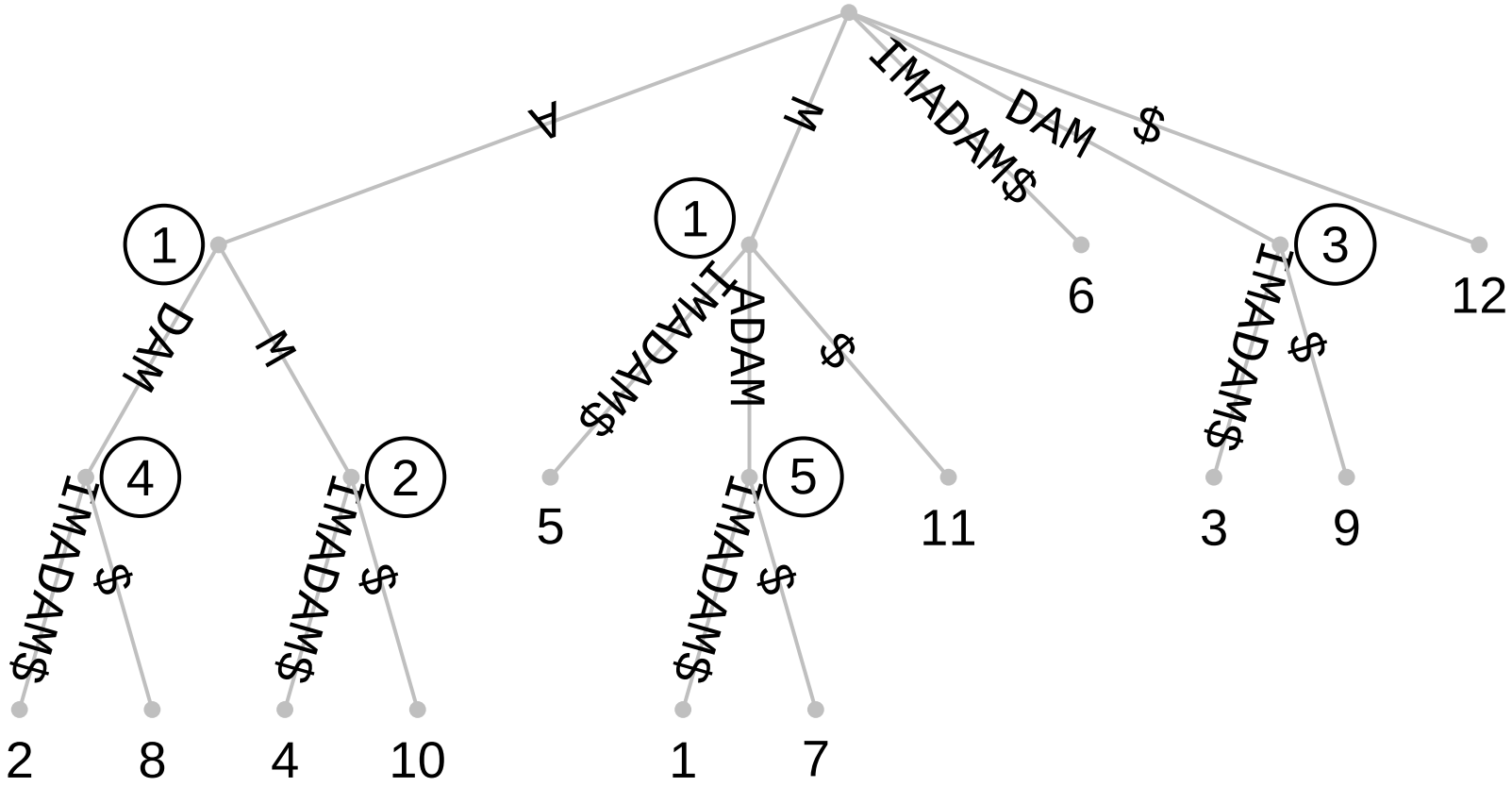
organism	trivial name	interspersed repeats
Homo sapiens	human	44.4%
Mus musculus	mouse	38.6%
Arabidopsis thaliana	mustard weed	10.5%
Caenorhabditis elegans	nematode worm	6.5%
Drosophila melanogaster	fruit fly	3.1%

ICB © 2006 Birkhäuser Verlag

Repeats Confuse Genome Assembly



String Depth

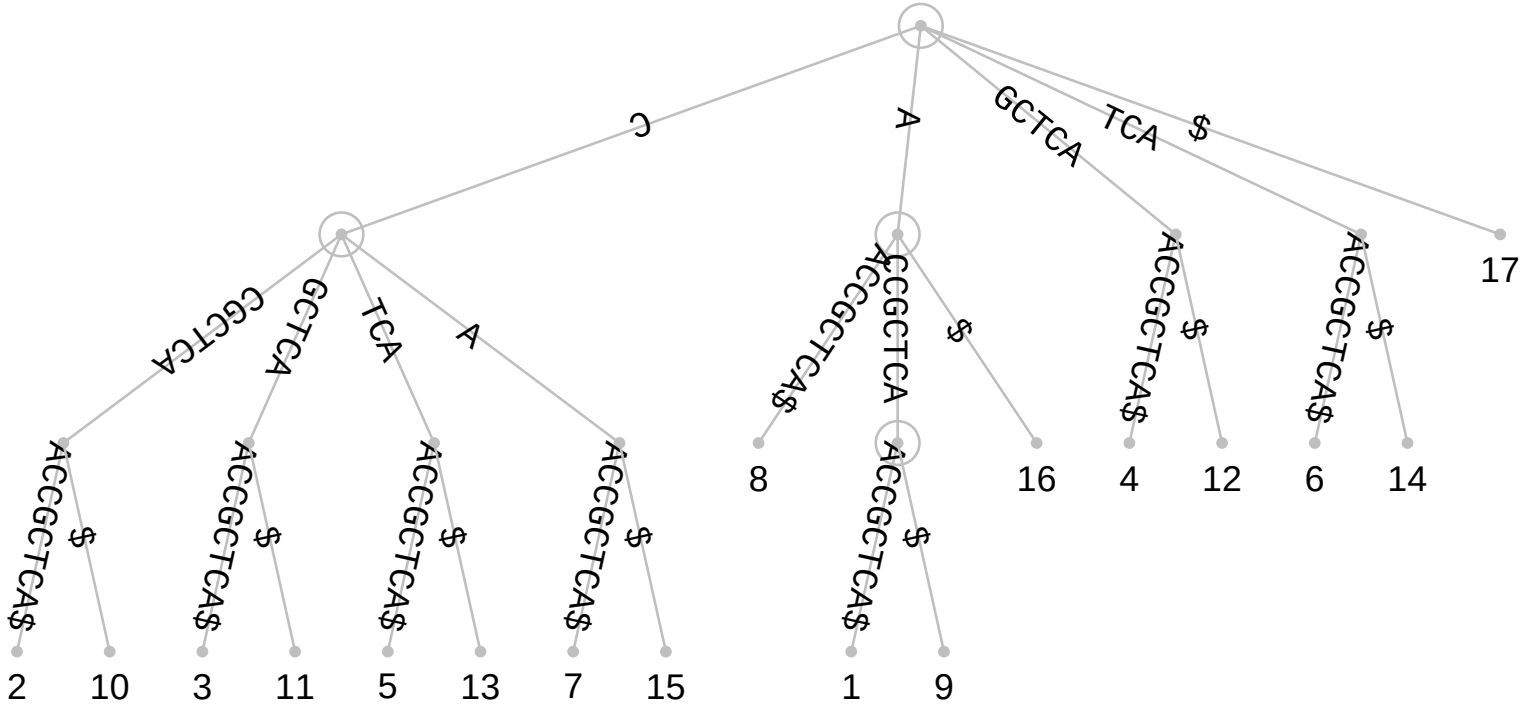


1	2	3	4	5	6	7	8	9	10	11	12
M	A	D	A	M	I	M	A	D	A	M	$

Repeated & Unique Substrings

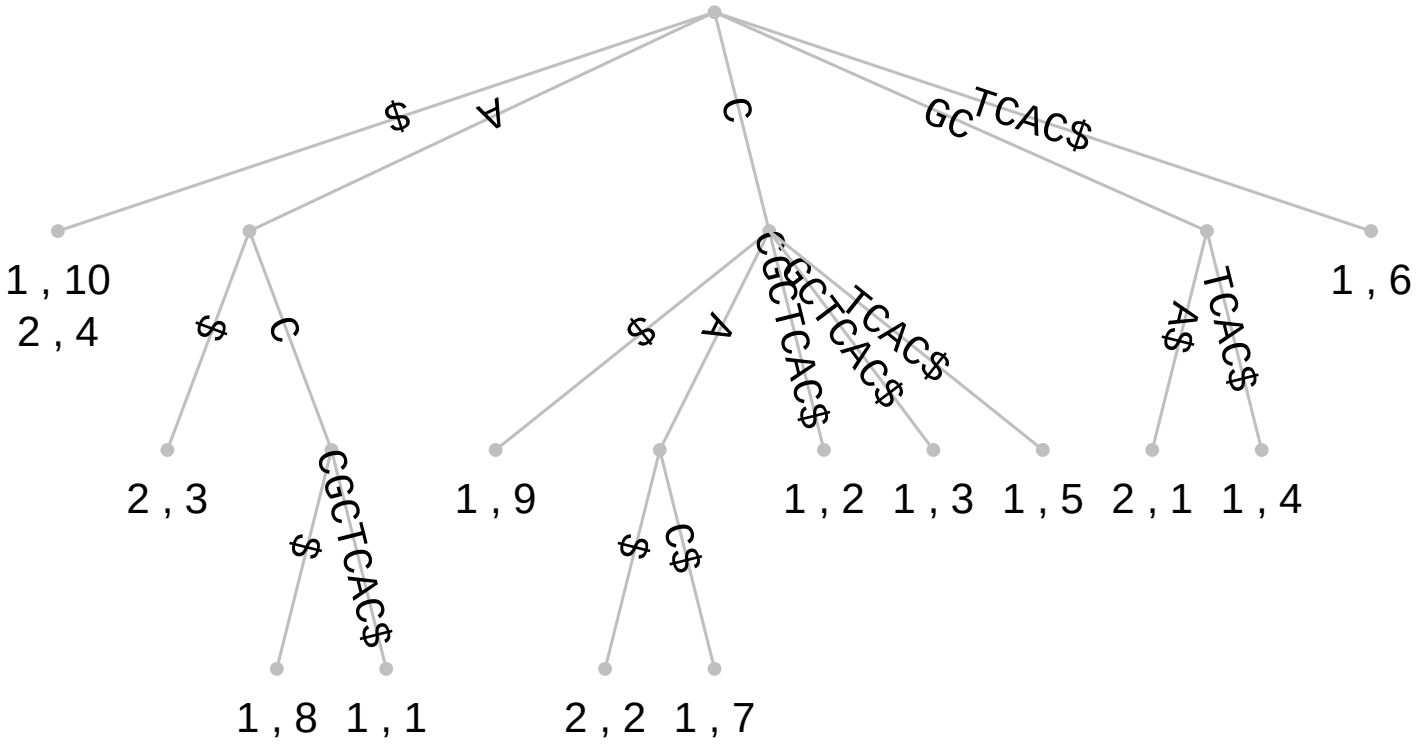
1. Path labels ending above internal node: repeat sequences
2. Path labels ending above leaf: unique sequences
3. Path labels ending just below an internal node connected to leaf: shortest unique substring (shustring) starting at leaf label
4. Path labels ending just above a left-divergent node: maximal repeat

Left-Divergent Nodes



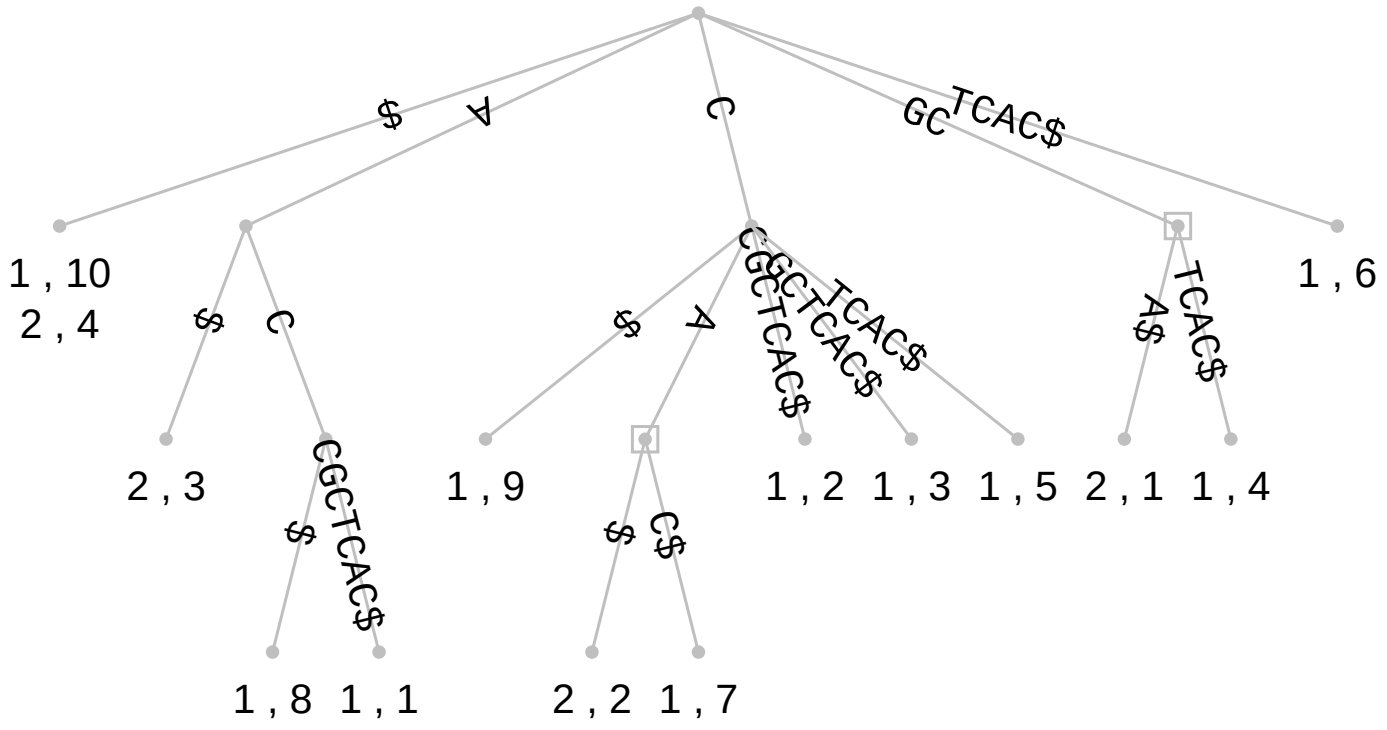
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	C	C	G	C	T	C	A	A	C	C	G	C	T	C	A	$

Generalized Suffix Tree



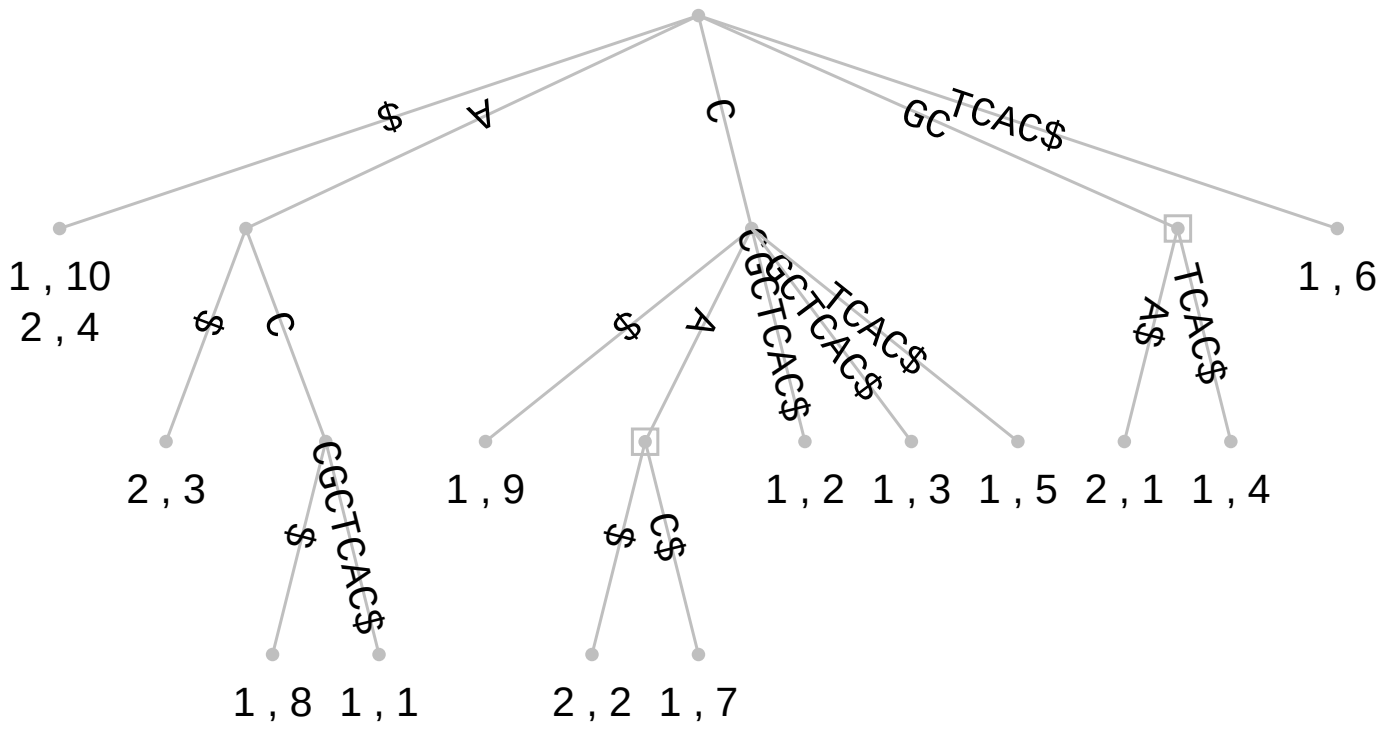
	1	2	3	4	5	6	7	8	9	10
T_1	A	C	C	G	C	T	C	A	C	$
T_2	G	C	A	$						

Longest Common Substring



	1	2	3	4	5	6	7	8	9	10
T_1	A	C	C	G	C	T	C	A	C	$
T_2	G	C	A	$						

k-Mismatches



	1	2	3	4	5	6	7	8	9	10
T_1	A	C	C	G	C	T	C	A	C	$
T_2	G	C	A	$						