Algorithms in Genome Research

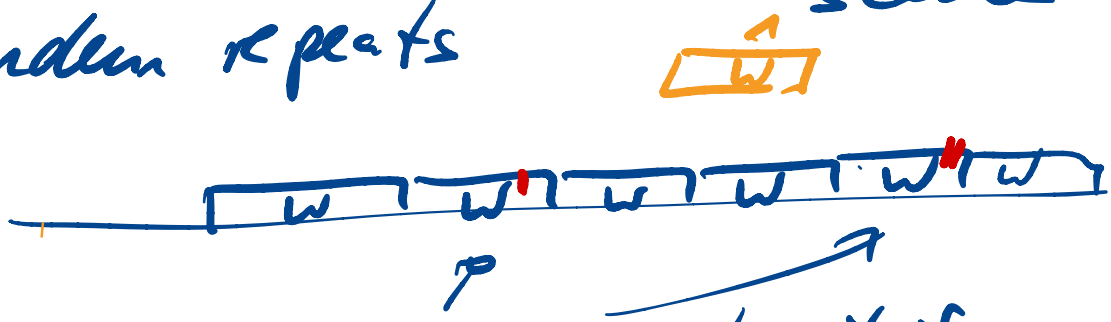
4. Dec. 2020

Patterns in non-coding genomic regions

e.g.
- binding sites ?
- repeat regions !
- small RNAs (miRNAs)
- other motifs that form secondary structures e.g. in mRNA
- horizontally transferred DNA

A (1) Repeat detection on a genomic scale

- tandem repeats



approximate X.V.

a) limited edit distance between neighbors

w_1 w_2 w_3 w_4 ...

$edist(w_i, w_{i+1}) \leq t$

b) global model $\hat{w}$

$edist(w_i, \hat{w}) \leq t$

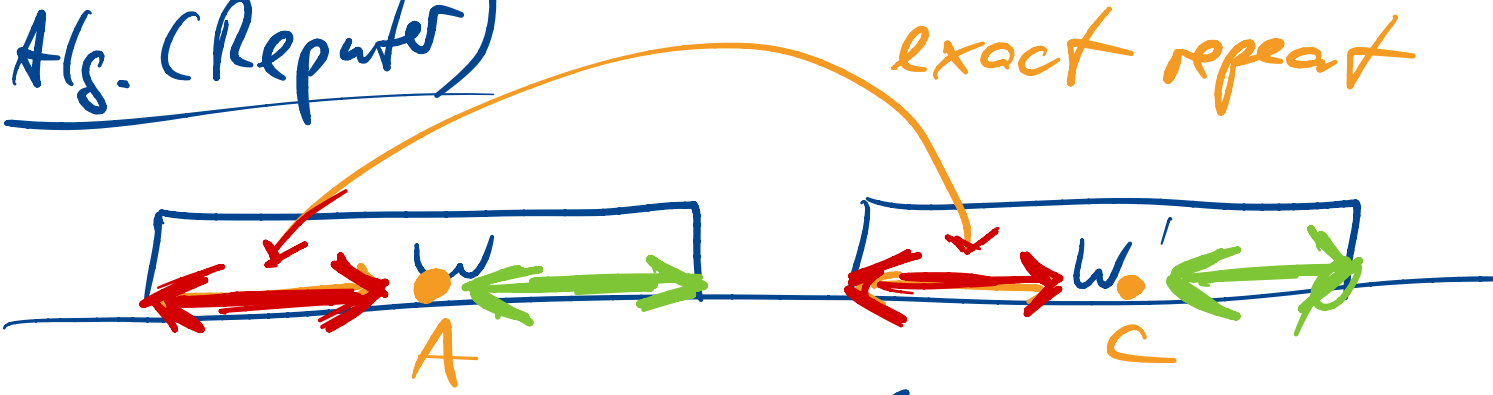
(2) interspersed repeats (SINES, LINES)

parameters: min. length ℓ

max. error e

i.e. $edist(W,W') \le e$

Alg. (Reputer)



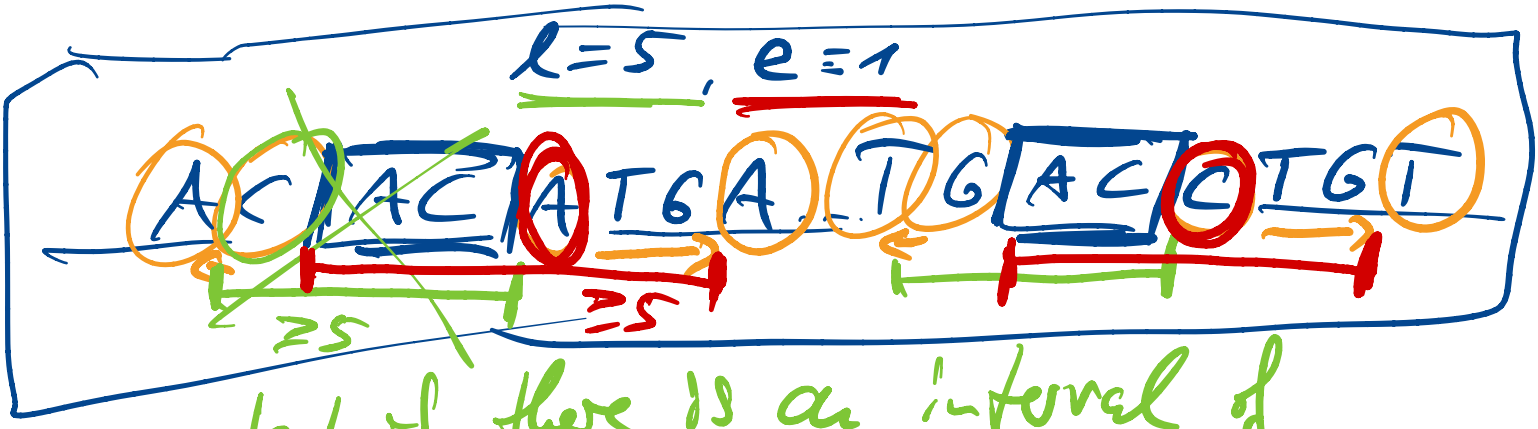
observe: there is always an exact repeat of seed length

$$s \ge \left\lfloor \frac{\ell}{e+1} \right\rfloor .$$

Algorithm:

1. find maximal exact repeats of length $s \ge \left\lfloor \frac{\ell}{e+1} \right\rfloor$.
   → using suffix trees / arrays /

2. extend each seed by up to e errors to the left and to the right. BUT



3. test if there is an interval of length at least ℓ with e errors.

How many seeds z?

$$E[z] = O\left(\frac{n^2}{4^s}\right)$$

Overall runtime: $O(n + ze)$

in practice: can be applied to big genomes with reasonable parameters e.g. $\ell = 100$, $e = 3$

B

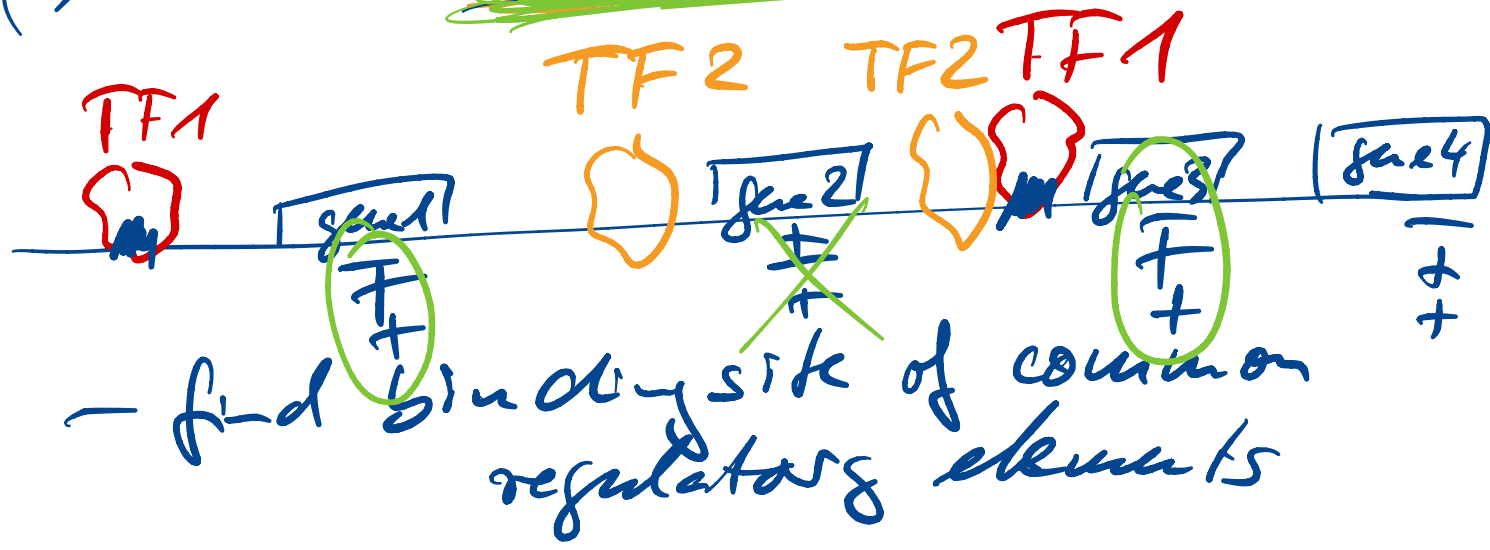
Finding Regulatory Regions

Promotor region

discovery

Here: ~~detection~~ of unknown binding sites

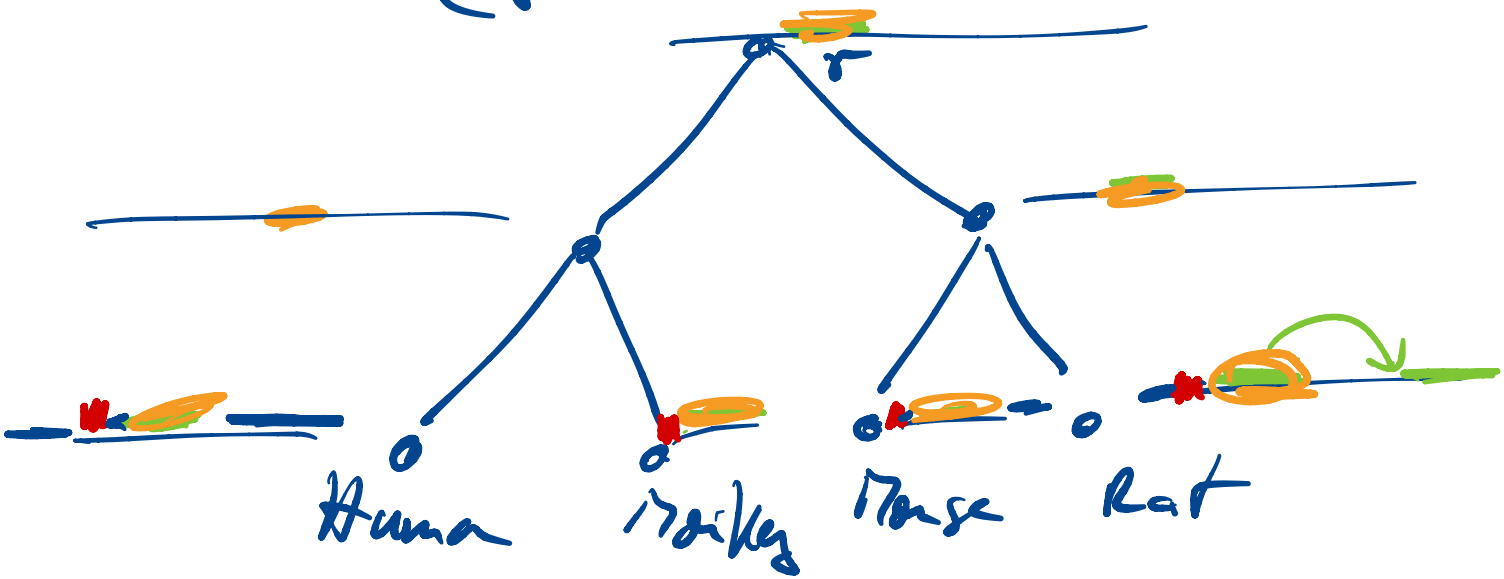
(1) examine co-regulated genes



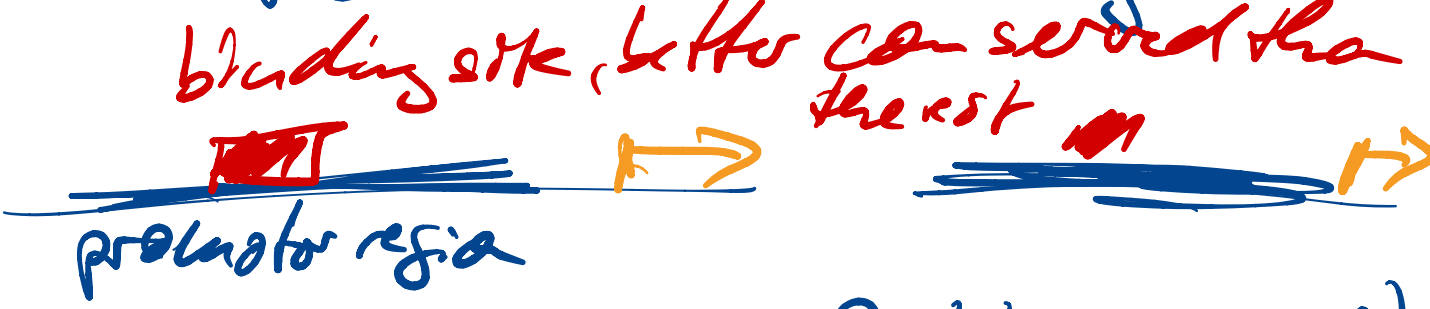
- find binding site of common regulatory elements

(2) examine orthologous genes (of different species)

paralogs



Known: binding sites (& genes)
are better conserved
than the surrounding areas



↳ Substring Parsimony Problem (SPP)

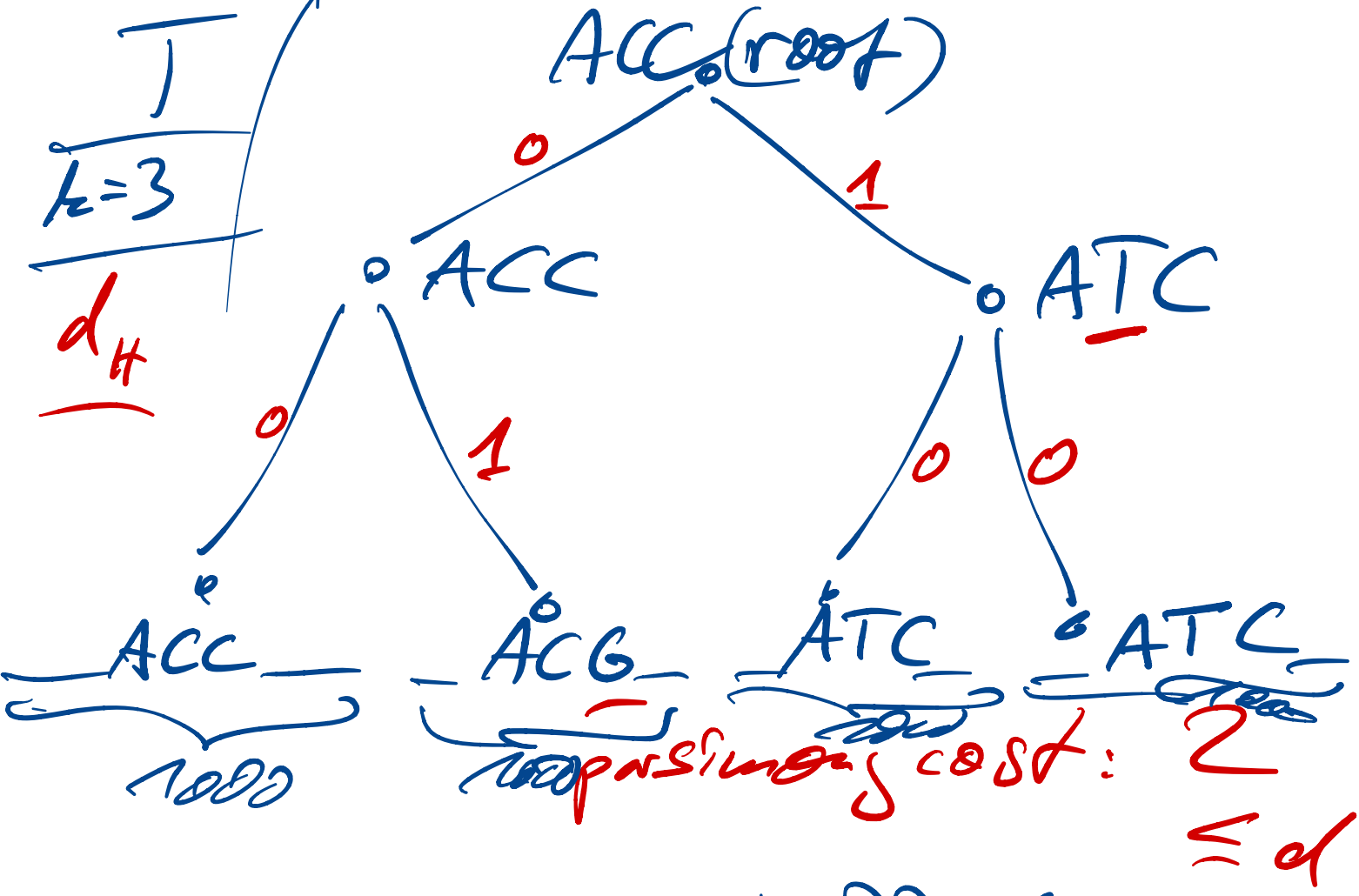
given: $S_1, \dots, S_n$ set of promotor regions of orthologous genes

T phylogenetic tree of the n species with root r

k length of the motif

d maximum # of errors allowed.

wanted: All sets of substrings $s_1, \dots, s_n$ of $S_1, \dots, S_n$, each of length k, such that the parsimony cost of $s_1, \dots, s_n$ along T is at most d.



Solution: A simple DP algorithm

bottom up processing of T:

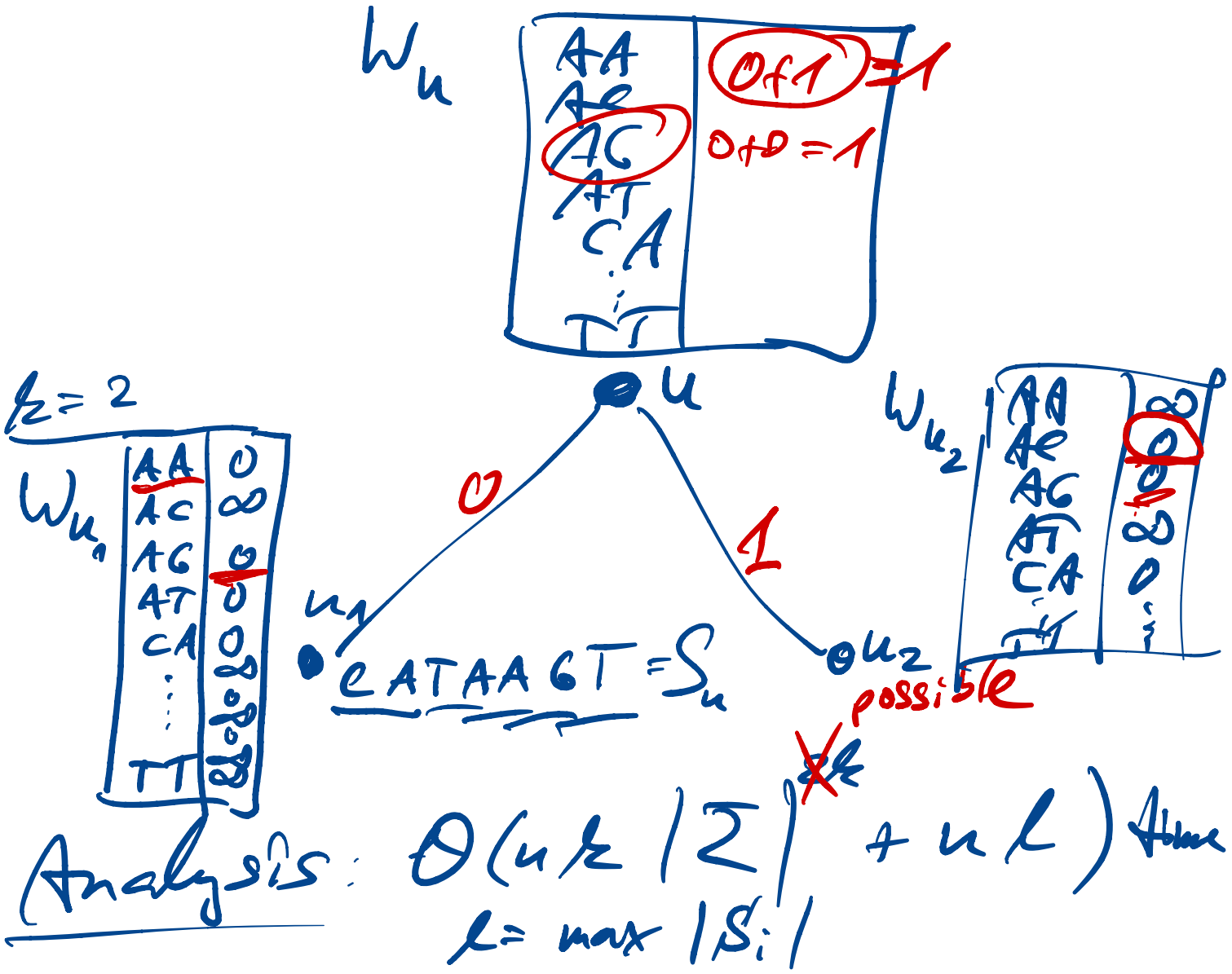
initialisation of leaf u

$W_u[s] = 0$ if s is a substring of S_u

$= \infty$ otherwise

recursive calculation of internal vertex u with children C(u)

$$W_u[s] = \sum_{v \in C(u)} \min_{t \in \Sigma^k} \left(W_v[t] + d(s,t) \right)$$


W_u
AA
AC
AT
CA
TT
0+1=1
0+0=1
u
0
1
u_1
u_2
$k=2$
W_{u_1}
W_{u_2}
CATAAGT = S_u
possible
Analysis: $O(uk|\Sigma| + ul)$ time
$l = \max |S_i|$