A parallel algorithm for the extraction of structured motifs

Alexandra M. Carvalho

ALGOS, INESC-ID

Lisbon, Portugal

joint work with

Ana T. Freitas, Arlindo L. Oliveira

ALGOS, INESC-ID

Lisbon, Portugal

and

Marie-France Sagot

INRIA Rhône-Alpes

Lyon, France

Plan of the talk

- Structured motifs extraction
  - the SMILE algorithm

    [L. Marsan and M.-F. Sagot, *J. Computational Biology*, 2000]

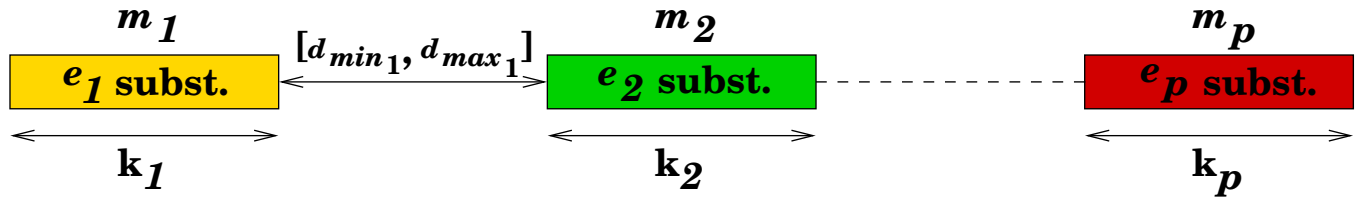
- Parallelization
  - The PARTITION UP TO ε problem
    - the SimpleCut algorithm
  - The tree partition problem
    - the PSMILE algorithm

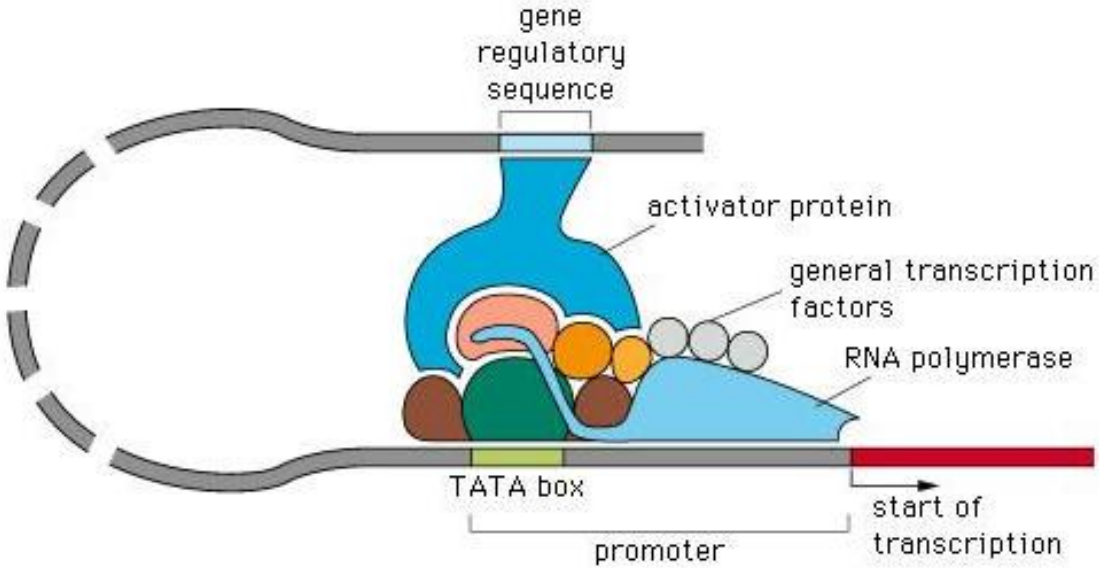
Structured motifs

Definition. ***structured model***

A structured model is a pair (m, d) where:

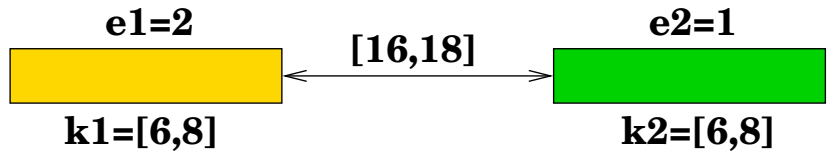
- $m = (m_i)_{1 \leq i \leq p}$, denoting the p boxes, where $m_i \in \Sigma^+$ and $\Sigma = \{A, C, G, T\}$
- $d = (d_{\min_i}, d_{\max_i}, \delta_i)_{1 \leq i \leq p-1}$, denoting the $p - 1$ intervals of distance





Structured motifs

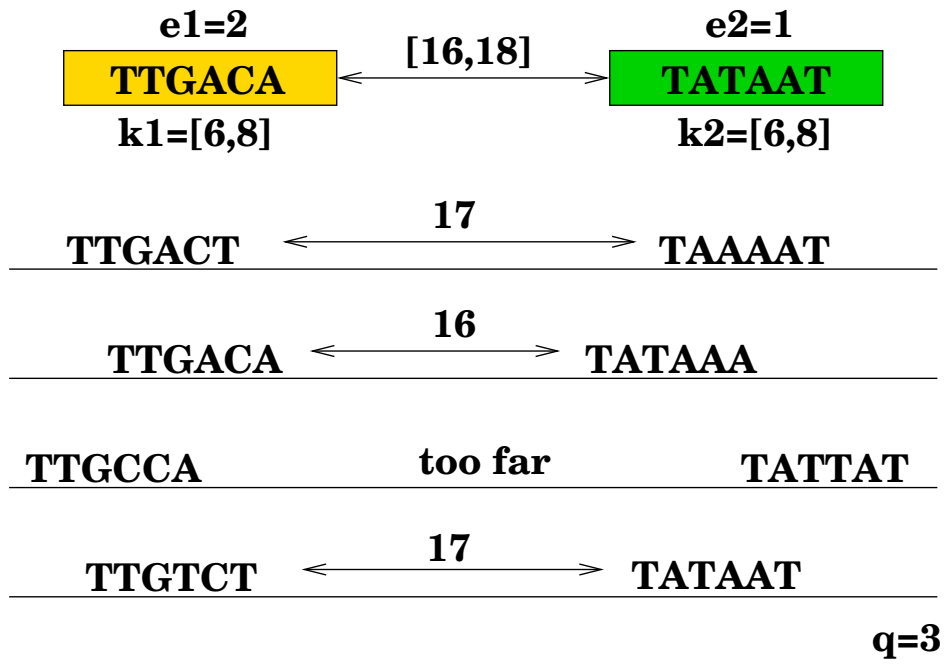
A structured model in a set of input sequences



q=3

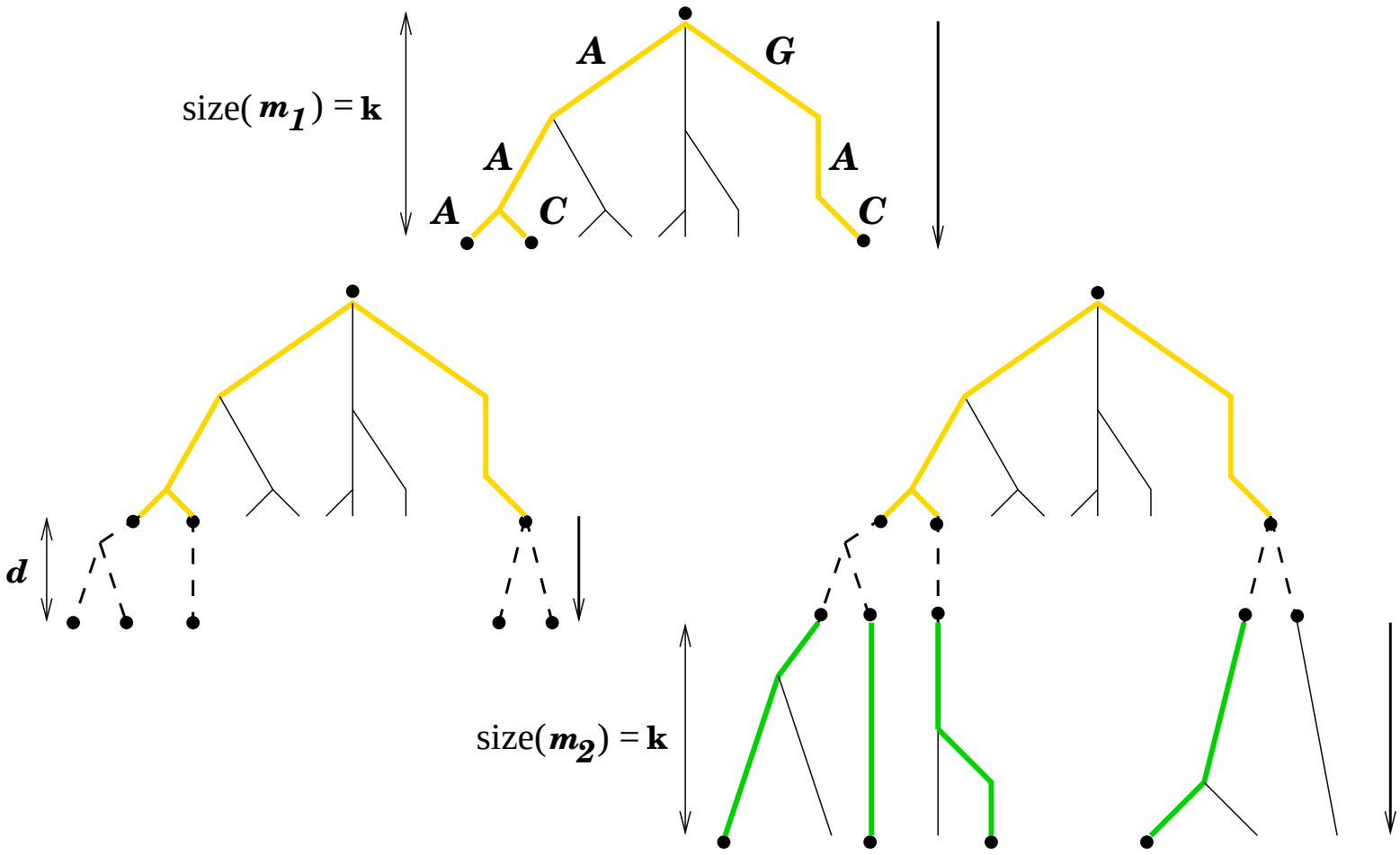
Structured motifs

A structured model in a set of input sequences



Extraction of Structured Models: SMILE

L. Marsan and M.-F. Sagot, *Journal of Computational Biology*, 2000



PARTITION UP TO ε

PARTITION UP TO ε problem:

- ℓ gold bars
- $w_i \geq 0$ is the the weight of the i-th gold bar
- any gold bar can be cut in c equal parts

Optimization version: The problem is how to share the gold between r persons, with the minimum number of gold bars z, in such a way that each person gets the same share of gold up to some weight $\varepsilon > 0$.

Decision version: The problem is to decide whether it is possible to share the gold between r persons, with z gold bars, in such a way that each person gets the same share of gold up to some weight $\varepsilon \geq 0$.

Proposition. The PARTITION UP TO ε problem is NP-complete in the strong sense.

PARTITION UP TO ε

SimpleCut (**Partition** i, **GoldBars** ℓ, **Persons** r, **Weights** w_j, **CutFactor** c, **WorkOverload** ε)

// compute the number of cuts to apply in all gold bars in order to have final gold bars with weight up to ε

1. find the smallest t such that $\frac{max\ \ w_j}{c^t} \leq \varepsilon$

// define ℓ *sets of virtual golds bars, where all virtual gold bars have the same weight*

2. for each $j \in \{1, ..., \ell\}$
3. let $V_j = \left[\sum_{k=1}^{j-1} w_k \times c^t, \sum_{k=1}^{j} w_k \times c^t\right)$

// divide all virtual gold bars in r *different intervals*

4. let $w = \sum_{j=1}^{\ell} w_j$, let $\gamma = w \times c^t \mod r$, let $\delta = \lfloor \frac{w \times c^t}{r} \rfloor$
5. let $I'_i = \begin{cases} [(i-1)(\delta+1),\ i(\delta+1)\ ) & \text{for all } i \leq \gamma \\ [\gamma(\delta+1) + (i-(\gamma+1))\delta,\ \gamma(\delta+1) + (i-\gamma)\delta\ ) & \text{otherwise} \end{cases}$

// transform the r *intervals of virtual gold bars in* r *intervals of real gold bars*

6. transform $I'_i = [a, b)$ into $I_i = [f(a), f(b))$ with $f : w \times c^t \rightarrow \ell \times c^t$:

$$f(x) = \begin{cases} (j-1) \times c^t + \frac{x - \inf(V_j)}{w_j} & \text{for all } x \in V_j \\ \ell \times c^t & \text{if } x = w \times c^t \end{cases}$$

Parallelization

Reducing the tree partition problem to the PARTITION UP TO ε problem

Input of the SimpleCut algorithm for the i-th processor:

- the ℓ gold bars matches the symbols of the alphabet Σ
  $\ell = \{A, C, G, T\}$ **for DNA sequences**
- the r persons matches the number of available processing units
- the weight w_j of each alphabet symbol is obtained by scanning the input sequences
- the c cut factor matches the size of the alphabet Σ
  $c = 4$ **for DNA sequences (for instance,** A **can be cut in** AA**,** AC**,** AG **and** AT**)**
- ε is an user parameter

Output of the SimpleCut algorithm for the i-th processor:

- the number t of cuts gives the depth $t + 1$ of the tree where the partition is defined
- an interval I_i corresponding to tree nodes at depth $t + 1$ assigned to the i-th processor

Parallelization

j	1	2	3	4
σ_j	A	C	T	G
w_j	2	1	1	2

$r = 5 \;\; \varepsilon = 1$

Parallelization

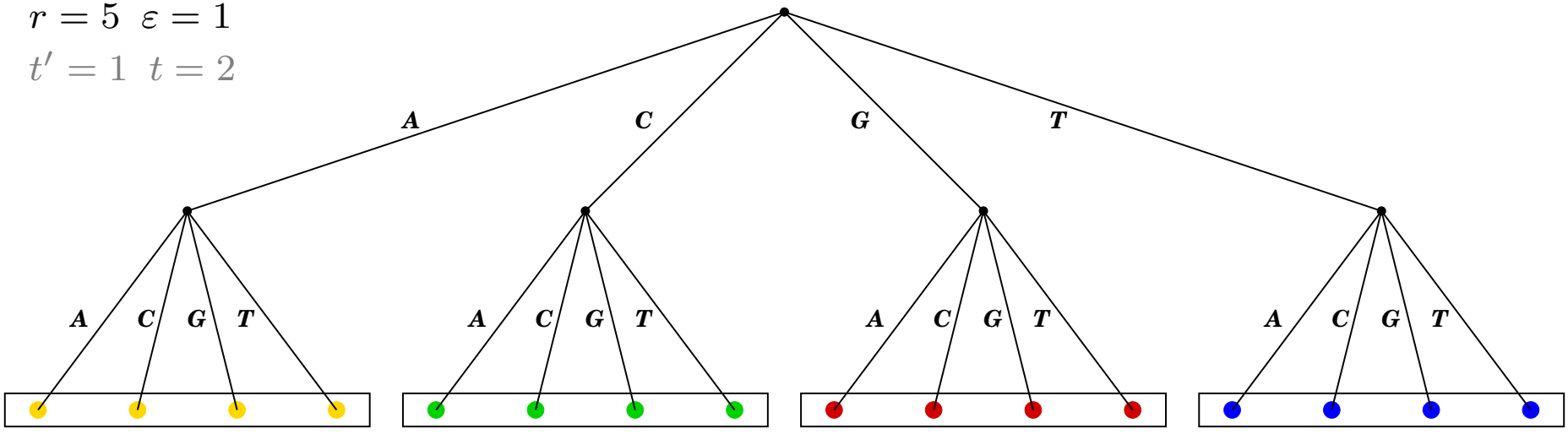
j	1	2	3	4
σ_j	A	C	T	G
w_j	2	1	1	2

find the smallest t' such that $\frac{max\ w_j}{c^{t'}} \leq \varepsilon$

t = min(**depth(**$\mathcal{M}$**) -** 1 **,**t'**)**

$r = 5 \quad \varepsilon = 1$

$t' = 1 \quad t = 2$



Parallelization

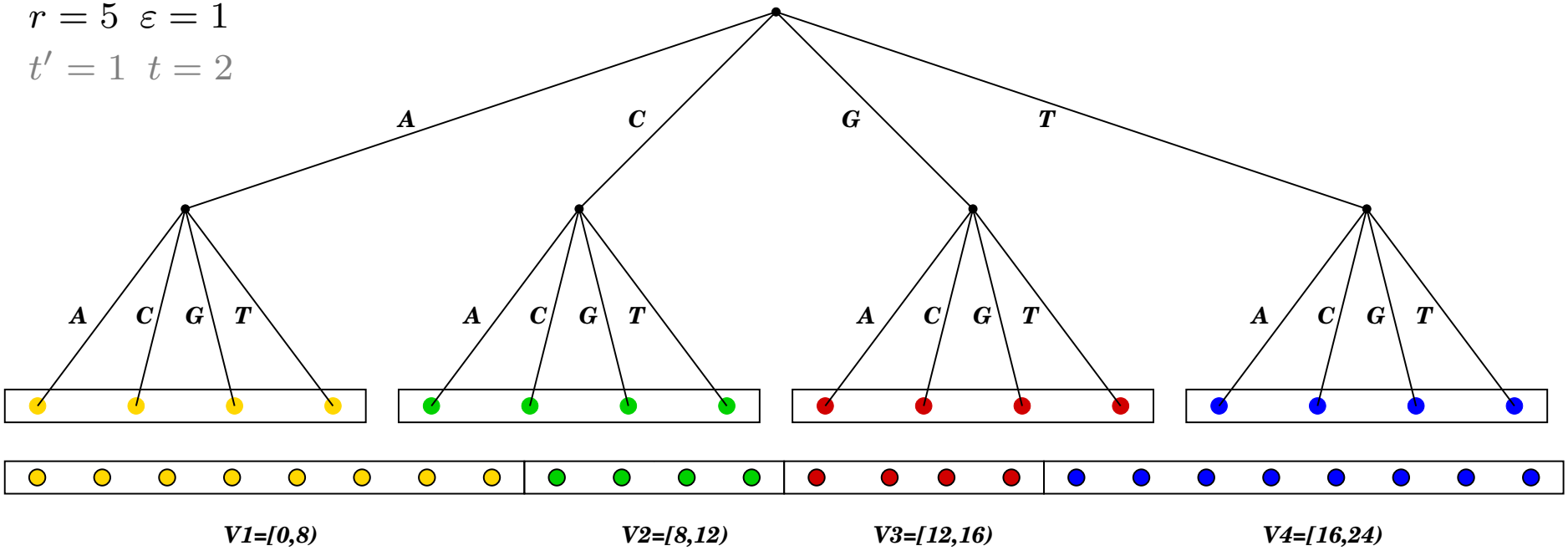
j	1	2	3	4
σ_j	A	C	T	G
w_j	2	1	1	2

for each $j \in 1, \ldots, \ell$

$$V_j = \left[\sum_{k=1}^{j-1} w_k \times c^t, \sum_{k=1}^{j} w_k \times c^t\right)$$

$r = 5 \;\; \varepsilon = 1$

$t' = 1 \;\; t = 2$



Parallelization

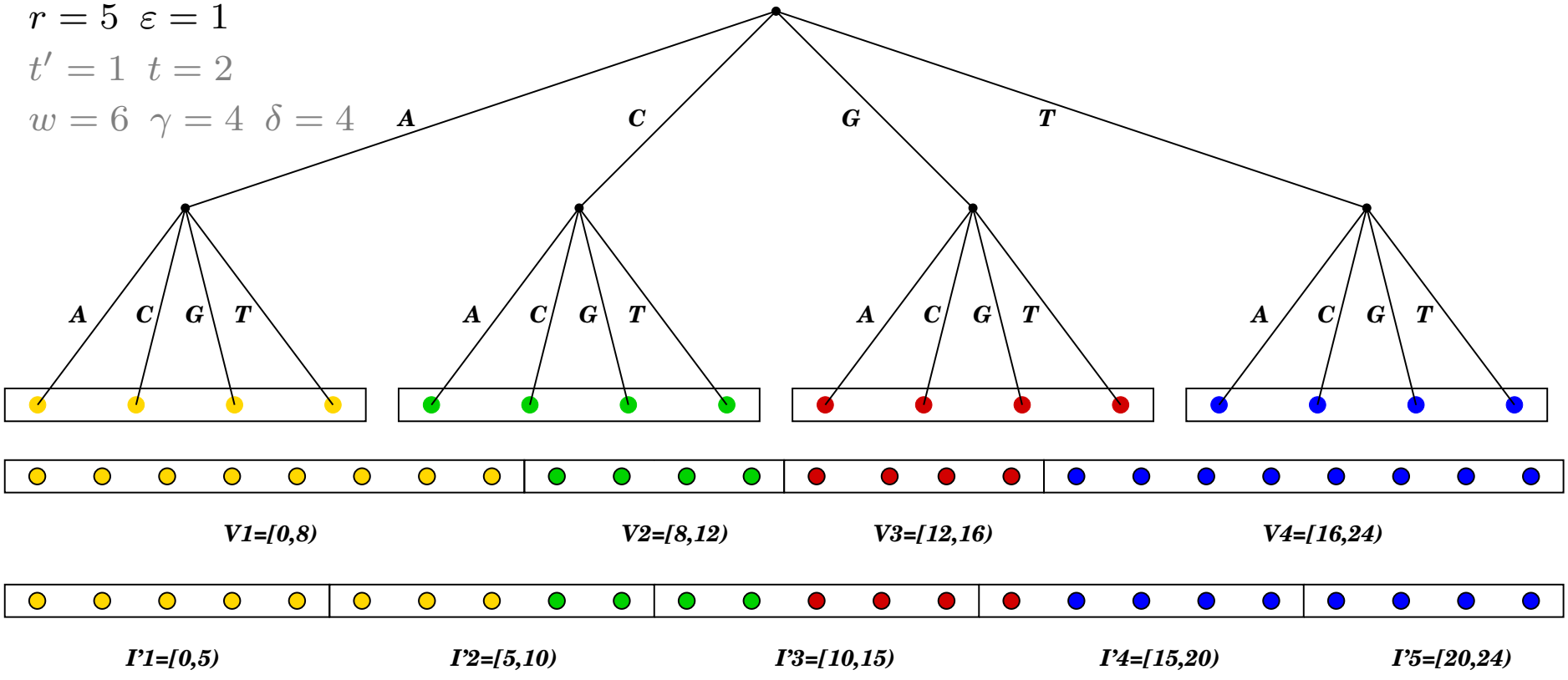
j	1	2	3	4
σ_j	A	C	T	G
w_j	2	1	1	2

$$I'_i = \begin{cases} [(i-1)(\delta+1),\ i(\delta+1)\,) & \text{for all } i \leq \gamma \\ [\gamma(\delta+1)+(i-(\gamma+1))\delta,\ \gamma(\delta+1)+(i-\gamma)\delta\,) & \text{otherwise} \end{cases}$$

$r = 5 \quad \varepsilon = 1$

$t' = 1 \quad t = 2$

$w = 6 \quad \gamma = 4 \quad \delta = 4$



Parallelization

j	1	2	3	4
σ_j	A	C	T	G
w_j	2	1	1	2

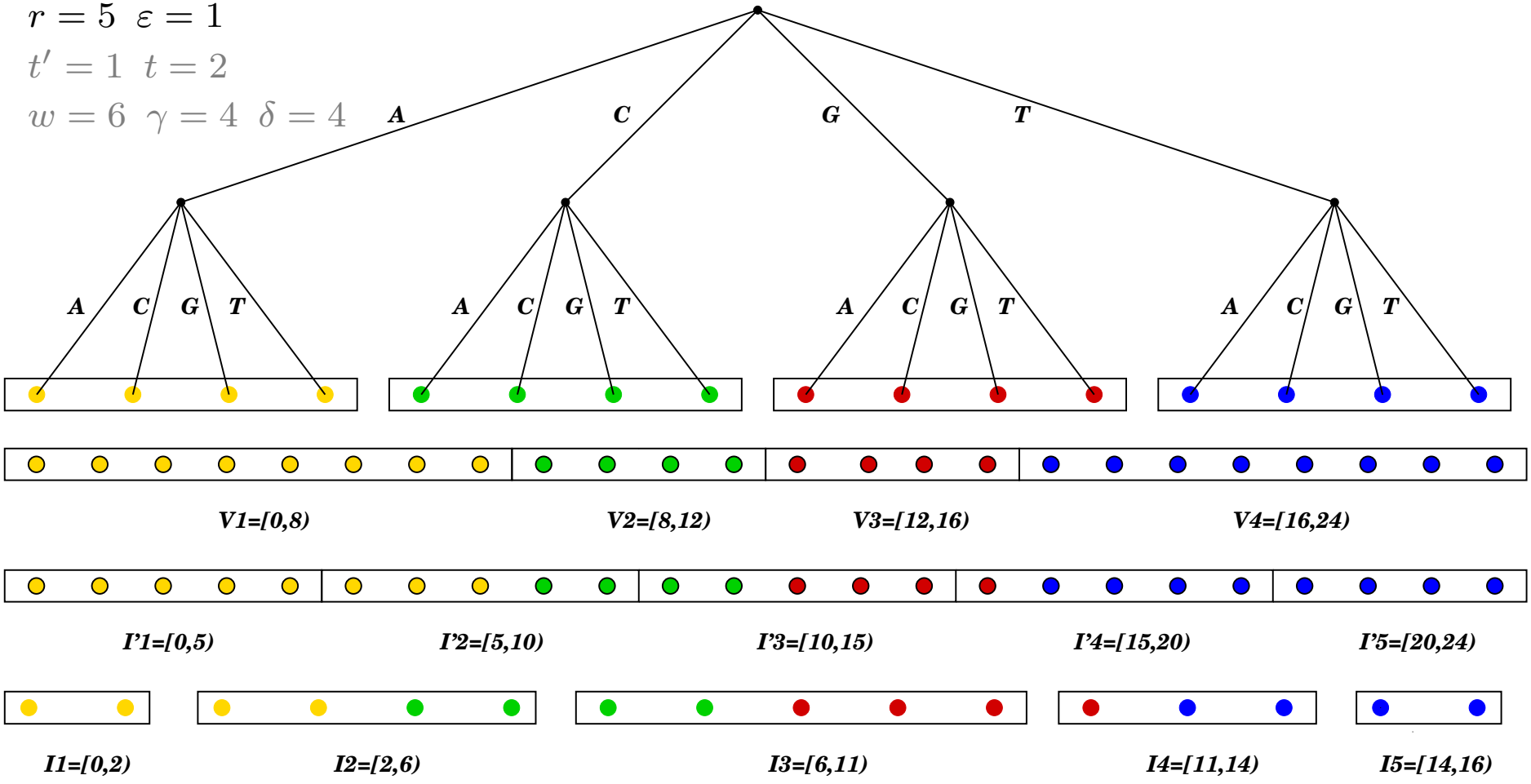
transform $I'_i = [a, b)$ into $I_i = [f(a), f(b))$ with

$$f(x) = \begin{cases} (j-1) \times c^t + \frac{x - \inf(V_j)}{w_j} & \texttt{for all } x \in V_j \\ \ell \times c^t & \texttt{if } x = w \times c^t \end{cases}$$

$r = 5 \;\; \varepsilon = 1$

$t' = 1 \;\; t = 2$

$w = 6 \;\; \gamma = 4 \;\; \delta = 4$



Parallelization

PSmile (**GridNode** i, **WorkOverload** ε)

1. compute weights $(w_i)_{1 \leq i \leq |\Sigma|}$;
2. build suffix tree $\mathcal{T}$;
3. let I_i= SimpleCut$(i, |\Sigma|, r, (w_i)_{1 \leq i \leq |\Sigma|}, |\Sigma|, \varepsilon)$;
4. call PExtractModels($\mathcal{T}$, I_i);

Proposition. Assume Σ fixed and $w_i = 1$ for $1 \leq i \leq |\Sigma|$. The parallel algorithm PSmile is work-efficient with respect to the sequential version when $r = O(\nu^{\frac{p}{2}}(e, k))$ and $\frac{\varepsilon}{w} \leq \frac{1}{r}$.

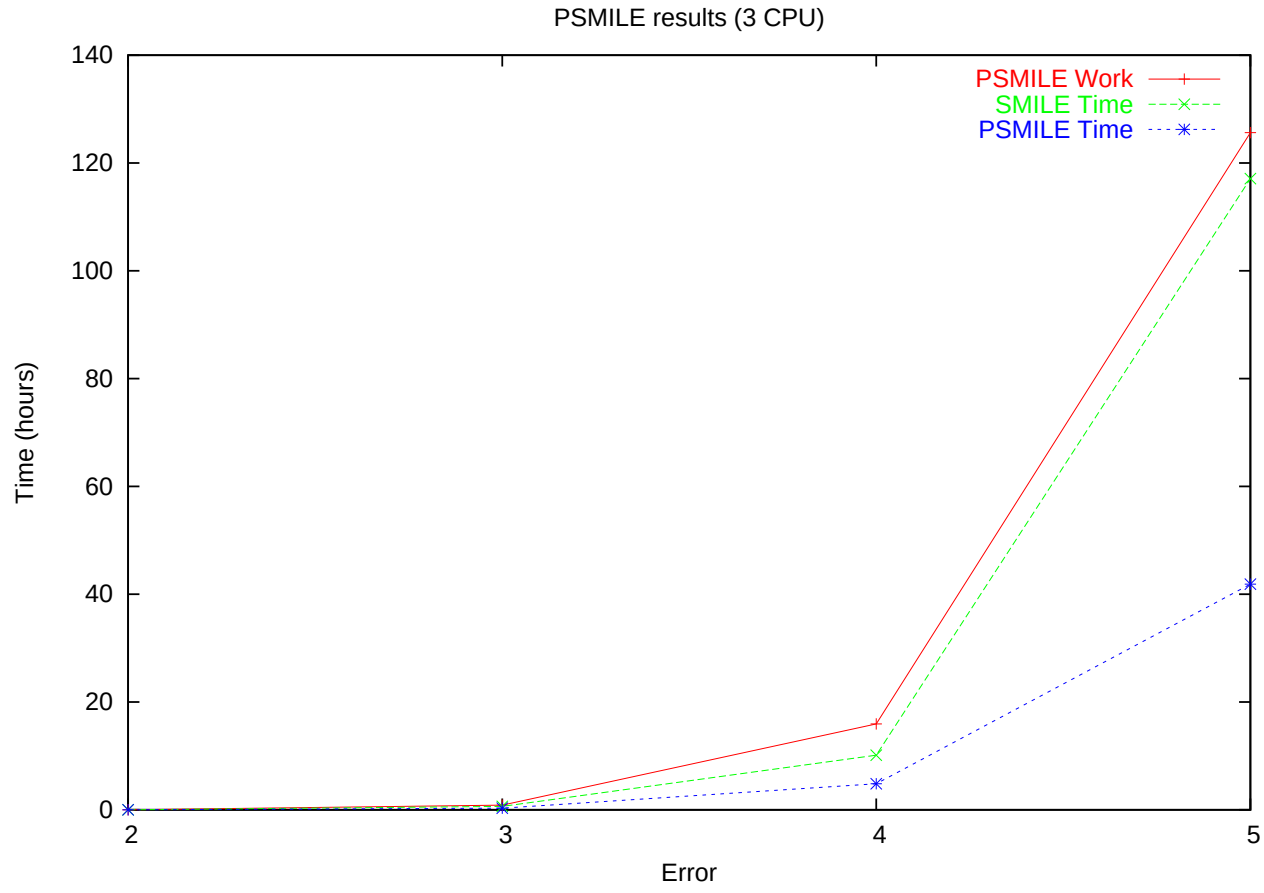
Experimental results

Infrastructure:

- Grid technology
- 3 CPU

Data set:

- *H. pylori* organism
- 1148 sequences
- 226928 nucleotides



Error	2	3	4	5
speed up	2.0	2.2	2.1	2.8

Ongoing and future work

- A more efficient sequential algorithm to extract structured models
  [A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, submitted, 2004]
- Parallelization of this new algorithm with the same technique
- Integration of the new parallel algorithm with a database of transcription factors and respective promoter consensus motifs for the yeast organism, in order to:
  - provide semi-automatic methods for processing experimental results
  - allow users to analyze complex interactions between gene networks and proteins