Wallbreaker

Schulz, Mihov, Mitankin, Gerdjikov

CIS Technische Universität München, IICT Bulgarian Academy of Science,FMI Sofia University,

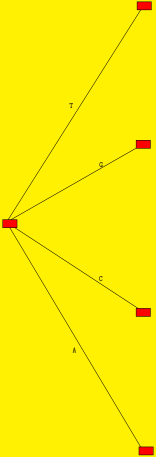
22 March 2013, Genoa

The Wall Effect

```
TGTTTCTCAATCCCACTTACTATTTTATCAAAAGTTATCTCCACCAAT,8
```

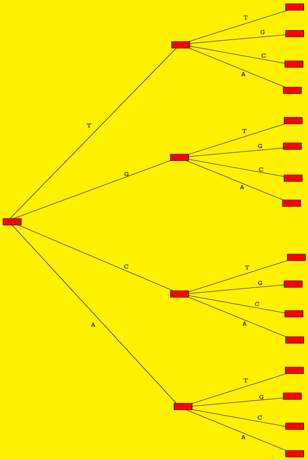
The Wall Effect

TGTTTCTCAATCCCACTTACTATTTTATCAAAAGTTATCTCCACCAAT,8



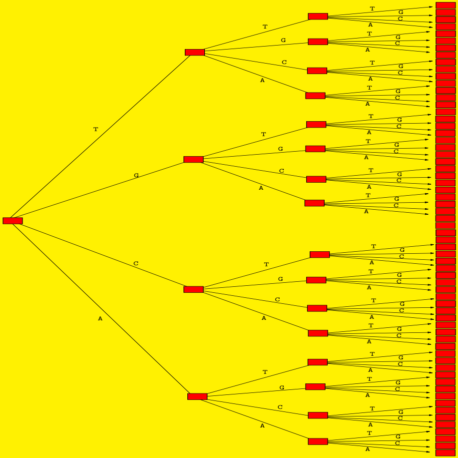
The Wall Effect

TGTTTCTCAATCCCACTTACTATTTTATCAAAAGTTATCTCCACCAAT,8



The Wall Effect

TGTTTCTCAATCCCACTTACTATTTTATCAAAAGTTATCTCCACCAAT,8

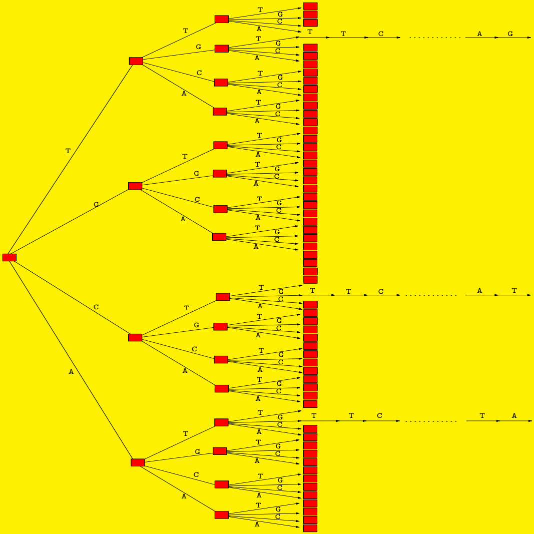


Through the Wall

TGTTTCTCAATCCCACTTACTATTTTATCAAAAGTTATCTCCACCAAT,8

Through the Wall

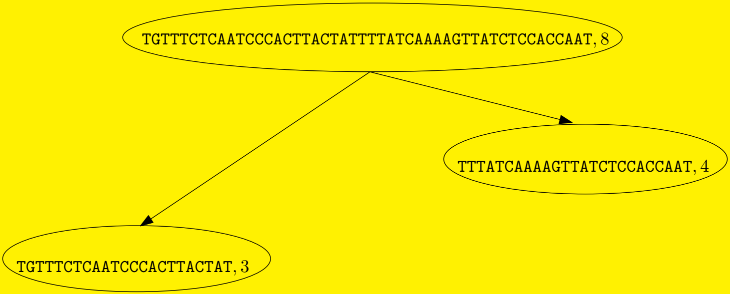
TGTTTCTCAATCCCACTTACTATTTTATCAAAAGTTATCTCCACCAAT,8



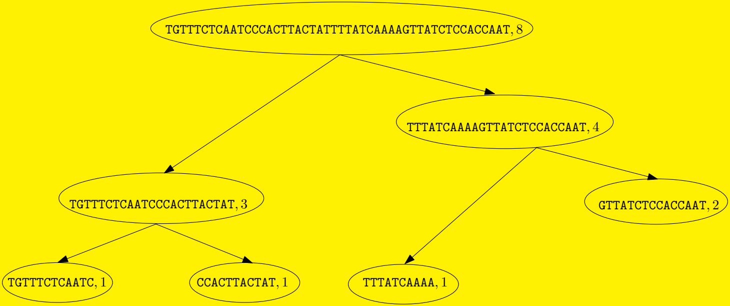
Divide and Conquer

TGTTTCTCAATCCCACTTACTATTTTATCAAAAGTTATCTCCACCAAT, 8

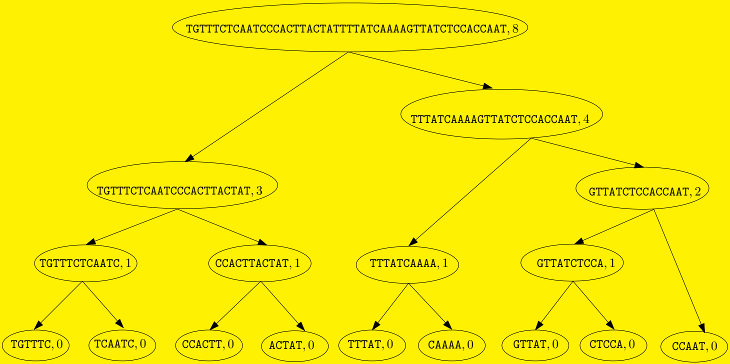
Divide and Conquer



Divide and Conquer

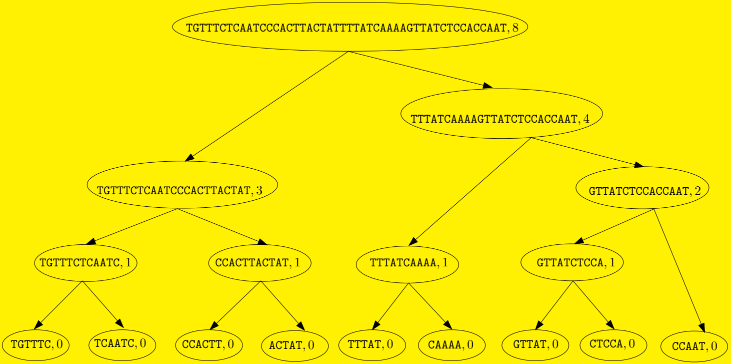


Divide and Conquer



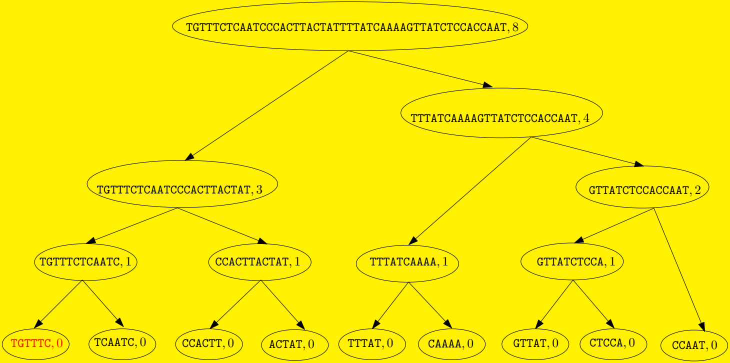
Answering Query

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT



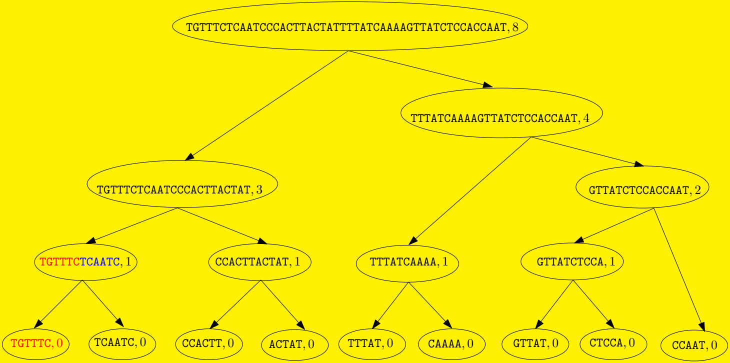
Answering Query

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT



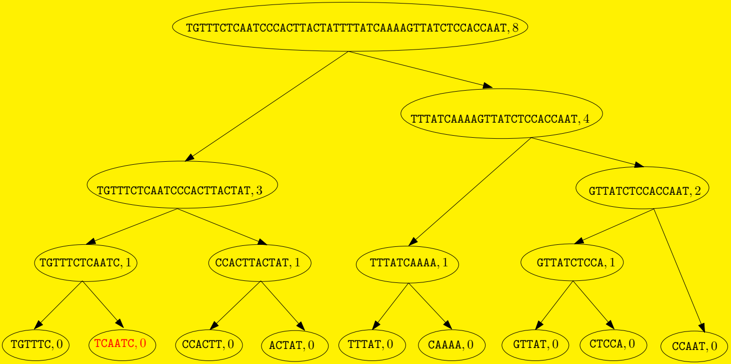
Answering Query

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT



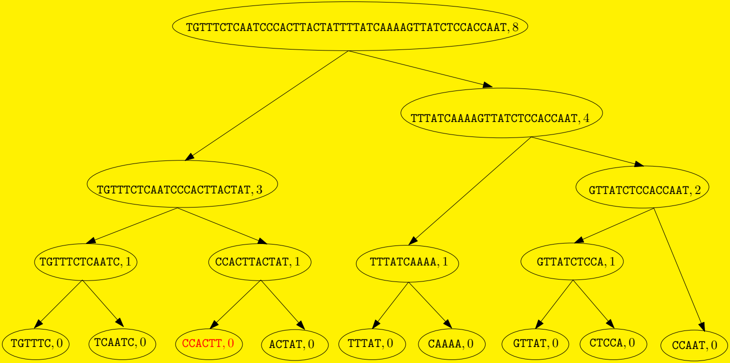
Answering Query

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT



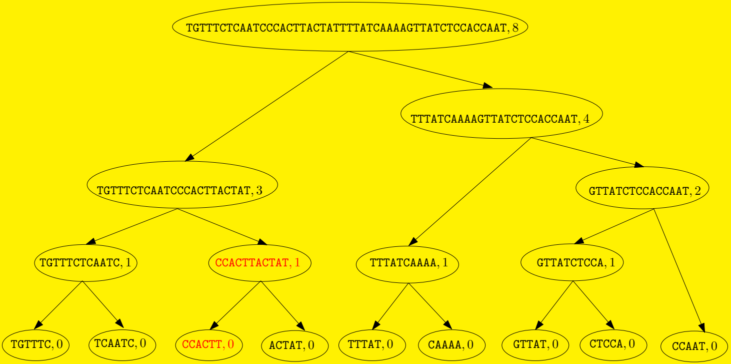
Answering Query 2

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT



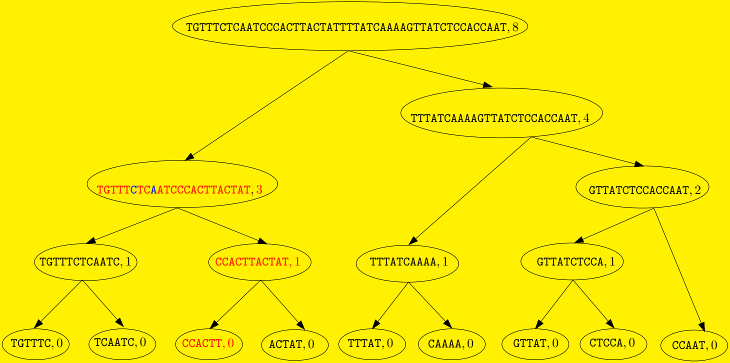
Answering Query 2

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTTCATC
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT



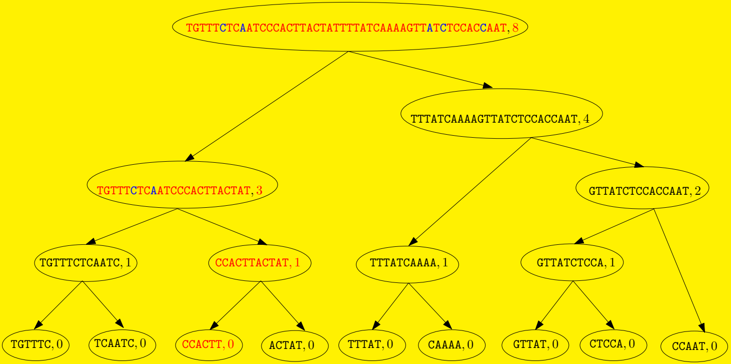
Answering Query 2

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT

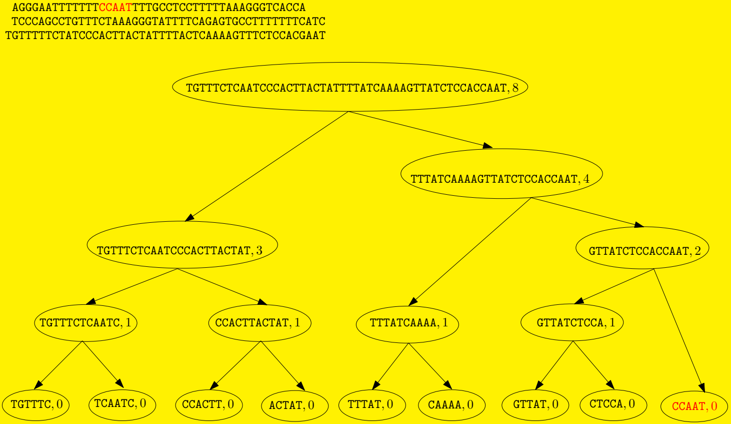


Answering Query 2

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTTCCACGAAT

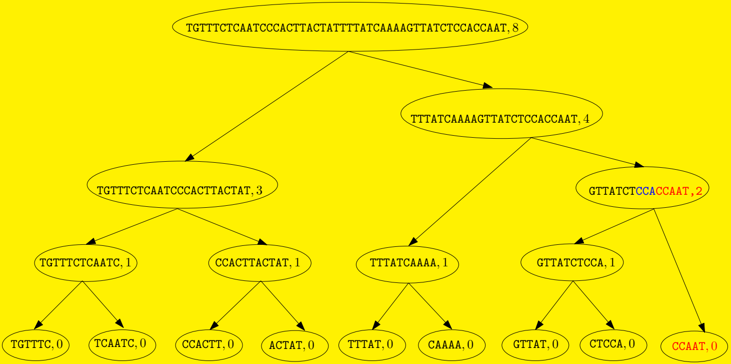


Answering Query 2



Answering Query 2

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT



Resources

1. efficient left and right extensions.

Resources

1. efficient left and right extensions.
2. filter(s) to supervise the search.

Efficient Representation of Infixes

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA,
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC,
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT

Efficient Representation of Infixes

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA,
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC,
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT

Efficient Representation of Infixes

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA,
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC,
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT

Efficient Representation of Infixes

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA,
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC,
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT

Efficient Representation of Infixes

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA,
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC,
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT

Efficient Representation of Infixes

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA,
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC,
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT

Efficient Representation of Infixes

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA,
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC,
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT

Definition

[Inenaga05] If $\mathcal{S}$ is a finite set of words and X is an infix in $\mathcal{S}$, then $\overleftrightarrow{X} = \alpha \circ X \circ \beta$ is the longest infix in $\mathcal{S}$ such that whenever for a word $W \in \mathcal{S}$, $W = U \circ X \circ V$ it is true that $U = U_1 \circ \alpha$ and $V = \beta \circ V_1$.

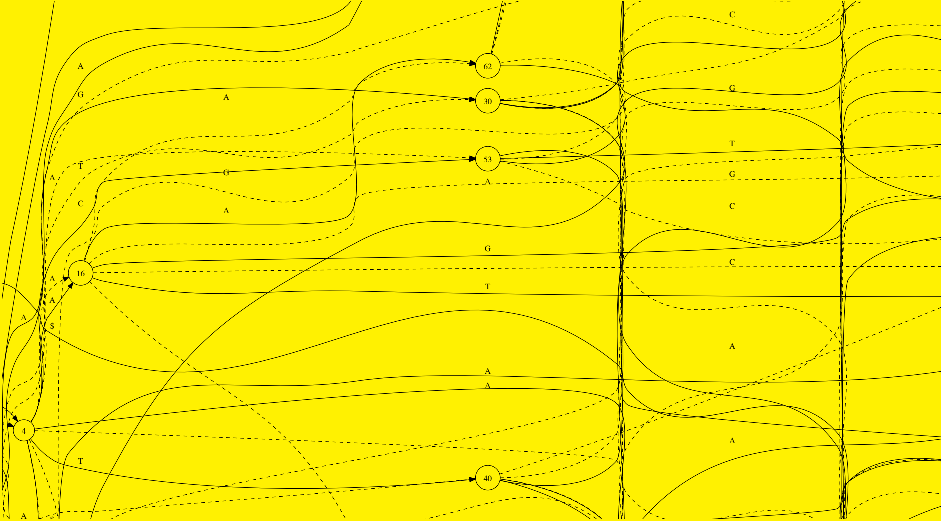
Efficient Representation of Infixes

AGGGAATTTTTTTCCAATTTTGCCTCCTTTTTAAAGGGTCACCA,
TCCCAGCCTGTTTCTAAAGGGTATTTTCAGAGTGCCTTTTTTTCATC,
TGTTTTTCTATCCCACTTACTATTTTACTCAAAAGTTTCTCCACGAAT

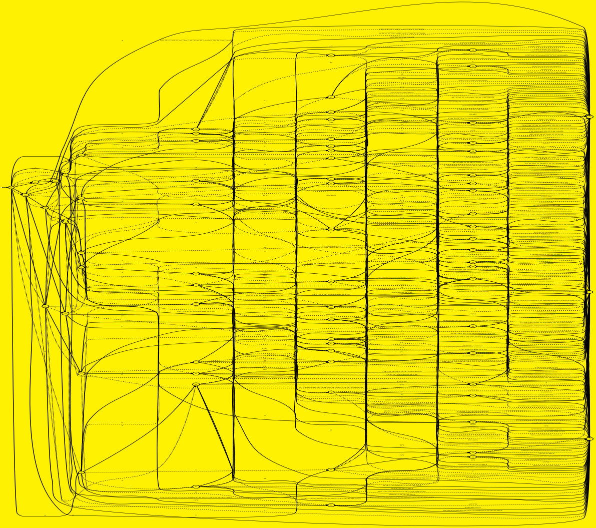
Definition

[Inenaga05] If $\mathcal{S}$ is a finite set of words and X is an infix in $\mathcal{S}$, then $\overleftrightarrow{X} = \alpha \circ X \circ \beta$ is the longest infix in $\mathcal{S}$ such that whenever for a word $W \in \mathcal{S}$, $W = U \circ X \circ V$ it is true that $U = U_1 \circ \alpha$ and $V = \beta \circ V_1$.
$X \equiv_{\mathcal{S}} Y$ iff $\overleftrightarrow{X} = \overleftrightarrow{Y}$.

SCDWAG for a Set of Words



SCDWAG for a Set of Words



SDWAG for a set of words, $\mathcal{S}$

1. states - $\{\overleftrightarrow{X} \mid X \in \mathit{Inf}(\mathcal{S})\}$
2. *right* transitions - $\{\langle \overleftrightarrow{X}, a, \overleftrightarrow{\overleftrightarrow{X} a} \rangle \mid \overleftrightarrow{X} a \in \mathit{Inf}(\mathcal{S})\}$

SDWAG for a set of words, $\mathcal{S}$

1. states - $\{\overleftrightarrow{X} \mid X \in \mathit{Inf}(\mathcal{S})\}$
2. *right* transitions - $\{\langle \overleftrightarrow{X}, a, \overleftrightarrow{\overleftrightarrow{X} a} \rangle \mid \overleftrightarrow{X} a \in \mathit{Inf}(\mathcal{S})\}$
3. *left* transitions - $\{\langle \overleftrightarrow{X}, a, \overleftrightarrow{a \overleftrightarrow{X}} \rangle \mid a \overleftrightarrow{X} \in \mathit{Inf}(\mathcal{S})\}$

Filters

1. Ukkonen's dynamic programming.

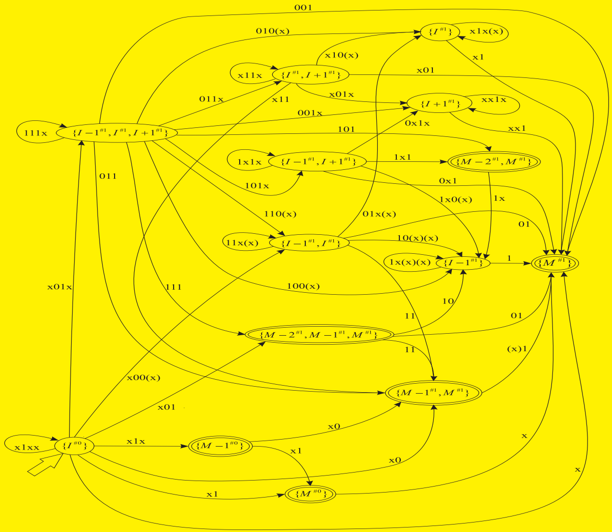
Filters

1. Ukkonen's dynamic programming.
2. Universal Levenshtein automata.

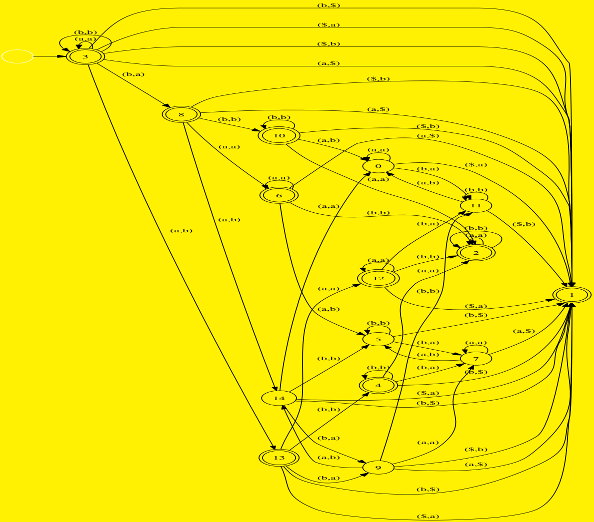
Filters

1. Ukkonen's dynamic programming.
2. Universal Levenshtein automata.
3. Synchronised Levenshtein automata.

Universal Levenshtein Automata



Synchronised Levenshtein Automata



Results

- Genome reads, 5% excerpt provided by the organisers, 750 000 strings.

Results

- Genome reads, 5% excerpt provided by the organisers, 750 000 strings.

Wallbreaker	space	time
construction	2623.13 MB	324 sec

Results

- Genome reads, 5% excerpt provided by the organisers, 750 000 strings.

Wallbreaker	space	time
construction	2623.13 MB	324 sec

- 100 000 randomly generated queries and thresholds in $\{0, 4, 8, 12, 16\}$; 16 threads.

Results

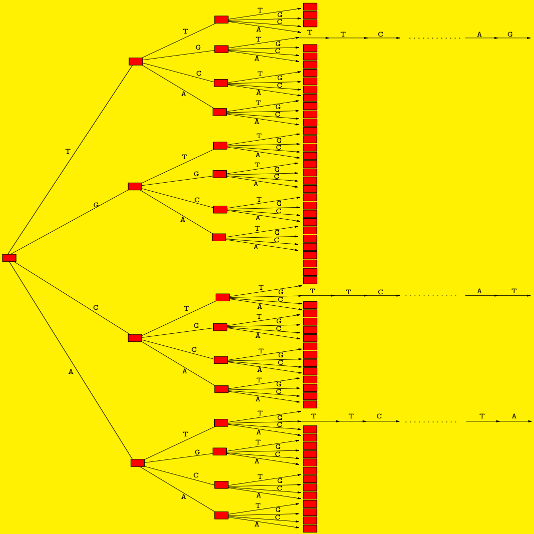
- Genome reads, 5% excerpt provided by the organisers, 750 000 strings.

Wallbreaker	space	time
construction	2623.13 MB	324 sec

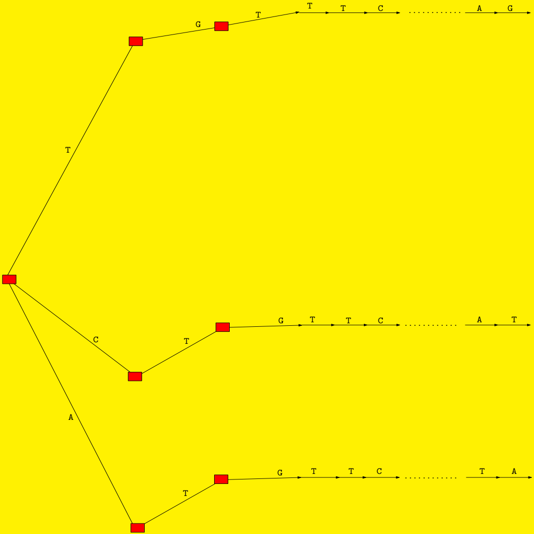
- 100 000 randomly generated queries and thresholds in $\{0, 4, 8, 12, 16\}$; 16 threads.

Wallbreaker	space	total time	average time
query	2623.13 MB	8.794 sec	0.000 008 794 sec

Breaking the Wall



Breaking the Wall



S.G., S. Mihov, P. Mitankin, K. Schulz, 2013.
Good parts first - a new algorithm for approximate search in lexica and string databases.
ArXiv e-prints, http://adsabs.harvard.edu/abs/2013arXiv1301.0722G.

S. Inenaga, H. Hoshino, A. Shinohara,M. Takeda, S. Arikawa, G. Mauri, G. Pavesi, 2005.
On-Line Construction of Compact Directed Acyclic Word Graphs.
Word Journal Of The International Linguistic Association, 146(2):1–12.

Stoyan Mihov and Klaus U. Schulz, 2004.
Fast approximate search in large dictionaries.
Computational Linguistic, 30(4):451–477.

Petar Mitankin and Stoyan Mihov and Klaus U. Schulz, 2011.
Deciding word neighborhood with universal neighborhood automata.
Theoretical Computer Science, 412(22):2340–2355.