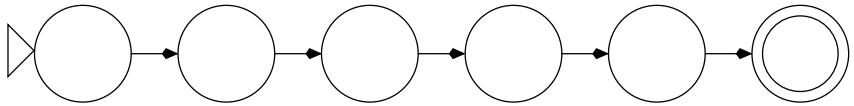
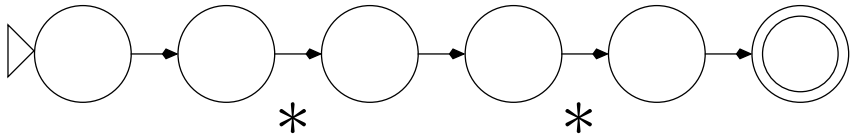
Approximate seeds, seeds with errors

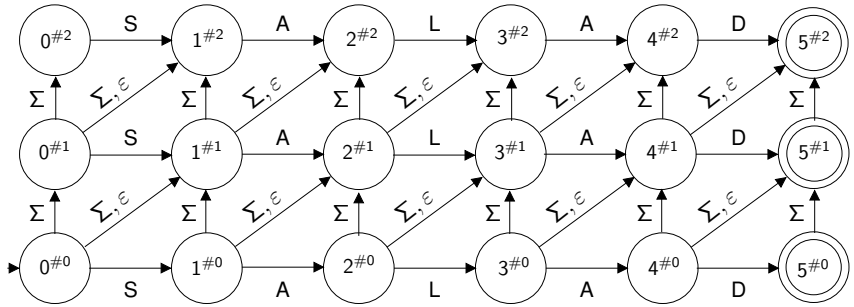
contiguous seed



spaced seed



seed with errors



Seeds with errors : Levenshtein distance

- alphabet Σ
- two words over Σ
- three edit operations

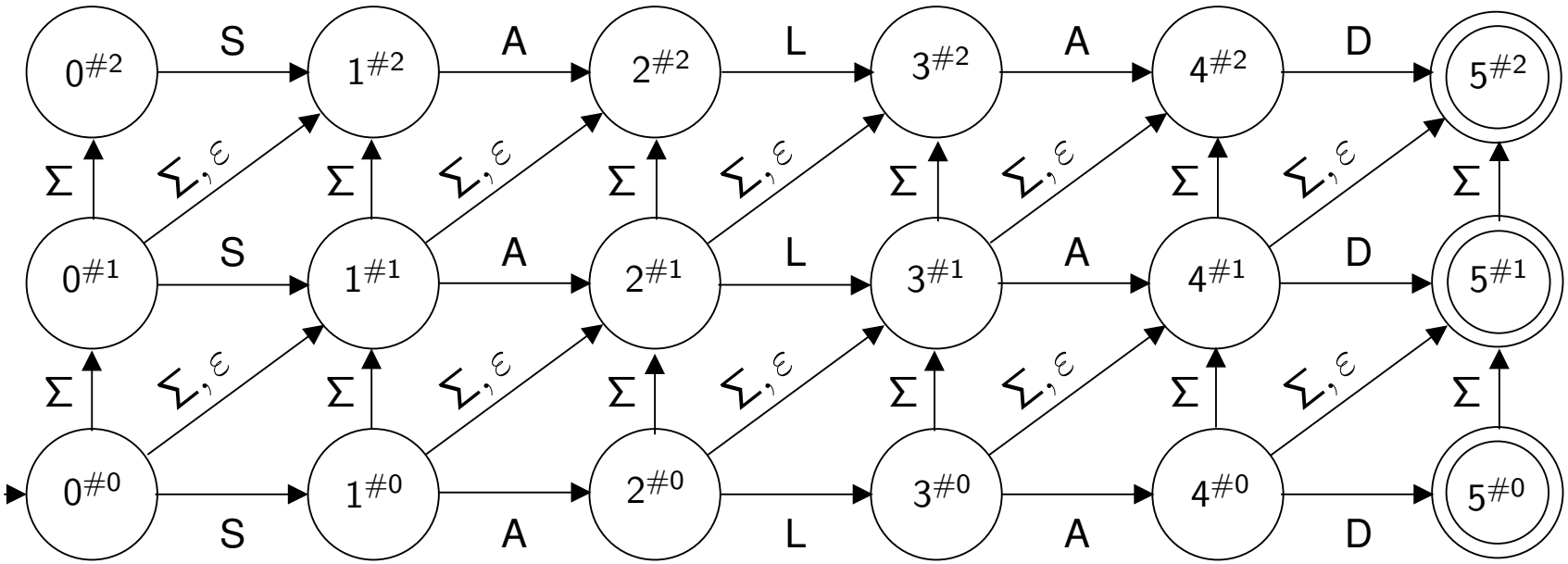
```
A B L        A - L        A B L
|   |        |   |        |   |
A L L        A B L        A - L

substitution  insertion    deletion
```

- Levenshtein distance : smallest number of operations needed to transform one word into another

```
B A L L A D -
  | |   | |
S A L - A D S
```

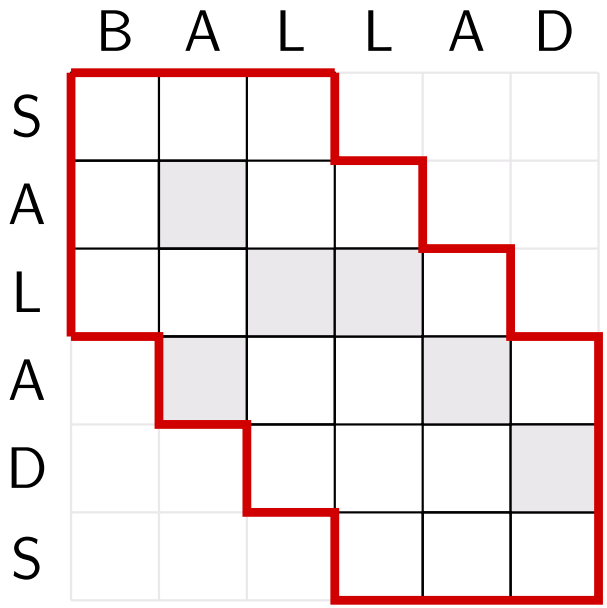
Levenshtein automaton



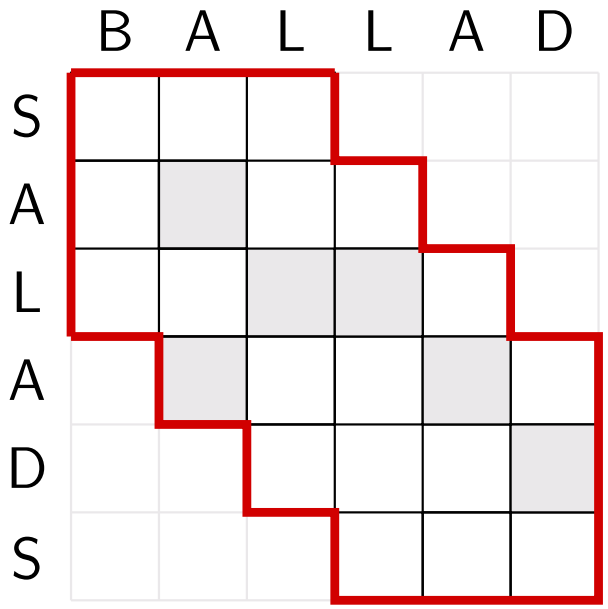
$\rightarrow$: identity $\quad \uparrow$: insertion $\quad \nearrow$: substitution (Σ), deletion (ε)

P=SALAD and $k = 2$

Bit vector representation



Bit vector representation



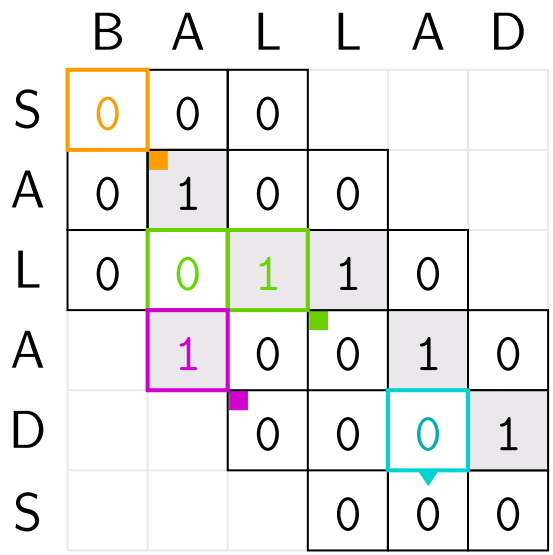
	$	$	B	A	L	L	A	D	$	$	$	$
S	0	0	0	0	0							
A		0	0	1	0	0						
L			0	0	1	1	0					
A				1	0	0	1	0				
D					0	0	0	1	0			
S						0	0	0	0	0		
$							0	0	1	1	1	
$								0	1	1	1	1

- P=BALLAD, V=SALAD, $k = 2$
- pattern $P \rightarrow \$^k P \$^{2k}$
- word $V \rightarrow V\$^{|P|-|V|+k}$
- sequence of $|P| + k$ bit vectors of length $2k + 1$

Nondeterministic Universal Levenshtein Automaton

	B	A	L	L	A	D
S	0	0	0			
A	0	1	0	0		
L	0	0	1	1	0	
A		1	0	0	1	0
D			0	0	0	1
S				0	0	0

Nondeterministic Universal Levenshtein Automaton



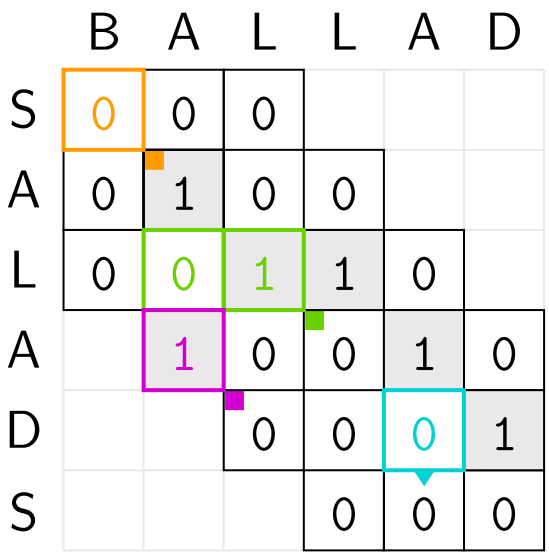
substitution

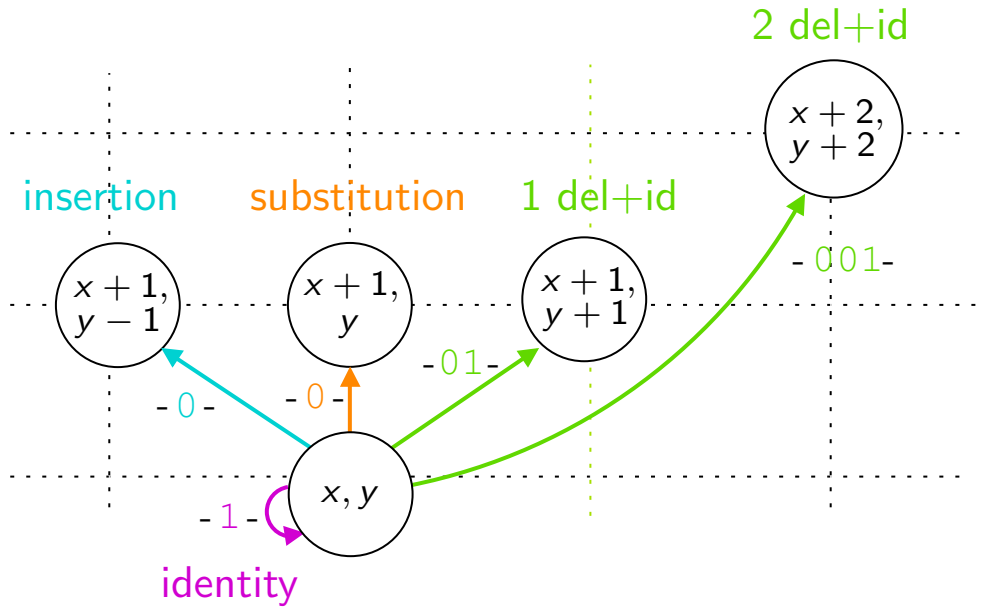
deletion (1 del + id)

identity

insertion

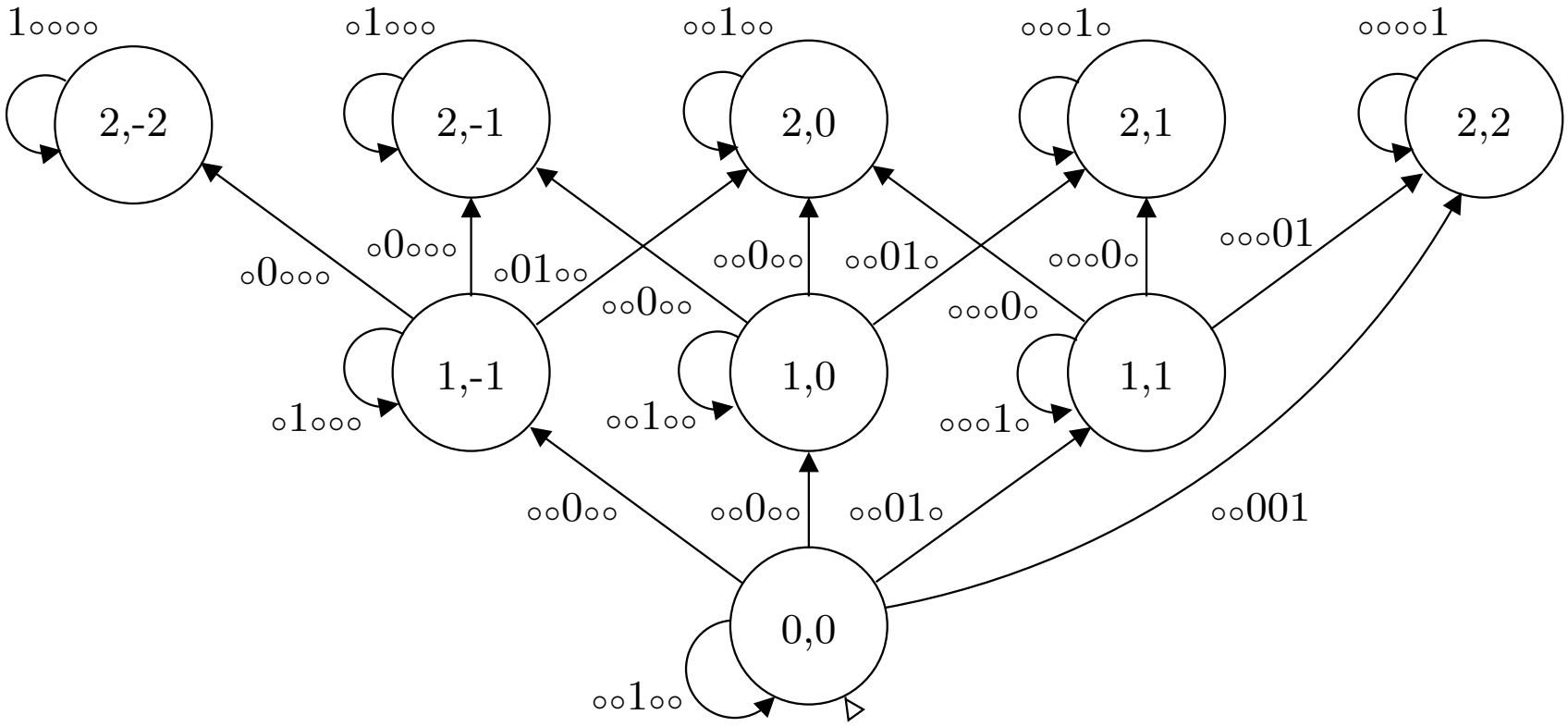
Nondeterministic Universal Levenshtein Automaton





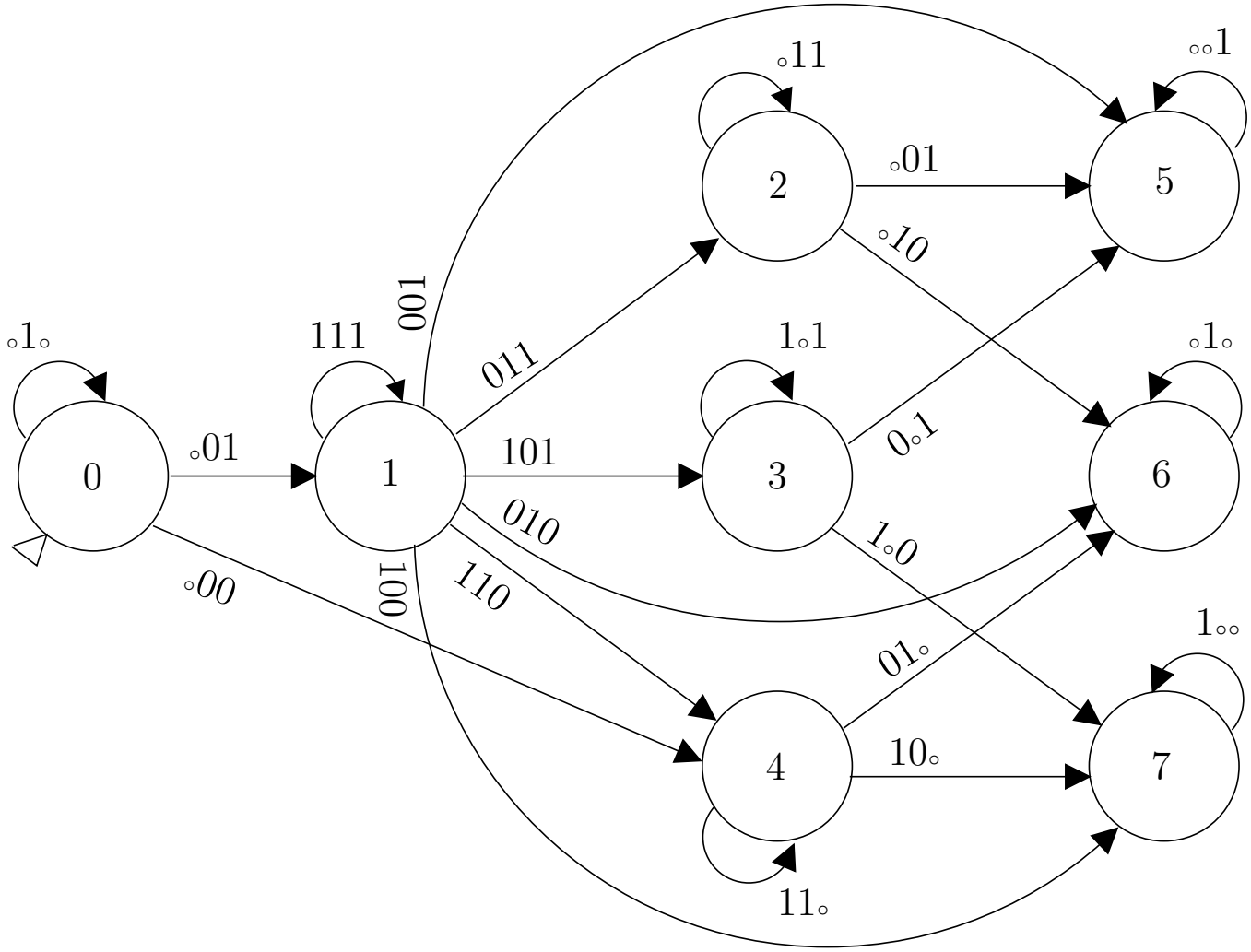
state (x, y) : "I am in the lane y and have made x errors so far"

Nondeterministic Universal Levenshtein Automaton



$\mathrm{NULA}(2)$ $(k = 2)$

Deterministic Universal Levenshtein Automaton



$\mathrm{DULA}(1)$ $(k = 1)$

Seeds with error : a new type of seeds

Seeds with error : a new type of seeds

3 errors within the seed

AUCAGUGCAAAUGCUCAAGA

Seeds with error : a new type of seeds

3 errors within the seed

AUCAGUGCAAAUGCUCAAGA

AUCAG UGCAA AUGCU CAAGA

4 parts of length 5
1 part out of 4
"selectivity" : 1/256

Seeds with error : a new type of seeds

3 errors within the seed

AUCAGUGCAAAUGCUCAAGA

AUCAG UGCAA AUGCU CAAGA — 4 parts of length 5

1 part out of 4

"selectivity" : 1/256

AUCA GUGC AAAU GCUC AAGA — 5 parts of length 4

Seeds with error : a new type of seeds

3 errors within the seed

`AUCAGUGCAAAUGCUCAAGA`

`AUCAG UGCAA AUGCU CAAGA`

4 parts of length 5
1 part out of 4
"selectivity" : 1/256

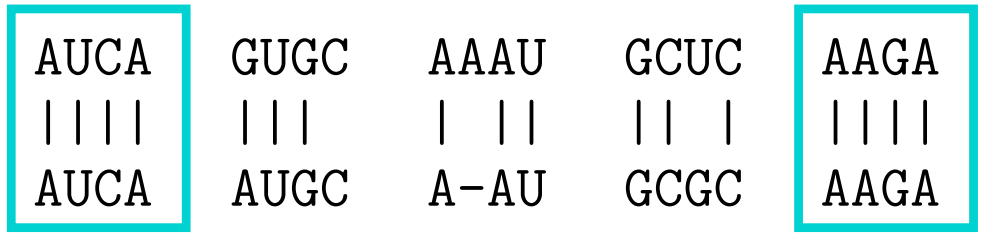
`AUCA GUGC AAAU GCUC AAGA`

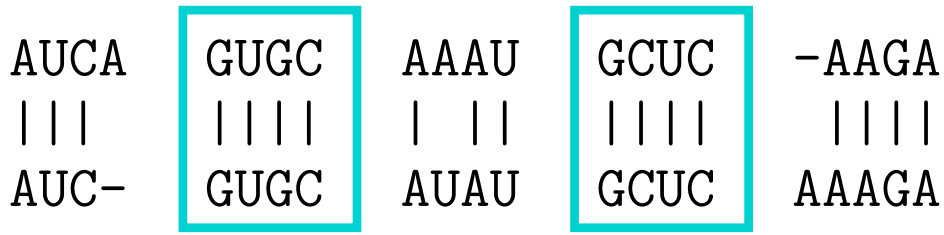
5 parts of length 4
2 parts out of 5
"selectivity" : 1/2100

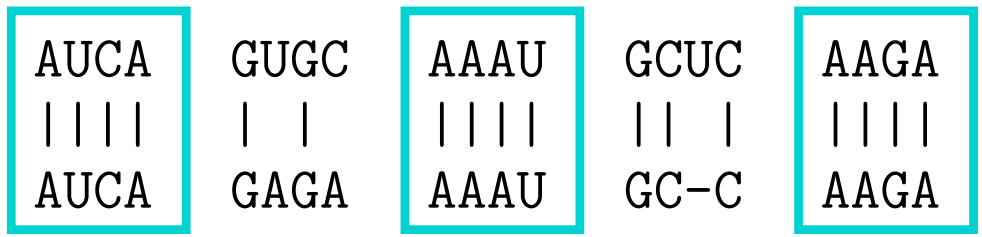
```
AUCA   GUGC   AAAU   GCUC   AAGA
||||   |||    | ||   || |   ||||
AUCA   AUGC   A-AU   GCGC   AAGA
```

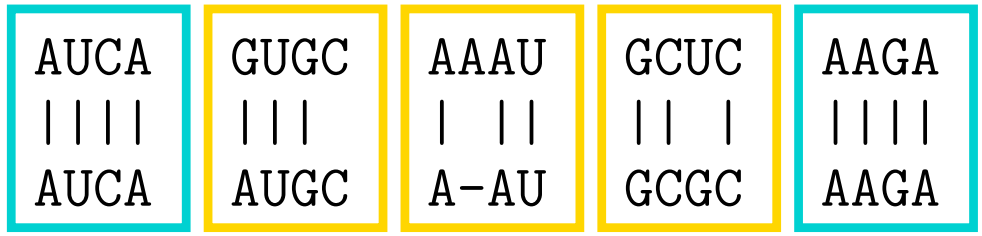
```
AUCA   GUGC   AAAU   GCUC   -AAGA
|||    ||||   | ||   ||||    ||||
AUC-   GUGC   AUAU   GCUC   AAAGA
```

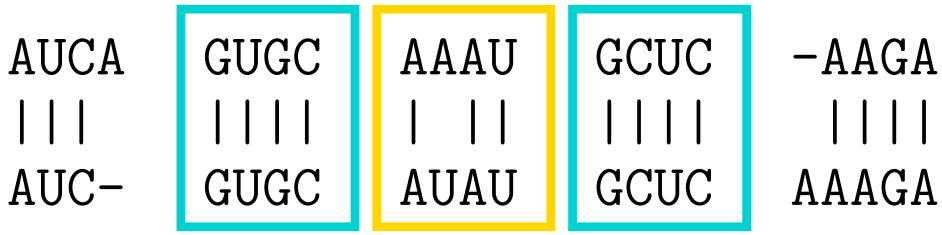
```
AUCA   GUGC   AAAU   GCUC   AAGA
||||   | |    ||||   || |   ||||
AUCA   GAGA   AAAU   GC-C   AAGA
```

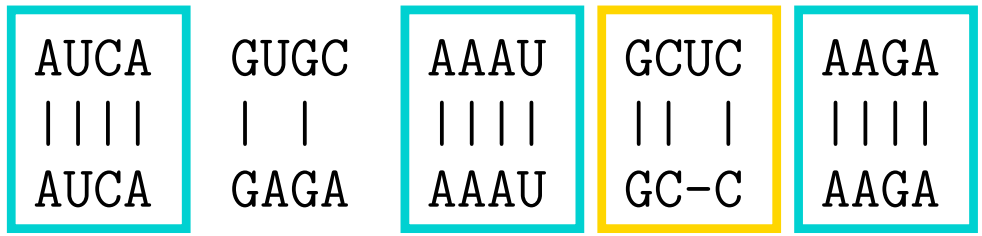

AUCA GUGC AAAU GCUC AAGA
AUCA AUGC A-AU GCGC AAGA

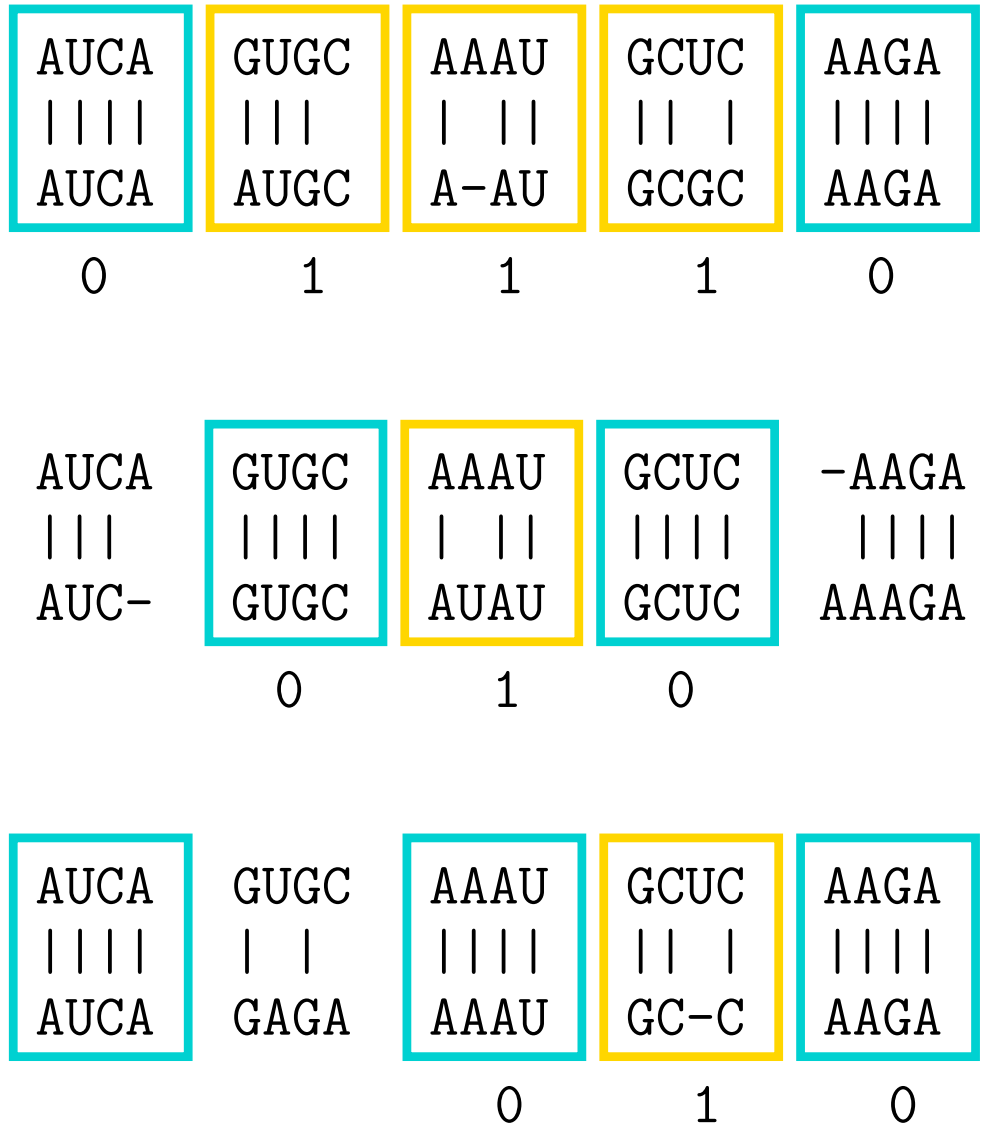

AUCA GUGC AAAU GCUC -AAGA
AUC- GUGC AUAU GCUC AAAGA


AUCA GUGC AAAU GCUC AAGA
AUCA GAGA AAAU GC-C AAGA


AUCA
||||
AUCA
GUGC
|||
AUGC
AAAU
| ||
A-AU
GCUC
|| |
GCGC
AAGA
||||
AAGA

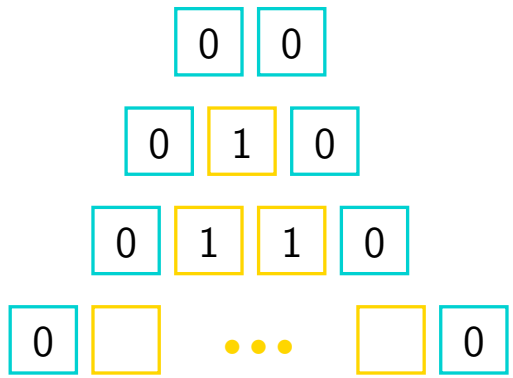

AUCA
|||
AUC-
GUGC
||||
GUGC
AAAU
| ||
AUAU
GCUC
||||
GCUC
-AAGA
||||
AAAGA


AUCA
||||
AUCA
GUGC
| |
GAGA
AAAU
||||
AAAU
GCUC
|| |
GC-C
AAGA
||||
AAGA


AUCA
||||
AUCA
GUGC
|||
AUGC
AAAU
| ||
A-AU
GCUC
|| |
GCGC
AAGA
||||
AAGA
0
1
1
1
0
AUCA
|||
AUC-
GUGC
||||
GUGC
AAAU
| ||
AUAU
GCUC
||||
GCUC
-AAGA
||||
AAAGA
0
1
0
AUCA
||||
AUCA
GUGC
| |
GAGA
AAAU
||||
AAAU
GCUC
|| |
GC-C
AAGA
||||
AAGA
0
1
0

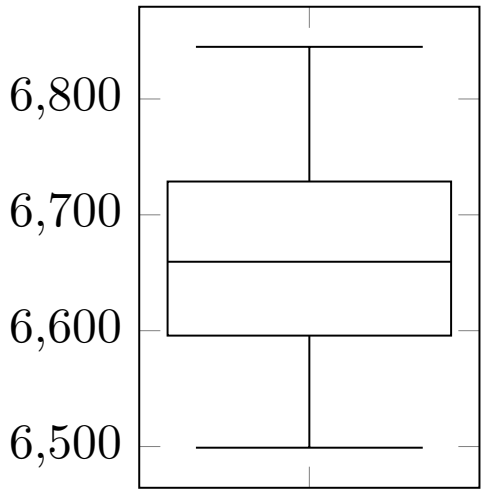
01^*0 seeds

- A 01^*0 seed is a pair of exact parts surrounding 0 or more parts with exactly 1 error.

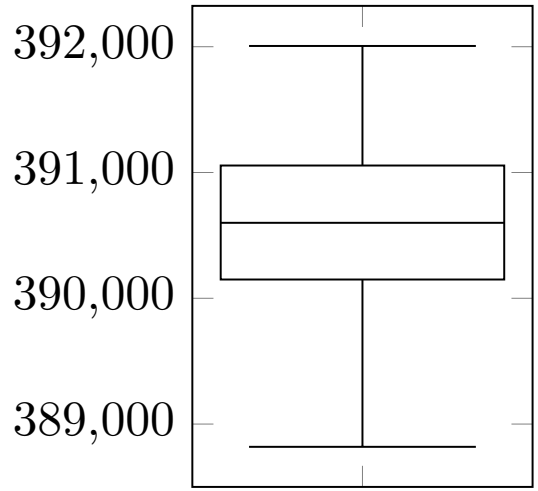


- Given a pattern partitioned in at least $k+2$ parts, then any occurrence with at most k errors contains such a 01^*0 seed.

01^*0 seeds – filtration efficiency



01^*0 seed



exact seed (5 parts of length 4)

number of seed occurrences per pattern

100 patterns of length 20, text of length 10^8, up to 3 errors