

# Bit-parallel string matching

We will discuss

- Ukkonens algorithm for  $k$ -differences matching
- Myers' bit vector algorithm

# Bit-parallel string matching (2)

1. Navarro and Raffinot (2002) *Flexible Pattern Matching in Strings*, sections 2.2.2 and 6.4.
2. G. Myers (1999) *A Fast Bit-Vector Algorithm for Approximate String Matching Based on Dynamic Programming*, Journal of the ACM, 46(3): 495-415
3. Heikki Hyyrö (2001) *Explaining and Extending the Bit-parallel Approximate String Matching Algorithm of Myers*, Technical report, University of Tampere, Finland.

# Shift-And versus Shift-Or

# Bit vector based approximate string matching

In this following we will focus on pairwise string alignments minimizing the *edit distance*. The algorithms covered are:

1. The classical dynamic programming algorithm, discovered and rediscovered many times since the 1960's. It computes a DP table indexed by the positions of the text and the pattern.
2. This can be improved using an idea due to Ukkonen, if we are interested in hits of a pattern in a text with *bounded* edit distance (which is usually the case).
3. The latter can be further speeded up using *bit parallelism* by an algorithm due to Myers. The key idea is to represent the *differences* between the entries in the DP matrix instead of their absolute values.

## **The ‘classical’ algorithm**

## The classical algorithm

We want to find all occurrences of a query  $P = p_1 p_2 \dots p_m$  that occur with  $k \geq 0$  differences (substitutions and indels) in a text  $T = t_1 t_2 \dots t_n$ .

The classic approach computes in time  $O(mn)$  a  $(m+1) \times (n+1)$  dynamic programming matrix  $C[0..m, 0..n]$  using the recurrence

$$C[i, j] = \min \left\{ \begin{array}{l} C[i-1, j-1] + \delta_{ij} \\ C[i-1, j] + 1 \\ C[i, j-1] + 1 \end{array} \right\},$$

where

$$\delta_{ij} = \begin{cases} 0, & \text{if } p_i = t_j, \\ 1, & \text{otherwise.} \end{cases}$$

Each location  $j$  in the last ( $m$ -th) row with  $C[m, j] \leq k$  is a solution to our query.

## The classical algorithm (2)

The matrix  $C$  is initialized:

- at the upper boundary by  $C[0, j] = 0$ , since an occurrence can start anywhere in the text, and
- at the left boundary by  $C[i, 0] = i$ , according to the edit distance.

## The classical algorithm (3)

**Example.** We want to find all occurrences with less than 2 differences of the query `annual` in the text `annealing`. The DP matrix looks as follows:

		A	N	N	E	A	L	I	N	G
	0	0	0	0	0	0	0	0	0	0
A	1	0	1	1	1	0	1	1	1	1
N	2	1	0	1	2	1	1	2	2	2
N	3	2	1	0	1	2	2	2	2	2
U	4	3	2	1	1	2	3	3	3	3
A	5	4	3	2	2	1	2	3	4	4
L	6	5	4	3	3	2	1	2	3	4

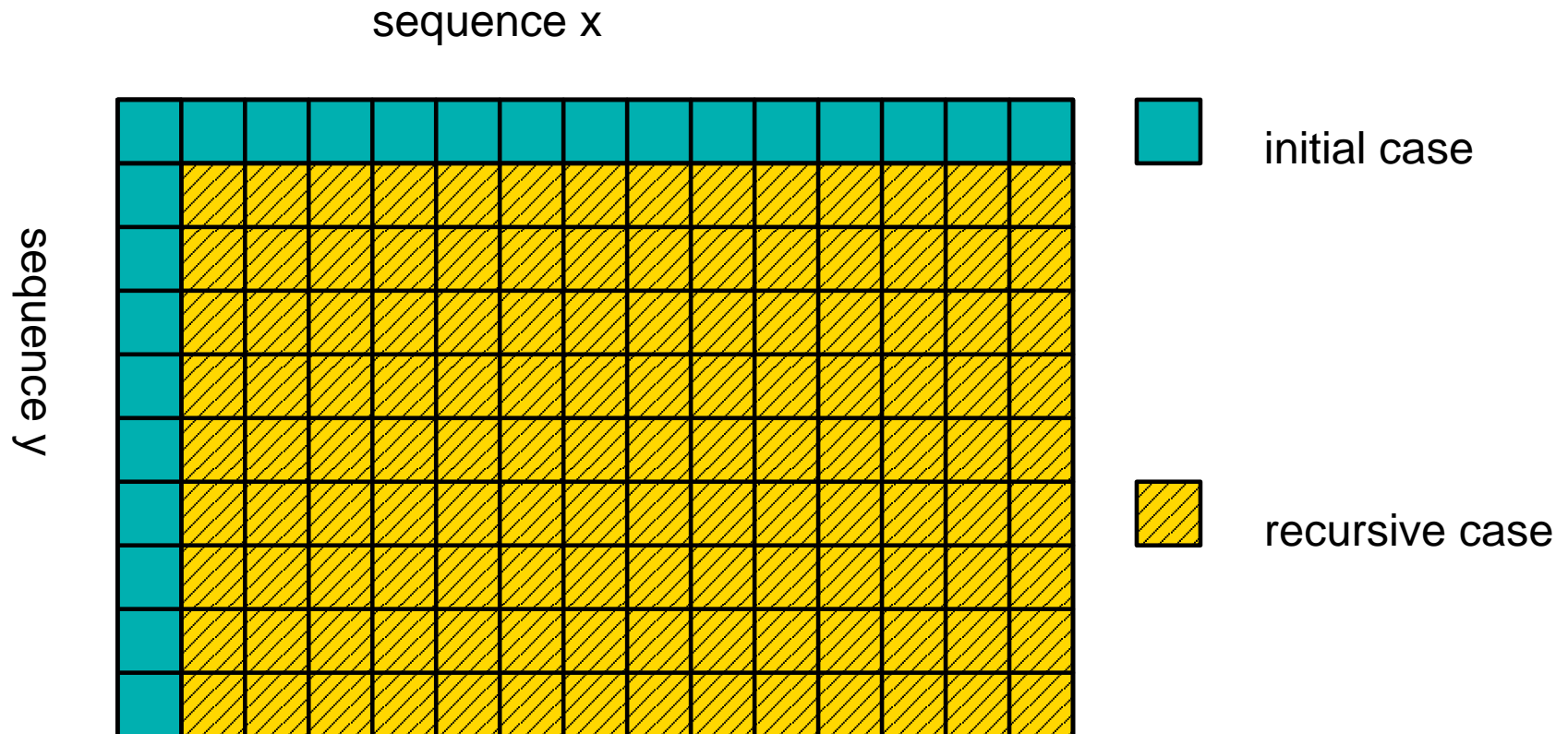


## Computing the score in linear space

A basic observation is that the *score* computation can be done in  $O(m)$  space because computing a column  $C_j$  only requires knowledge of the previous column  $C_{j-1}$ . Hence we can scan the text from left to right updating our *column vector* and reporting a match every time  $C[m, j] \leq k$ .

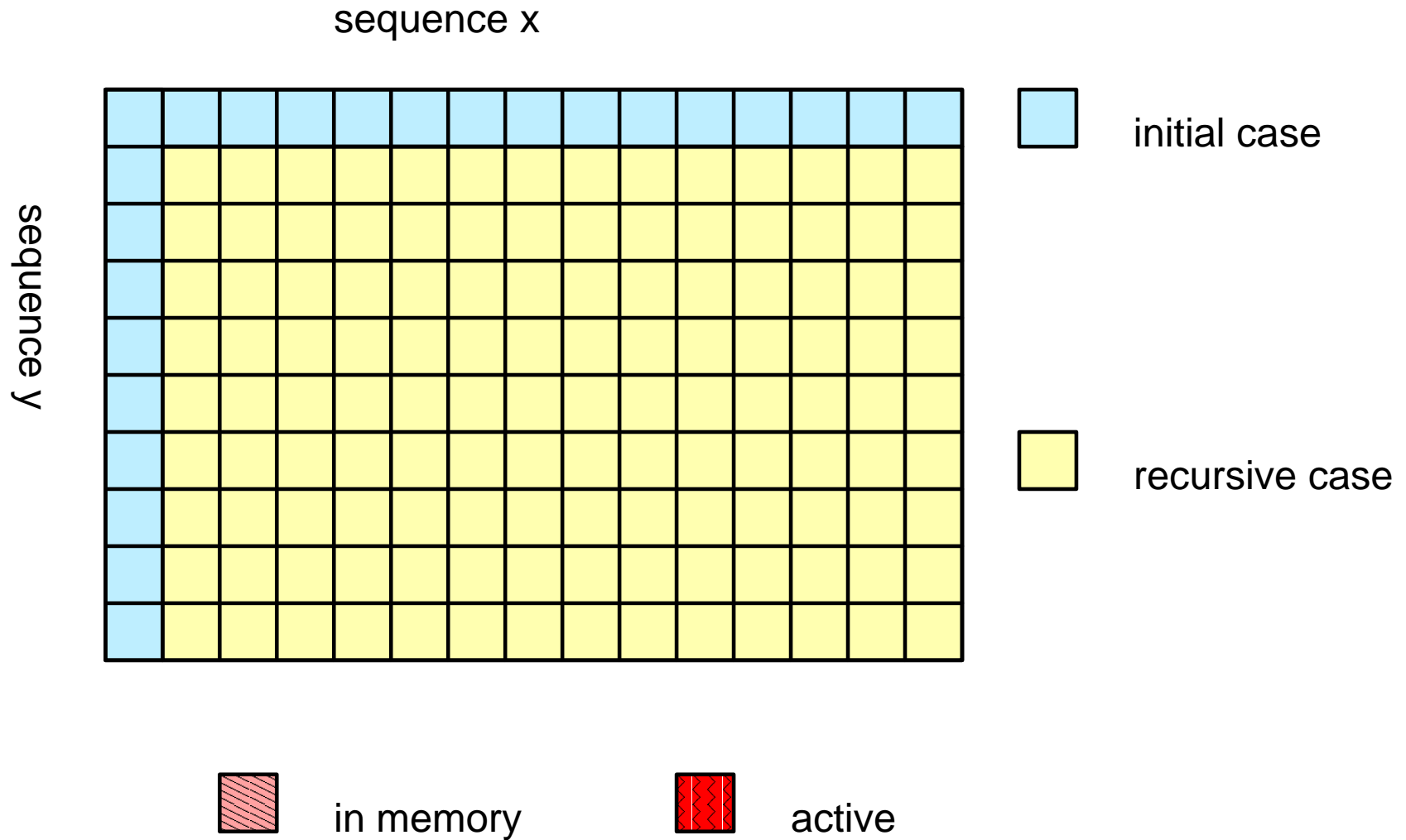
# Computing the score in linear space

(2)



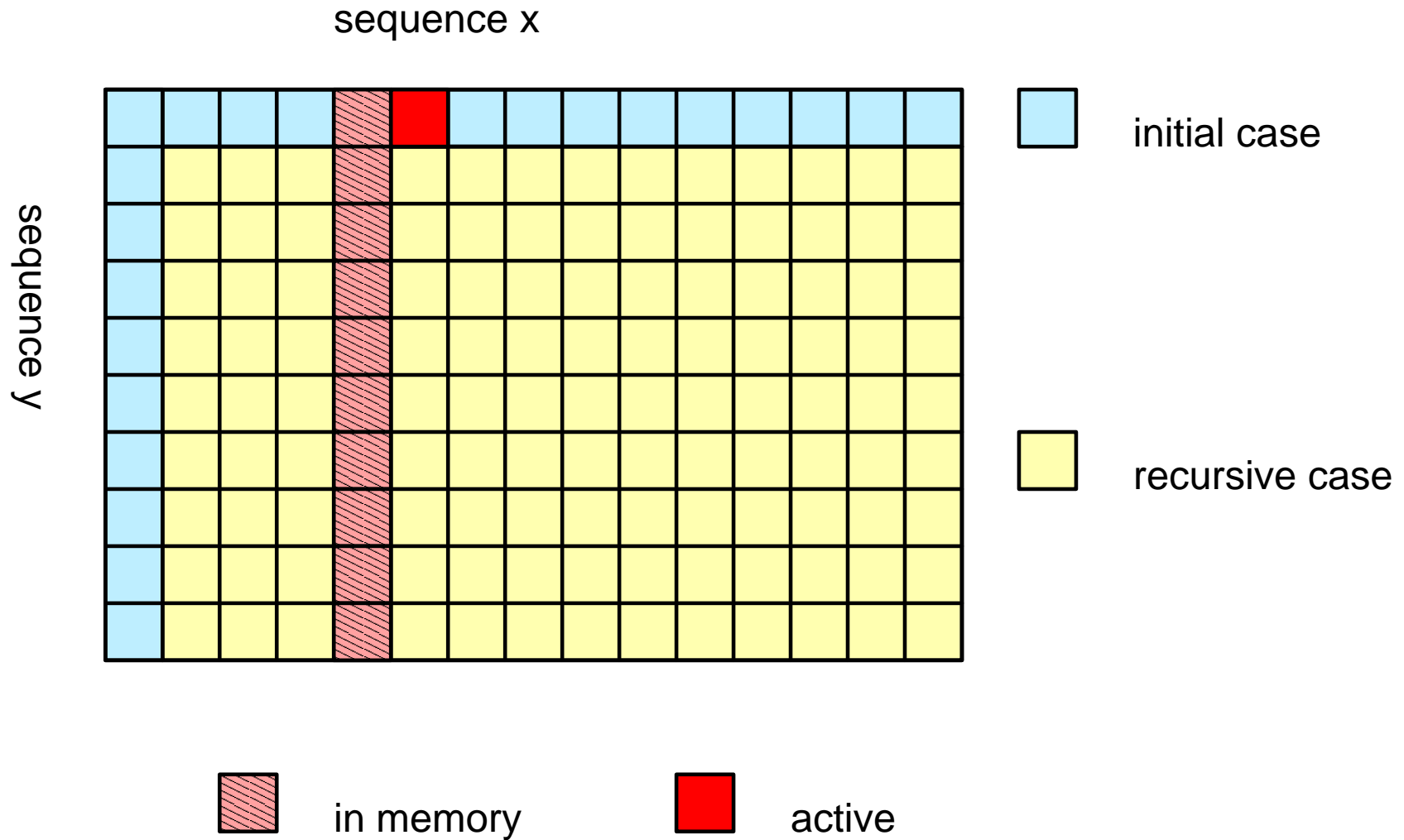
# Computing the score in linear space

(3)



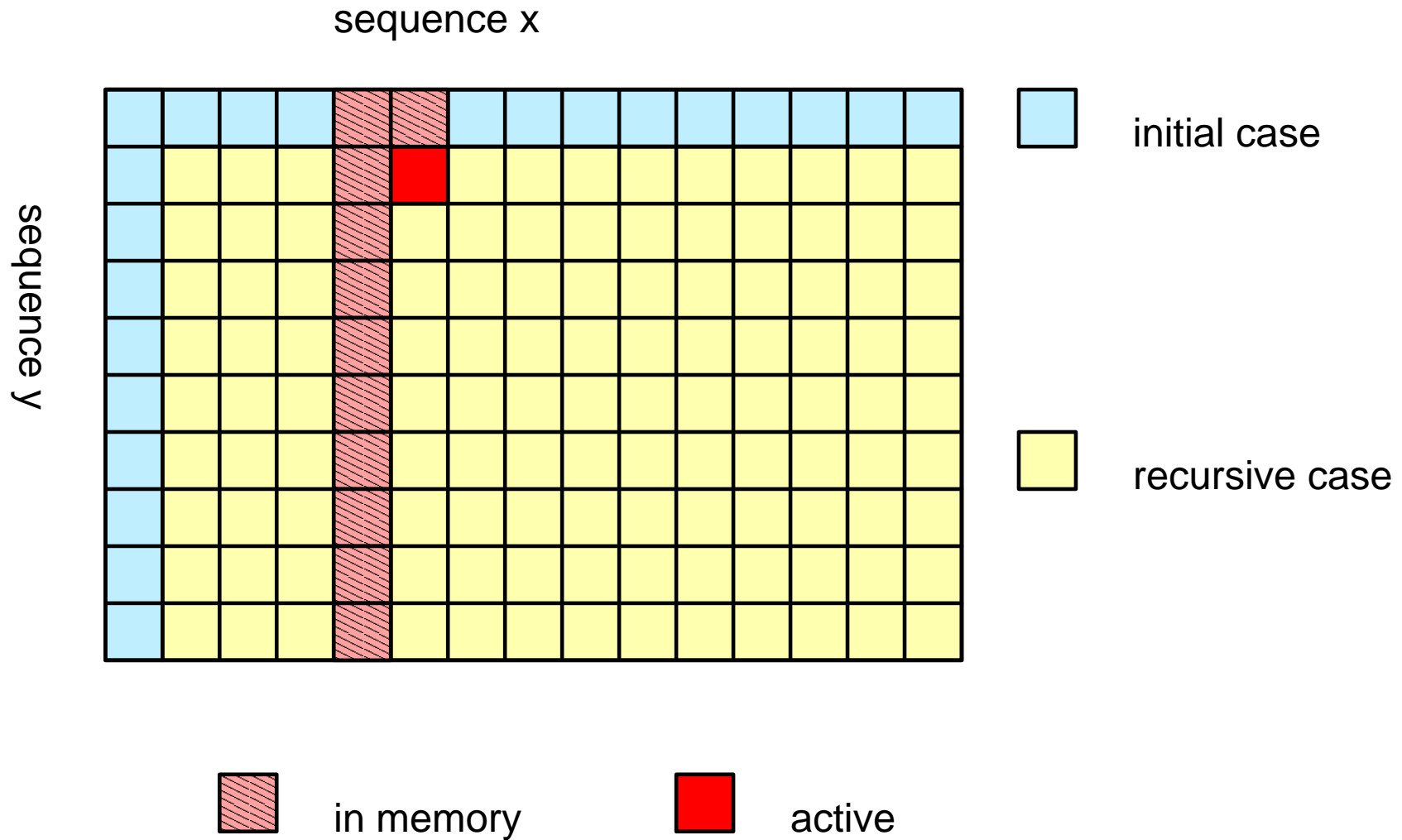
# Computing the score in linear space

(4)



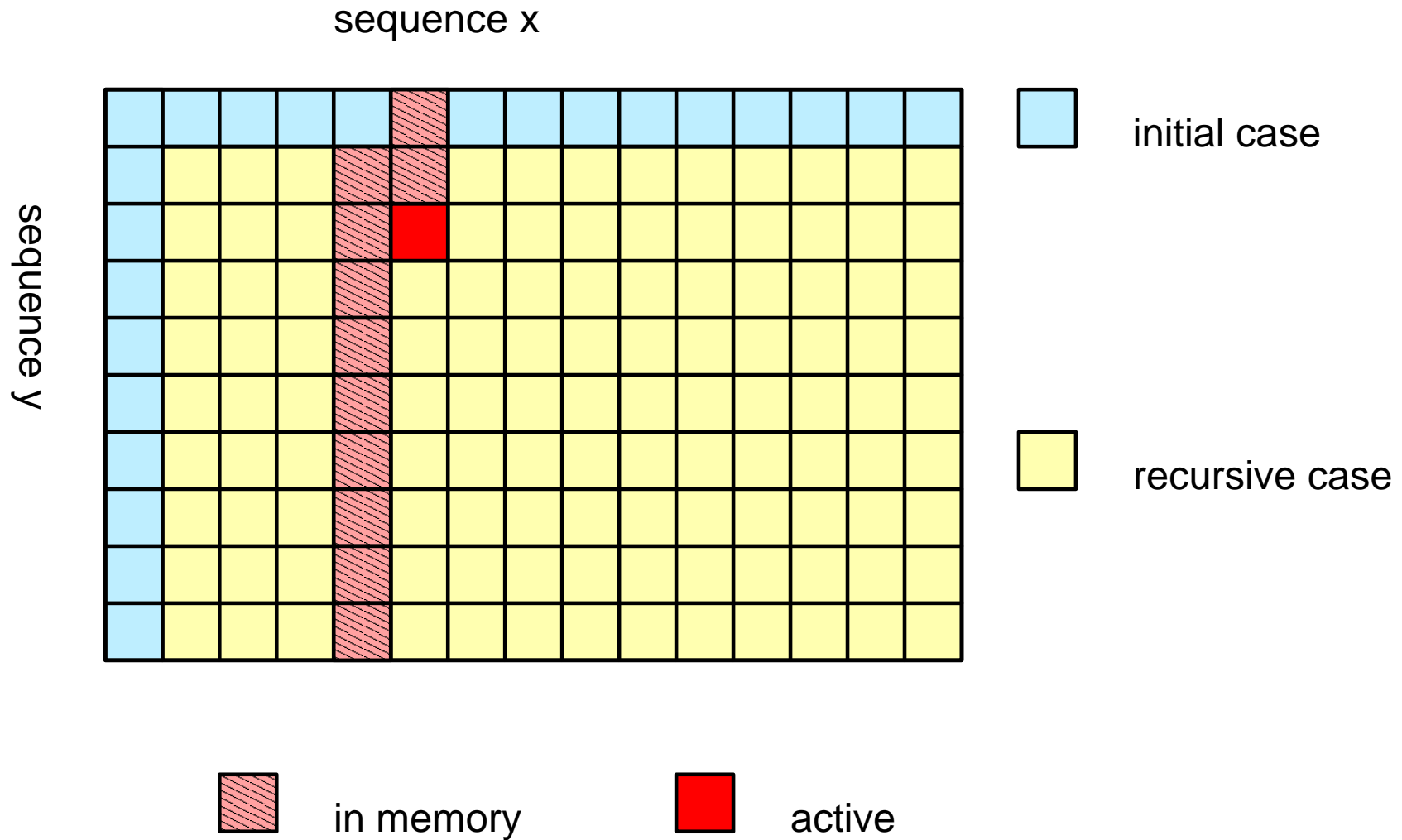
# Computing the score in linear space

(5)



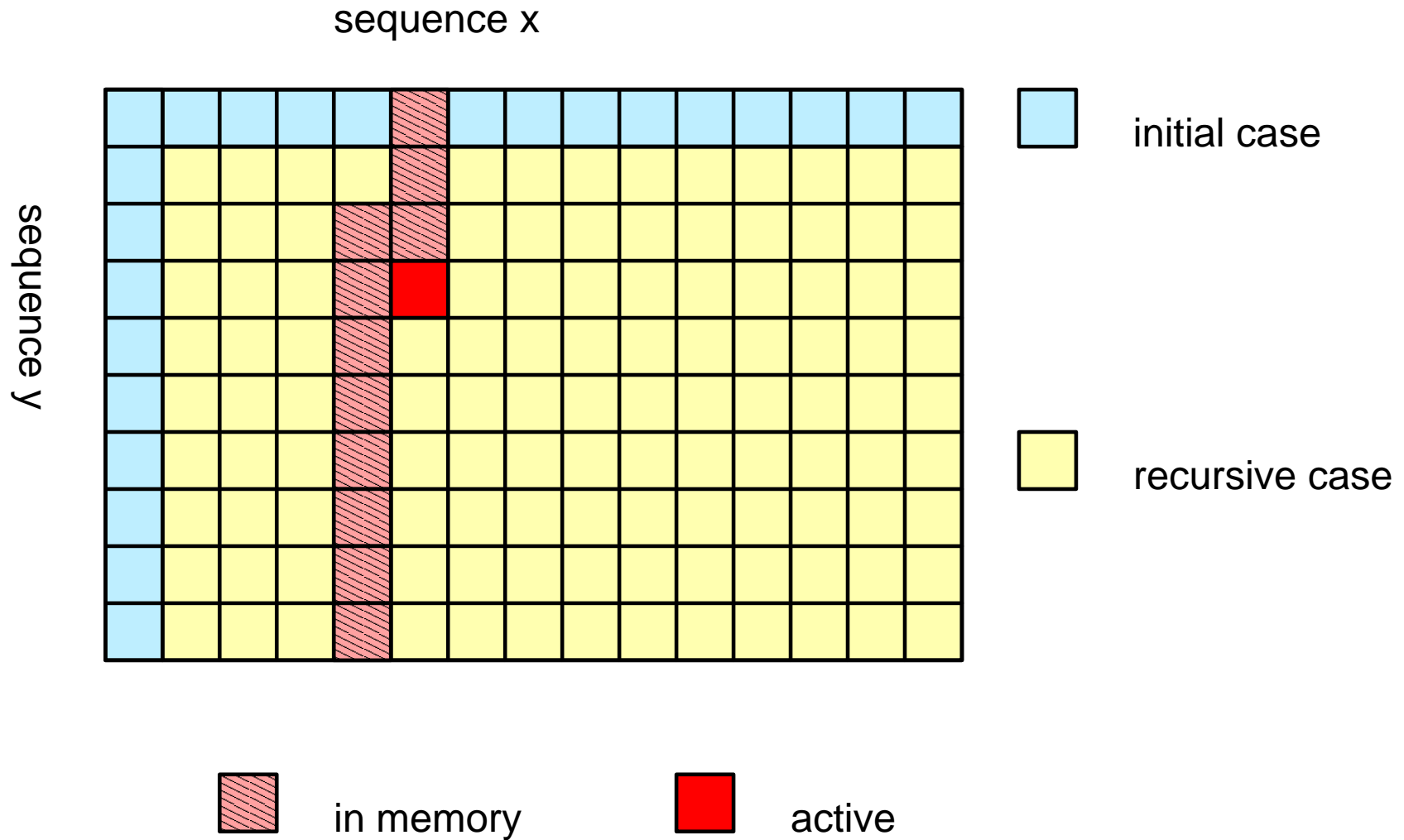
# Computing the score in linear space

(6)



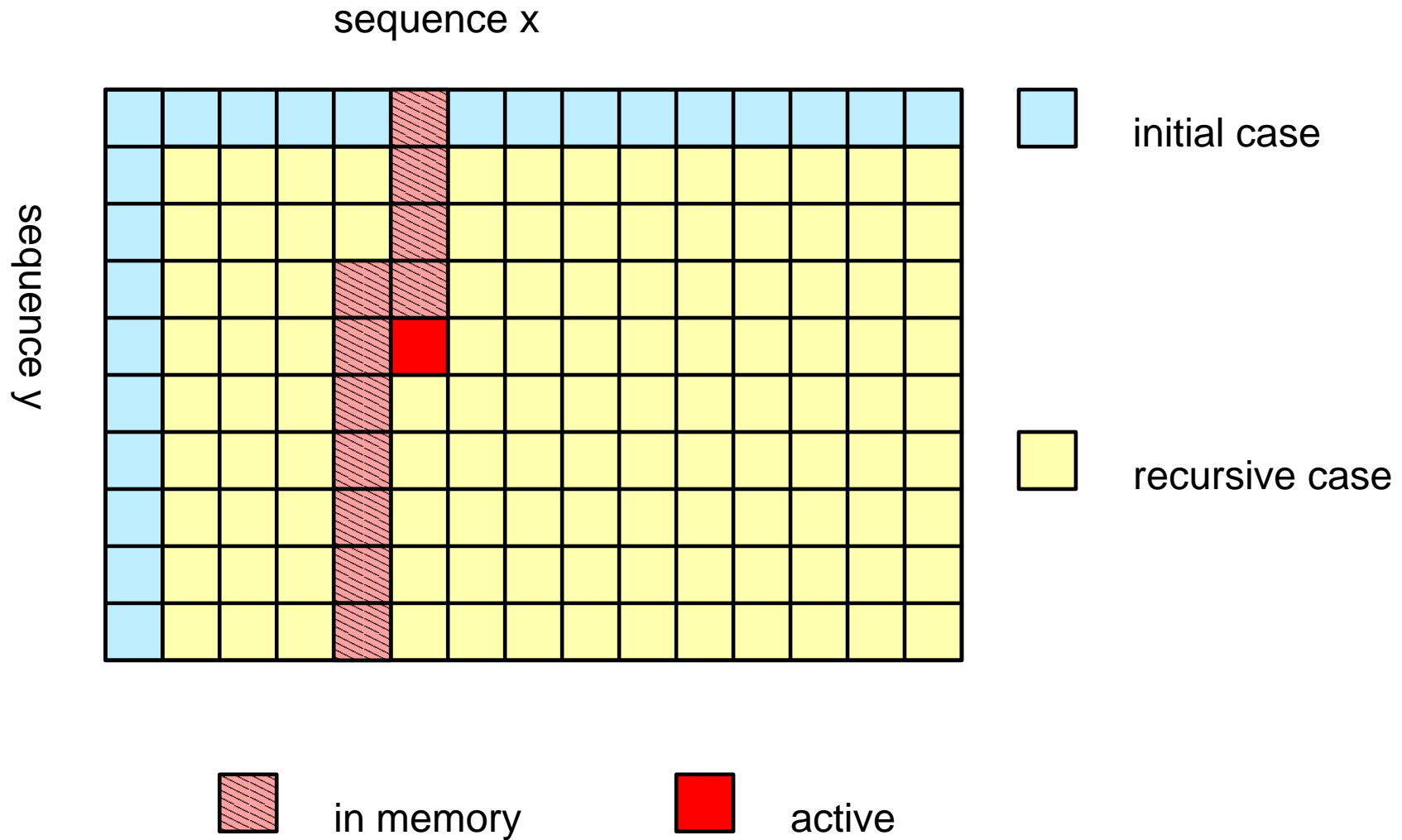
# Computing the score in linear space

(7)



# Computing the score in linear space

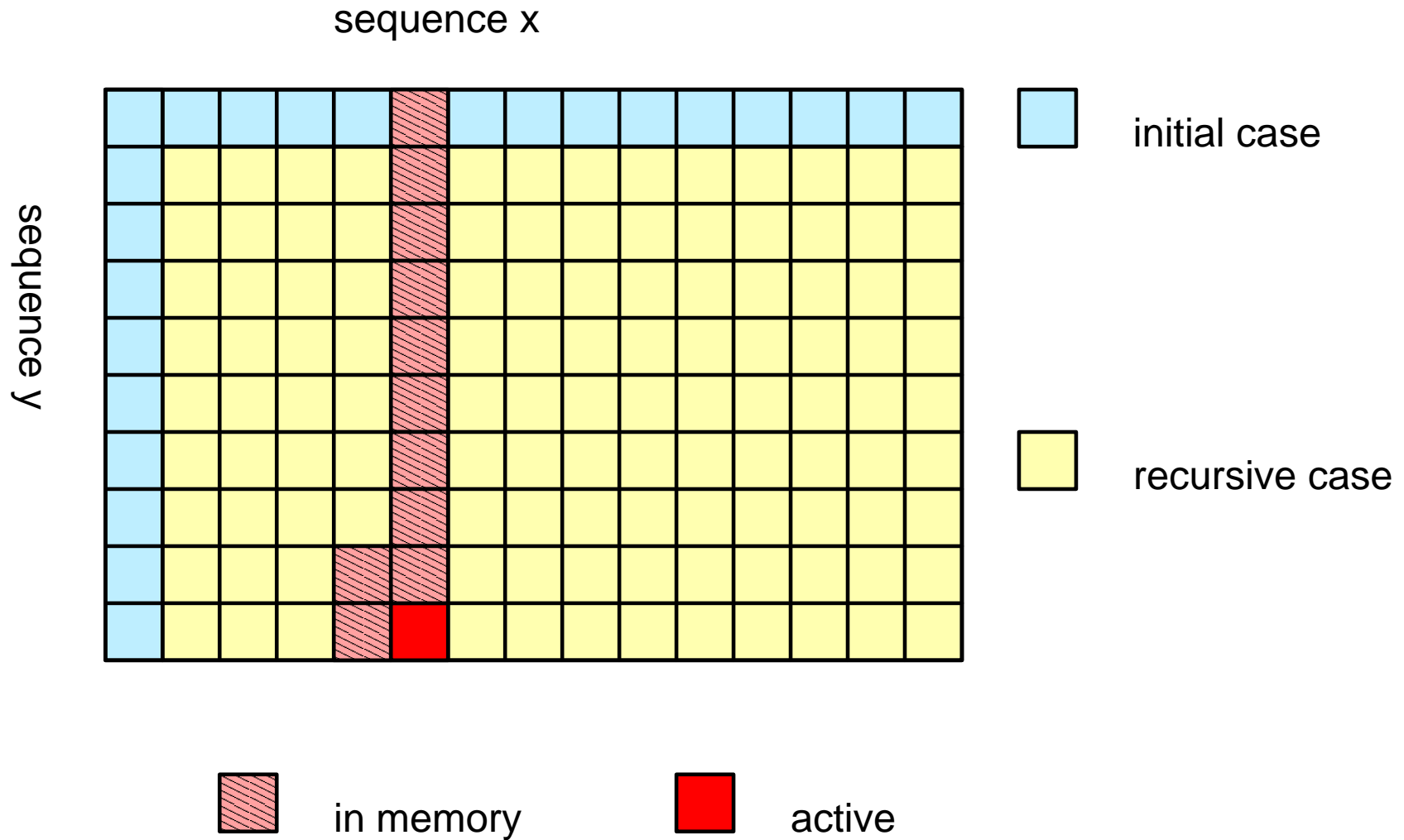
(8)





# Computing the score in linear space

(9)



## Computing the *score* in linear space (10)

The approximate matches of the pattern can be output on-the-fly when the final value of the column vector is (re-)calculated.

The algorithm still requires in  $O(mn)$  time to run, but it uses only  $O(m)$  memory.

Note that in order to obtain the actual *alignment* (not just its score), we need to trace back by some other means from the point where the match was found, as the preceding columns of the DP matrix are no longer available.

# Ukkonen's algorithm

# Ukkonen's algorithm

Ukkonen studied the properties of the dynamic programming matrix and came up with a simple twist to the classical algorithm that retains all of its flexibility while reducing the running time to  $O(kn)$  (as opposed to  $O(mn)$ ) on average.

The idea is that since the pattern will normally not match the text, the entries of each column read from top to bottom will quickly reach  $k + 1$ . Let us call an entry of the DP matrix *active* if its value is at most  $k$ . Ukkonen's algorithm maintains an index *lact* pointing to the *last active* cell and updates it accordingly. Due to the properties of *lact* it can avoid working on subsequent cells.

In the exercises you will prove that the value of *lact* can *decrease* in one iteration by more than one, but it can never *increase* by more than one.

# Pseudocode of Ukkonen's algorithm

```

1 // Preprocessing
2 for  $i \in 0 \dots m$  do  $C_i = i$ ; od
3  $lact = k + 1$ ;
4 // Searching
5 for  $pos \in 1 \dots n$  do
6    $C_p = 0$ ;  $C_n = 0$ ;
7   for  $i \in 1 \dots lact$  do
8     if  $p_i = t_{pos}$ 
9       then  $C_n = C_p$ ;
10    else
11      if  $C_p < C_n$  then  $C_n = C_p$ ; fi
12      if  $C_i < C_n$  then  $C_n = C_i$ ; fi
13       $C_n++$ ;
14    fi
15     $C_p = C_i$ ;  $C_i = C_n$ ;
16  od
17 // Updating lact
18 while  $C_{lact} > k$  do  $lact--$ ; od
19 if  $lact = m$  then report occurrence
20   else  $lact++$ ;
21 fi
22 od

```

	$pos - 1$	$pos$
$i - 1$	$C_p$	[ <i>was:</i> $C_n$ ]
$i$	$C_i$	$C_n$

## Running time of Ukkonen's algorithm

Ukkonen's algorithm behaves like a standard dynamic programming algorithm, except that it maintains and uses in the main loop the variable *lact*, the last active cell.

The value of *lact* can *decrease* in one iteration by more than one, but it can never *increase* more than one (why?). Thus the total time over the run of the algorithm spent for updating *lact* is  $O(n)$ . In other words, *lact* is maintained in amortized constant time per column.

One can show that on average the value of *lact* is bounded by  $O(k)$ . Thus Ukkonen's modification of the classical DP algorithm has an average running time of  $O(kn)$ .

## **Myers' bit vector algorithm**

# Encoding and parallelizing the DP matrix

Next we will look at Myers' bit-parallel algorithm which – in combination with Ukkonen's trick – yields a remarkably fast algorithm, which was used e. g. for the overlap computations performed at Celera as the starting point for genome assembly.

For simplicity, we assume that  $m$  is smaller than  $w$  the length of a machine word. It will be an exercise to extend the algorithm to handle  $m \geq w$ .

The main idea of following the bit-vector algorithm is to parallelize the dynamic programming matrix. We will compute the column as a whole in a series of bit-level operations. In order to do so, we need to

1. encode the dynamic programming matrix using bit vectors, and
2. resolve the dependencies (especially *within* the columns).



## Encoding the DP matrix

The binary encoding is done by considering the *differences* between consecutive rows and columns instead of their *absolute* values. We introduce the following nomenclature for these differences (“deltas”):

$$\text{horizontal adjacency property } \Delta h_{i,j} = C_{i,j} - C_{i,j-1} \in \{-1, 0, +1\}$$

$$\text{vertical adjacency property } \Delta v_{i,j} = C_{i,j} - C_{i-1,j} \in \{-1, 0, +1\}$$

$$\text{diagonal property } \Delta d_{i,j} = C_{i,j} - C_{i-1,j-1} \in \{0, +1\}$$

**Exercise.** *Prove that these deltas are indeed within the claimed ranges.*

## Encoding the DP matrix (2)

horizontal adjacency property  $\Delta h_{i,j} = C_{i,j} - C_{i,j-1} \in \{-1, 0, +1\}$

vertical adjacency property  $\Delta v_{i,j} = C_{i,j} - C_{i-1,j} \in \{-1, 0, +1\}$

diagonal property  $\Delta d_{i,j} = C_{i,j} - C_{i-1,j-1} \in \{0, +1\}$

The delta vectors are encoded as bit-vectors by the following boolean variables:

■  $VP_{ij} \equiv (\Delta v_{i,j} = +1)$ , the vertical positive delta vector

■  $VN_{ij} \equiv (\Delta v_{i,j} = -1)$ , the vertical negative delta vector

■  $HP_{ij} \equiv (\Delta h_{i,j} = +1)$ , the horizontal positive delta vector

■  $HN_{ij} \equiv (\Delta h_{i,j} = -1)$ , the horizontal negative delta vector

■  $D0_{ij} \equiv (\Delta d_{i,j} = 0)$ , the diagonal zero delta vector

## Encoding the DP matrix (3)

The deltas and bits are defined such that

$$\Delta v_{i,j} = VP_{i,j} - VN_{i,j}$$

$$\Delta h_{i,j} = HP_{i,j} - HN_{i,j}$$

$$\Delta d_{i,j} = 1 - D0_{i,j}.$$

It is also clear that these values “encode” the entire DP matrix  $C[0..m, 0..n]$  by  $C(i, j) = \sum_{r=1}^i \Delta v_{r,j}$ . Below is our example matrix with  $\Delta v_{i,j}$  values.

$\Delta v_{i,j}$ :

		A	N	N	E	A	L	I	N	G
	0	0	0	0	0	0	0	0	0	0
A	1	0	1	1	1	0	1	1	1	1
N	1	1	-1	0	1	1	0	1	1	1
N	1	1	1	-1	-1	1	1	0	0	0
U	1	1	1	1	0	0	1	1	1	1
A	1	1	1	1	1	-1	-1	0	1	1
L	1	1	1	1	1	1	-1	-1	-1	0

## Encoding the DP matrix (4)

We denote by  $score_j$  the edit distance of a pattern occurrence ending at text position  $j$ . The key ideas of Myers' algorithm are as follows:

1. Instead of computing  $C$  we compute the  $\Delta$  values, which in turn are represented as bit-vectors.
2. We compute the matrix column by column as in Ukkonen's version of the DP algorithm.
3. We maintain the value  $score_j$  using the fact that  $score_0 = m$  and  $score_j = score_{j-1} + \Delta h_{m,j}$ .

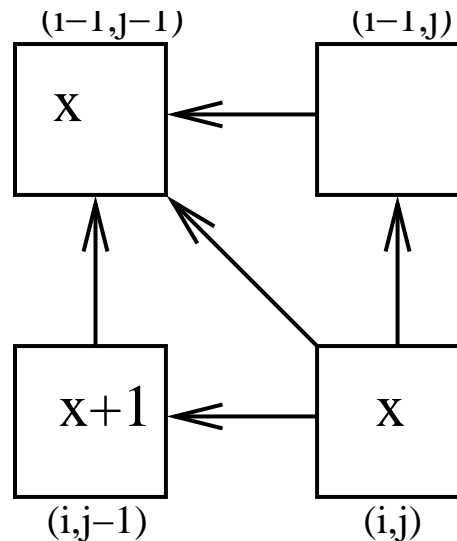
In the following slides we will make some observations about the dependencies of the bit-vectors.

# Observations

Lets have a look at the  $\Delta$ s. The first observation is that

$$HN_{i,j} \Leftrightarrow VP_{i,j-1} \text{ AND } D0_{i,j}.$$

**Proof.** If  $HN_{i,j}$  then  $\Delta h_{i,j} = -1$  by definition, and hence  $\Delta v_{i,j-1} = 1$  and  $\Delta d_{i,j} = 0$ . This holds true, because otherwise the range of possible values ( $\{-1, 0, +1\}$ ) would be violated.



## Observations (2)

By symmetry we have:

$$VN_{i,j} \Leftrightarrow HP_{i-1,j} \text{ AND } D0_{i,j} .$$

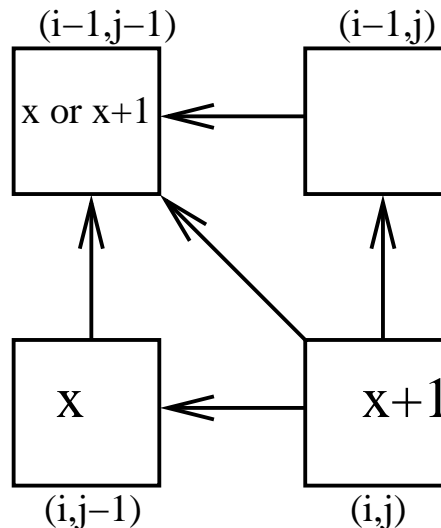
(( Last observation was:  $HN_{i,j} \Leftrightarrow VP_{i,j-1} \text{ AND } D0_{i,j}$  ))

## Observations (3)

The next observation is that

$$HP_{i,j} \Leftrightarrow VN_{i,j-1} \text{ OR NOT } (VP_{i,j-1} \text{ OR } D0_{i,j}) .$$

**Proof.** If  $HP_{i,j}$  then  $VP_{i,j-1}$  cannot hold without violating the ranges. Hence  $\Delta v_{i,j-1}$  is  $-1$  or  $0$ . In the first case  $VN_{i,j-1}$  is true, whereas in the second case we have neither  $VP_{i,j-1}$  nor  $D0_{i,j}$



## Observations (4)

Again by symmetry we have

$$VP_{i,j} \Leftrightarrow HN_{i-1,j} \text{ OR NOT } (HP_{i-1,j} \text{ OR } D0_{i,j}) .$$

(( Last observation was:  $HP_{i,j} \Leftrightarrow VN_{i,j-1} \text{ OR NOT } (VP_{i,j-1} \text{ OR } D0_{i,j})$  ))



## Observations (5)

Finally,  $D0_{i,j} = 1$  iff  $C[i, j]$  and  $C[i-1, j-1]$  have the same value. This can be true for three possible reasons, which correspond to the three cases of the DP recurrence:

1.  $p_i = t_j$ , that is the query at position  $i$  matches the text at position  $j$ .
2.  $C[i, j]$  is obtained by propagating a lower value from the left, that is  $C[i, j] = 1 + C[i, j-1]$ . Then we have  $VN_{i,j-1}$ .
3.  $C[i, j]$  is obtained by propagating a lower value from above, that is  $C[i, j] = 1 + C[i-1, j]$ . Then we have  $HN_{i-1,j}$ .

So writing this together yields:

$$D0_{i,j} \Leftrightarrow (p_i = t_j) \text{ OR } VN_{i,j-1} \text{ OR } HN_{i-1,j}.$$

## Observations (6)

Taking everything together we have the following equivalences:

$$HN_{i,j} \Leftrightarrow VP_{i,j-1} \text{ AND } D0_{i,j} \quad (1)$$

$$VN_{i,j} \Leftrightarrow HP_{i-1,j} \text{ AND } D0_{i,j} \quad (2)$$

$$HP_{i,j} \Leftrightarrow VN_{i,j-1} \text{ OR NOT } (VP_{i,j-1} \text{ OR } D0_{i,j}) \quad (3)$$

$$VP_{i,j} \Leftrightarrow HN_{i-1,j} \text{ OR NOT } (HP_{i-1,j} \text{ OR } D0_{i,j}) \quad (4)$$

$$D0_{i,j} \Leftrightarrow (p_i = t_j) \text{ OR } VN_{i,j-1} \text{ OR } HN_{i-1,j} \quad (5)$$

Can these be used to update the five bit-vectors as the algorithms searches through the text?

## Resolving circular dependencies

In principle the overall strategy is clear: We traverse the text from left to right and keep track of five bit-vectors  $VN_j$ ,  $VP_j$ ,  $HN_j$ ,  $HP_j$ ,  $D0_j$ , each containing the bits for  $1 \leq i \leq m$ . For moderately sized patterns ( $m \leq w$ ) these fit into a machine word. The reduction to linear space works similar as for  $C$ .

It is clear that we initialize  $VP_0 = 1^m$ ,  $VN_0 = 0^m$ , so perhaps we can compute the other bit-vectors using the above observations.

There remains a problem to be solved:  $D0_{i,j}$  depends on  $HN_{i-1,j}$  which in turn depends on  $D0_{i-1,j}$  – a value we have not computed yet. But there is a solution.

Let us have a closer look at  $D0_{i,j}$  and expand it.

$$\begin{aligned} D0_{i,j} &= (p_i = t_j) \text{ OR } VN_{i,j-1} \text{ OR } \underline{HN_{i-1,j}} \\ HN_{i-1,j} &= VP_{i-1,j-1} \text{ AND } D0_{i-1,j} \\ \Rightarrow \quad D0_{i,j} &= (p_i = t_j) \text{ OR } VN_{i,j-1} \text{ OR } (VP_{i-1,j-1} \text{ AND } D0_{i-1,j}) \end{aligned}$$

# Resolving (not so) circular dependencies

(2)

The formula

$$D0_{i,j} = \underline{(p_i = t_j) \text{ OR } VN_{i,j-1}} \text{ OR } (\underline{VP_{i-1,j-1}} \text{ AND } D0_{i-1,j})$$

is of the form

$$D0_i = X_i \text{ OR } (Y_{i-1} \text{ AND } D0_{i-1}),$$

where

$$\begin{aligned} X_i &:= (t_i = p_i) \text{ OR } VN_i \\ \text{and } Y_i &:= VP_i. \end{aligned}$$

Here we omitted the index  $j$  from notation for clarity. We solve for  $D_i$ . Unrolling now the first few terms we get:

$$D0_1 = X_1,$$

$$D0_2 = X_2 \text{ OR } (X_1 \text{ AND } Y_1),$$

$$D0_3 = X_3 \text{ OR } (X_2 \text{ AND } Y_2) \text{ OR } (X_1 \text{ AND } Y_1 \text{ AND } Y_2).$$

In general we have:

$$D0_i = \text{OR}_{r=1}^i (X_r \text{ AND } Y_r \text{ AND } Y_{r+1} \text{ AND } \dots \text{ AND } Y_{i-1}).$$

## Resolving (not so) circular dependencies

(3)

$$D0_i = \text{OR}_{r=1}^i (X_r \text{ AND } Y_r \text{ AND } Y_{r+1} \text{ AND } \dots \text{ AND } Y_{i-1}).$$

To put this in words, let  $s < i$  be such that  $Y_s \dots Y_{i-1} = 1$  and  $Y_{s-1} = 0$ . Then  $D0_i = 1$  if  $X_r = 1$  for some  $s \leq r \leq i$ . That is, the first consecutive block of 1s in  $Y$  that is right of  $i$  must be covered by a 1 in  $X$ . Here is an example:

$$Y = 00011111000011$$

$$X = 00001010000101$$

$$D0 = 001111110000111$$

A position  $i$  in  $D0$  is set to 1 if one position to the right in  $Y$  is a 1 and in the consecutive runs of 1s from there on is also an  $X_r$  set to 1.

## Computing $D0$

Now we solve the problem to compute  $D0$  from  $X$  and  $Y$  using bit-vector operations.

The first step is to compute the  $\&$  of  $X$  and  $Y$  and add  $Y$ . The result will be that every  $X_r = 1$  that is aligned to a  $Y_r = 1$  will be propagated by the addition one position to the left of a block of 1s.

Again our example:

$$\begin{aligned} Y &= 00011111000011 \\ X &= 00001010000101 \\ X \& Y &= 00001010000001 \\ (X \& Y) + Y &= 00\underline{1}01001000\underline{1}00 \\ &\dots \\ D0 &= 001111110000111 \end{aligned}$$

Note the two 1s that got propagated to the end of the runs of 1s in  $Y$ .

However, note also the one 1 that is not in the solution but was introduced by adding a 1 in  $Y$  to a 0 in  $(X \& Y)$ .

## Computing $D0$ (2)

As a remedy we *XOR* the term with  $Y$  so only the bits that changed during the propagation stay turned on. Again our example:

$$Y = 00011111000011$$

$$X = 00001010000101$$

$$X \& Y = 00001010000001$$

$$(X \& Y) + Y = 00101001000100$$

$$(X \& Y) + Y \wedge Y = 00110110000111$$

...

$$D0 = 00111110000111$$

Now we are almost done. The only thing left to fix is the observation that there may be several  $X_r$  bits under the same block of  $Y$ . Of those all but the first remain unchanged and hence will not be marked by the *XOR*.

## Computing $D0$ (3)

To fix this and to account for the case  $X_1 = 1$  we *OR* the final result with  $X$ . Again our example:

$$Y = 00011111000011$$

$$X = 00001010000101$$

$$X \& Y = 00001010000001$$

$$(X \& Y) + Y = 00101001000100$$

$$((X \& Y) + Y) \wedge Y = 00110110000111$$

$$(((X \& Y) + Y) \wedge Y) \mid X = 00111110000111$$

$$= D0 = 00111110000111$$

Now we can substitute  $X$  back by  $(p_i = t_j)$  OR  $VN$  and  $Y$  by  $VP$ .



## Preprocessing the alphabet

The only thing left now in the computation of  $D0$  is the expression  $p_i = t_j$ .

Of course we cannot check for all  $i$  whether  $p_i$  is equal  $t_j$  for all  $i$ . This would take time  $O(m)$  and defeat our goal. However, we can preprocess the query pattern and the alphabet  $\Sigma$ . We use  $|\Sigma|$  many bit-vectors  $B[\alpha] \mid \forall \alpha \in \Sigma$  with the property that  $B[\alpha]_i = 1$  if  $p_i = \alpha$ . These vectors can easily be precomputed in time  $O(|\Sigma| m)$ .

Now we have everything together.

# Myers' bit-vector algorithm

---

```
1 // Preprocessing
2 for  $c \in \Sigma$  do  $B[c] = 0^m$  od
3 for  $j \in 1 \dots m$  do  $B[p_j] = B[p_j] \mid 0^{m-j}10^{j-1}$  od
4  $VP = 1^m$ ;  $VN = 0^m$ ;
5  $score = m$ ;
```

---

# Myers' bit-vector algorithm

(2)

```
1 // Searching
2 for  $pos \in 1 \dots n$  do
3    $X = B[t_{pos}] \mid VN$ ;
4    $D0 = ((VP + (X \& VP)) \wedge VP) \mid X$ ;
5    $HN = VP \& D0$ ;
6    $HP = VN \mid \sim (VP \mid D0)$ ;
7    $X = HP \ll 1$ ;
8    $VN = X \& D0$ ;
9    $VP = (HN \ll 1) \mid \sim (X \mid D0)$ ;
10 // Scoring and output
11 if  $HP \& 10^{m-1} \neq 0^m$ 
12   then  $score += 1$ ;
13   else if  $HN \& 10^{m-1} \neq 0^m$ 
14     then  $score -= 1$ ;
15   fi
16 fi
17 if  $score \leq k$  report occurrence at  $pos$  fi;
18 od
```

## Myers' bit-vector algorithm (3)

Note that this algorithm easily computes the *edit distance* if we add in line 7 the following code fragment:  $| 0^{m-1} 1$ . This because in this case there is a horizontal increment in the 0-th row. In the case of reporting all occurrences each column starts with 0.

## The example

If we try to find `annual` in `annealing` we first run the preprocessing resulting in:

a	0	1	0	0	0	1
l	1	0	0	0	0	0
n	0	0	0	1	1	0
u	0	0	1	0	0	0
*	0	0	0	0	0	0

in addition

$VN = 000000$

$VP = 111111$

*score*      6

## The example

(2)

We read the first letter of the text, an **a**.

*Reading*   *a*   010001

*D0*   =   111111

*HN*   =   111111

*HP*   =   000000

*VN*   =   000000

*VP*   =   111110

*score*   5

## The example

(3)

We read the second letter of the text, an **n**.

*Reading*    *n*    000110

*D0*    =    111110

*HN*    =    111110

*HP*    =    000001

*VN*    =    000010

*VP*    =    111101

*score*        4

## The example

(4)

We read the third letter of the text, an [n](#).

*Reading*    *n*    000110

*D0*    =    111110

*HN*    =    111100

*HP*    =    000010

*VN*    =    000100

*VP*    =    111001

*score*        3



## The example

(5)

We read the fourth letter of the text, an *e*.

*Reading*   *e*   000000

*D0*   =   000100

*HN*   =   000000

*HP*   =   000110

*VN*   =   000100

*VP*   =   110001

*score*   3

## The example

(6)

We read the fifth letter of the text an **a**.

*Reading*   *e*   010001

*D0*   =   110111

*HN*   =   110011

*HP*   =   001100

*VN*   =   010000

*VP*   =   100110

*score*     2

and so on..... In the exercises you will extend the algorithm to handle queries larger than  $w$ .

## Summary

- Approximate string matching is an important topic in computational molecular biology.
- Comparing high fidelity DNA or Protein strings does not require amino acids substitution matrices. Edit distance is usually sufficient.
- The basic DP can be greatly improved by using bit parallelism.
- The proposed bit-vector algorithm encodes the differences in the DP matrix for the edit distance.
- The bit-vectors represent the three possible values of differences ( $-1, 0, 1$ )
- The bit-vector algorithm can easily be expanded to handle patterns that are longer than the word length and yields together with the Ukkonen trick a very fast algorithm for DP based alignment.