Algorithms in bioinformatics

Anton Bankevich
Center for Algorithmic Biotechnology
SPbSU

Bioinformatics vs Algorithmic Biology

Algorithmic biology:

development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Bioinformatics:

field that develops methods and software tools for understanding biological data

Bioinformatics vs Algorithmic Biology

Algorithmic biology:

Bioinformatics:

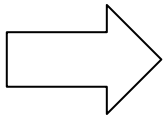
Bioinformatics vs Algorithmic Biology

Algorithmic biology:

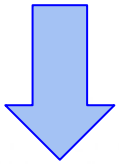
Bioinformatics:

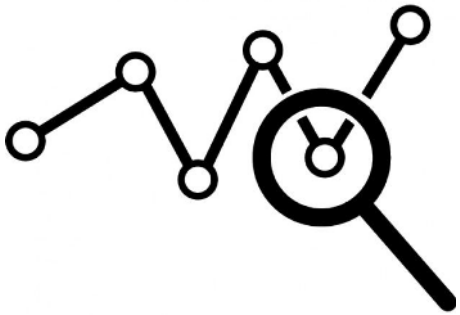
Discovery

Technology

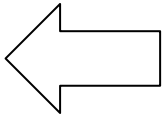


Data





Analysis



Discovery

Data

What is biological data?

- Sequencing reads
- Mass spectrometry
- Expression measurements

Questions about sequences

- Are they the same?
- Is one of them a part of another?
- What is common between sequences?

Questions about sequences

- Are they similar?
- Is one of them similar to a part of another?
- What is similar in the sequences?

Alignment

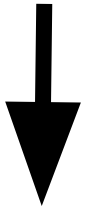
AACGCTAACGGTAA

AACCGCGAACTAA

Alignment

AACGCTAACGGTAA

AACCGCGAACTAA



AAC-GCTAACGGTAA

AACCGCGAAC--TAA

Needleman–Wunsch algorithm

A C G T T A G

A C C T - A G

Score(x, x) = 1 (match)

Score(x, y) = -1 (mismatch)

Score(x, -) = Score(-, x) = -1 (Indel)

Score(ACGTTAG, ACCTAG) = 3

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
A								
C								
C								
T								
A								
G								

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1							
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1							
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1							
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1						
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1						
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1/-2						
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1/-2						
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1/-2/-2						
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1						
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1						
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	-2					
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	-2/0					
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	-2/0/-3					
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0					
C	-2							
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
C	-2	0						
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
C	-2	0						
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
C	-2	0	2					
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
C	-2	0	2/-1					
C	-3							
T	-4							
A	-5							
G	-6							

Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
C	-2	0	2/-1/-1					
C	-3							
T	-4							
A	-5							
G	-6							

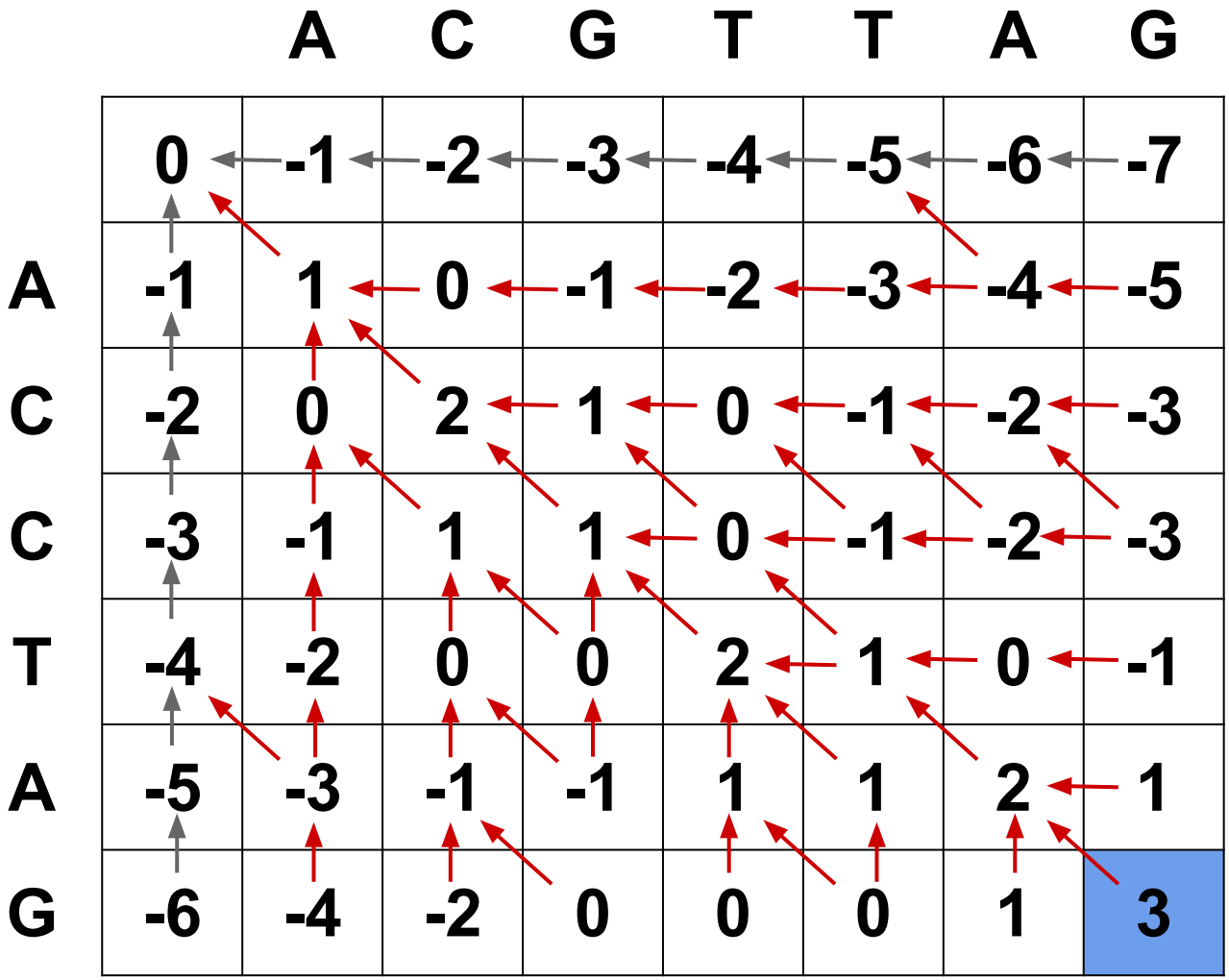
Needleman–Wunsch algorithm

		A	C	G	T	T	A	G
	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
C	-2	0	2					
C	-3							
T	-4							
A	-5							
G	-6							

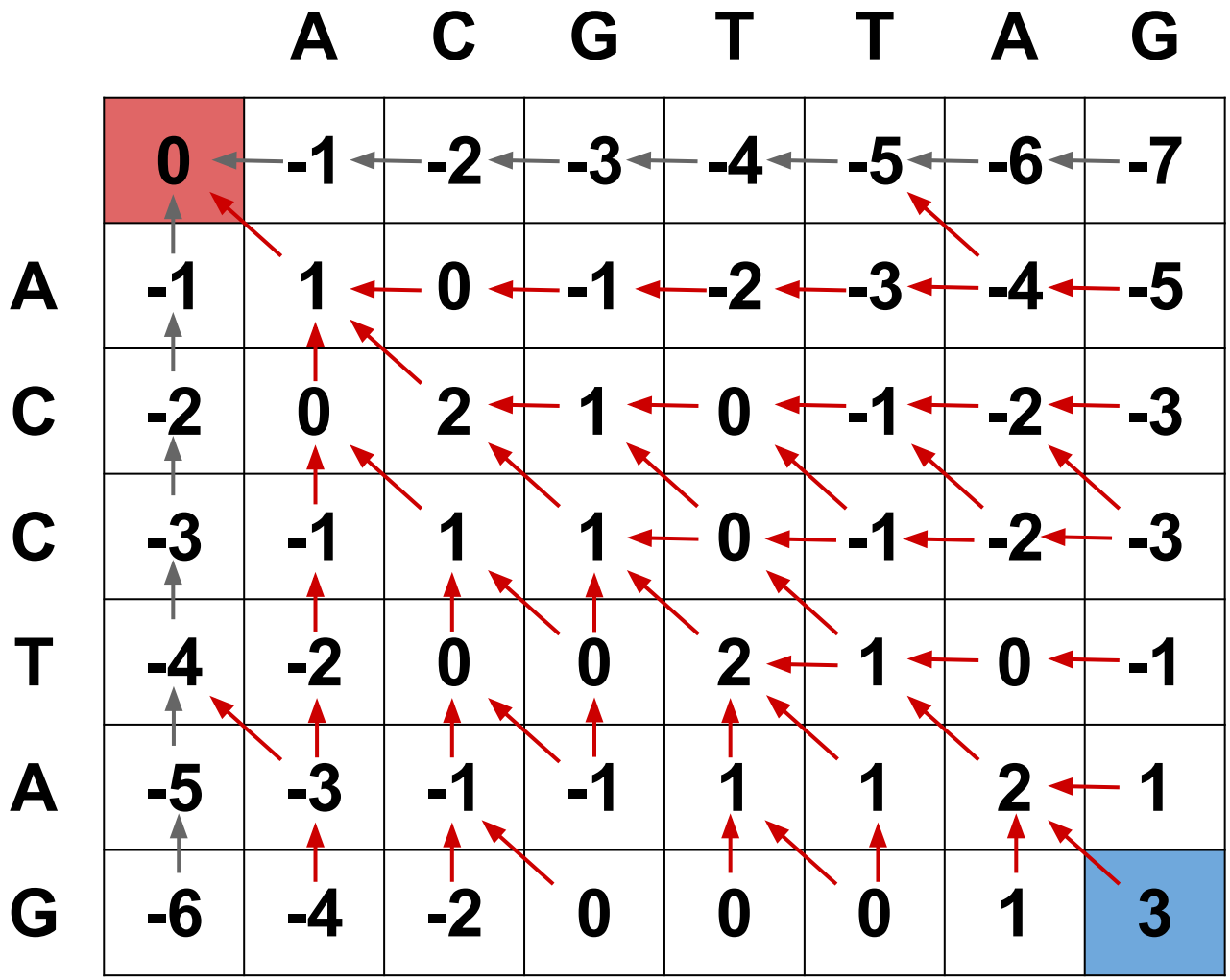
Needleman–Wunsch algorithm

$$M(i, j) = \max \begin{cases} M(i-1, j-1) + \text{Score}(X(i), Y(j)) \\ M(i-1, j) + \text{Score}(X(i), \text{‘-’}) \\ M(i, j-1) + \text{Score}(\text{‘-’}, Y(j)) \end{cases}$$

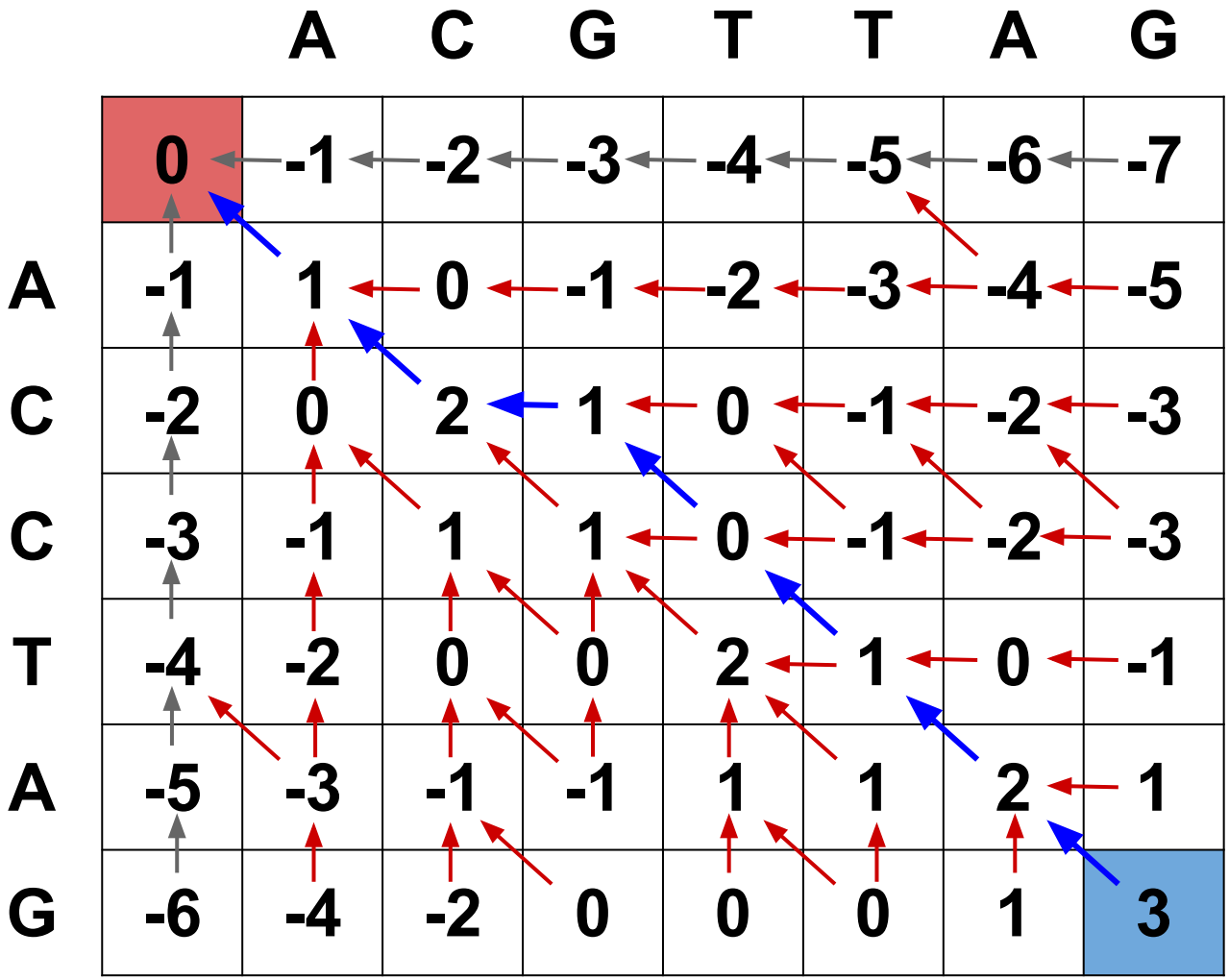
Needleman–Wunsch algorithm



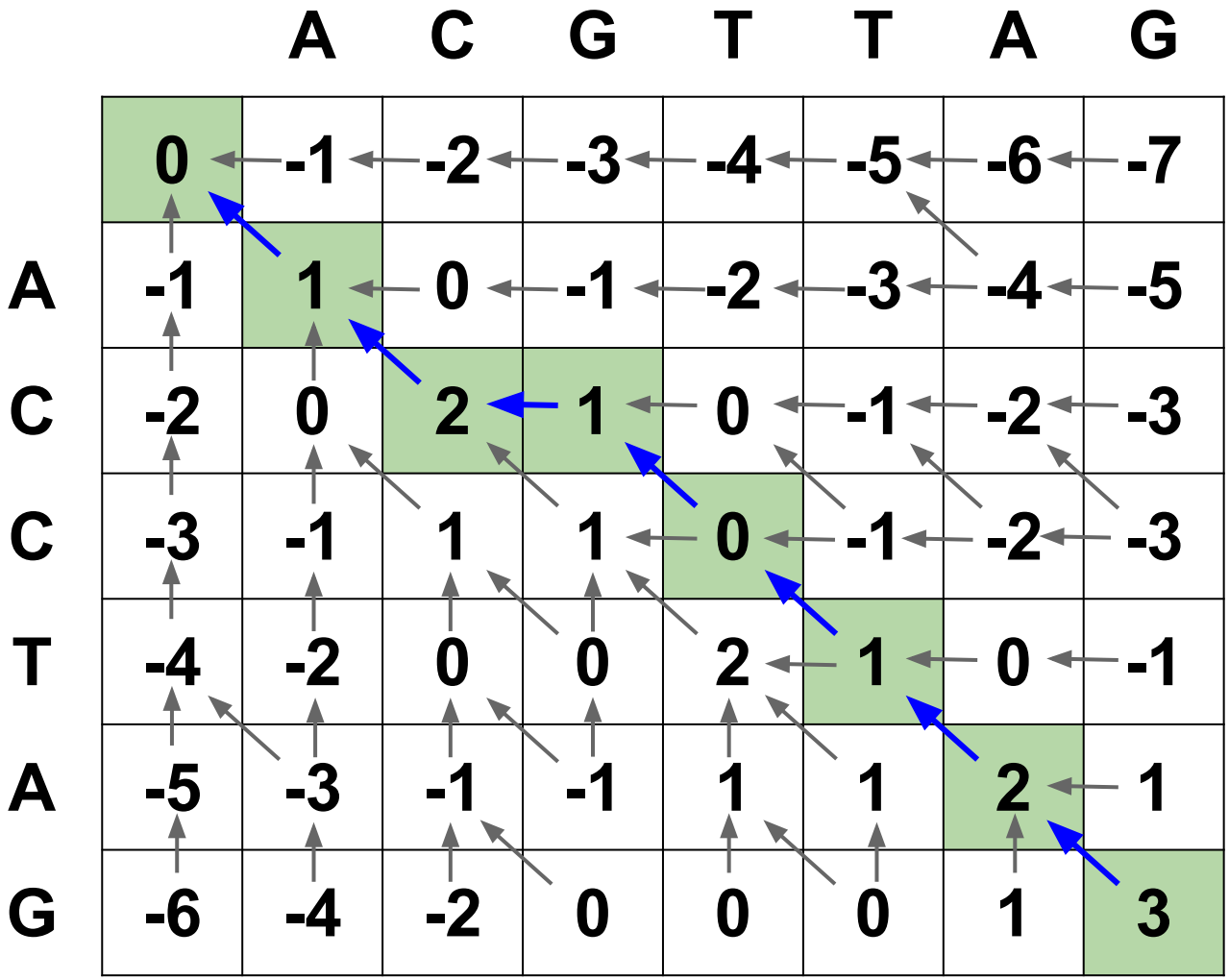
Needleman–Wunsch algorithm



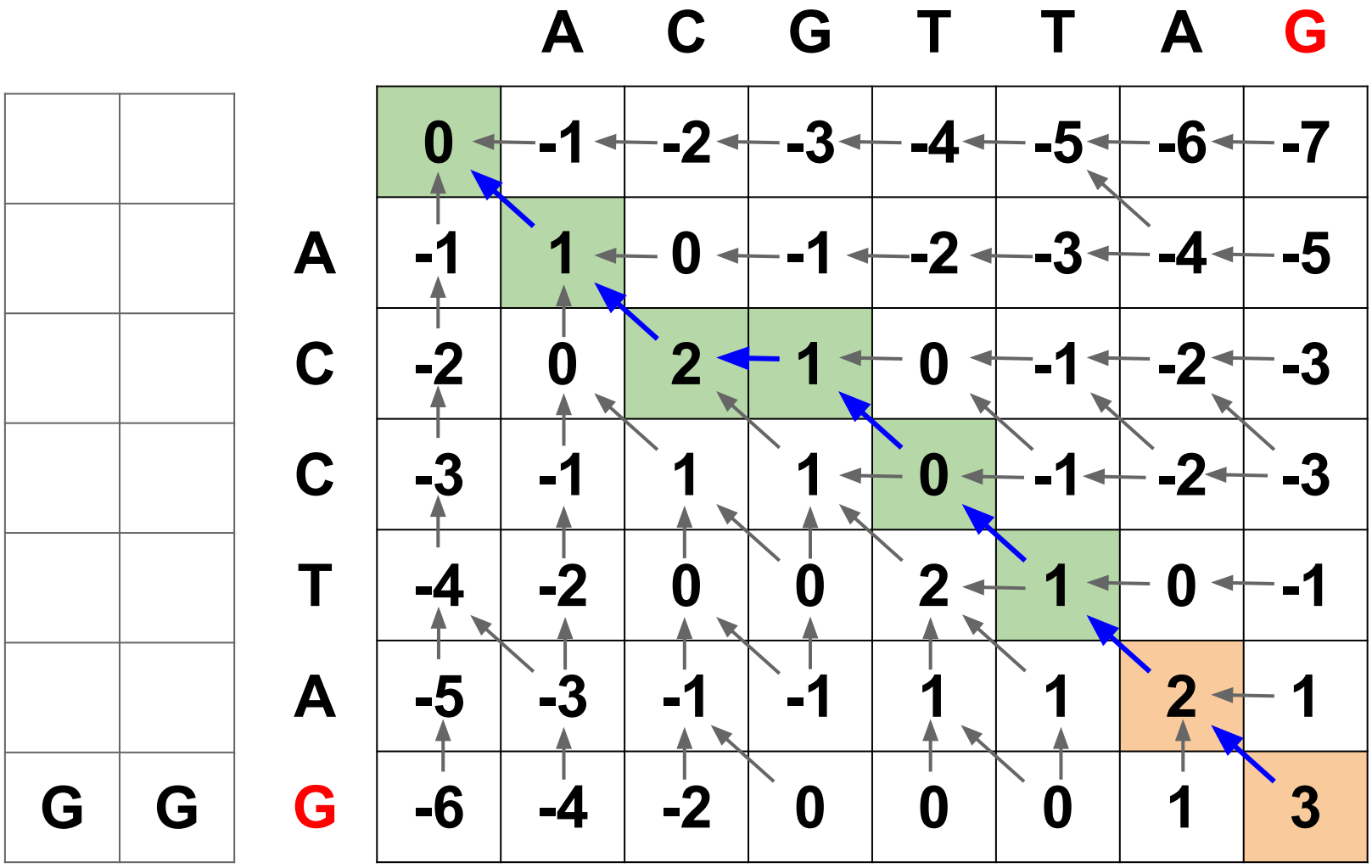
Needleman–Wunsch algorithm



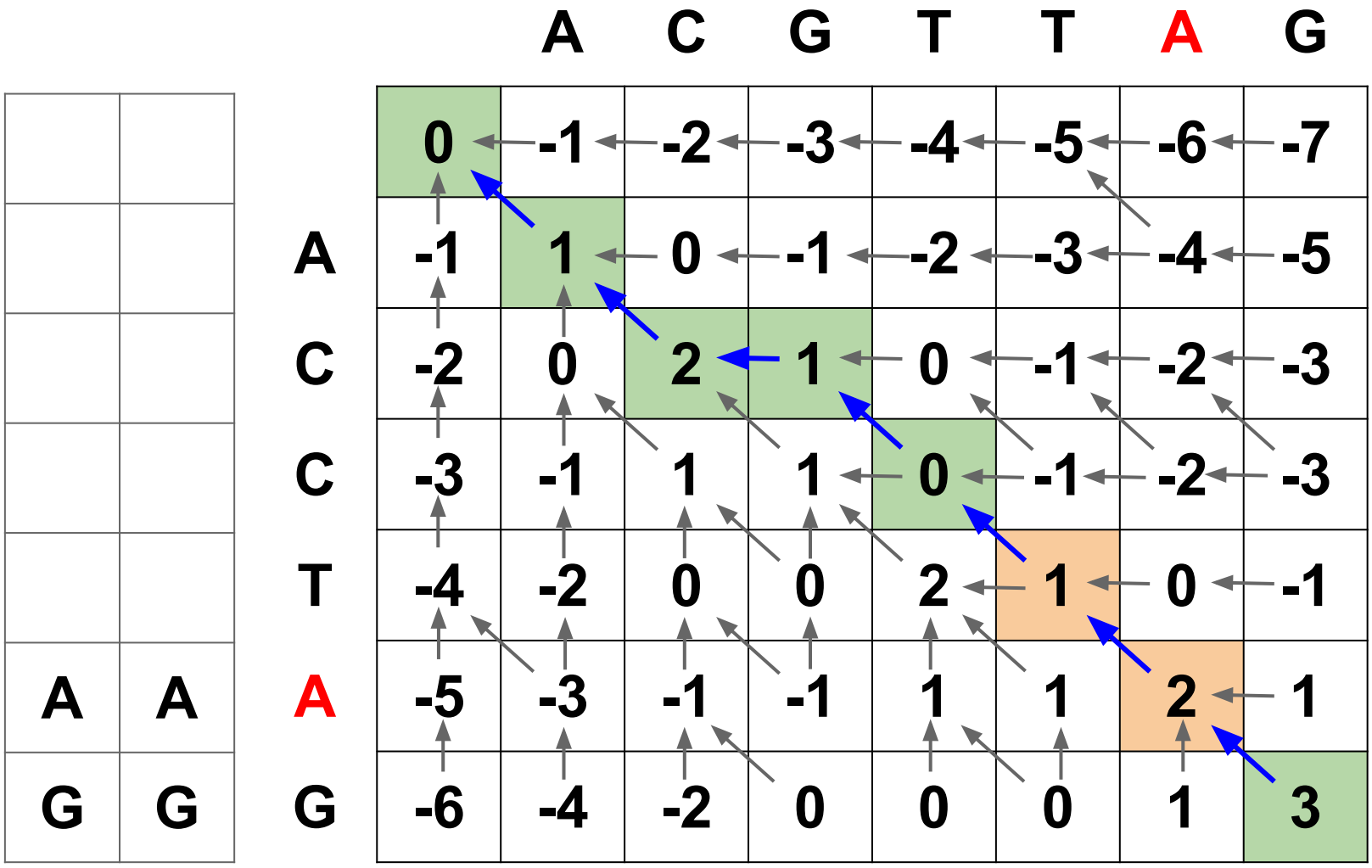
Needleman–Wunsch algorithm



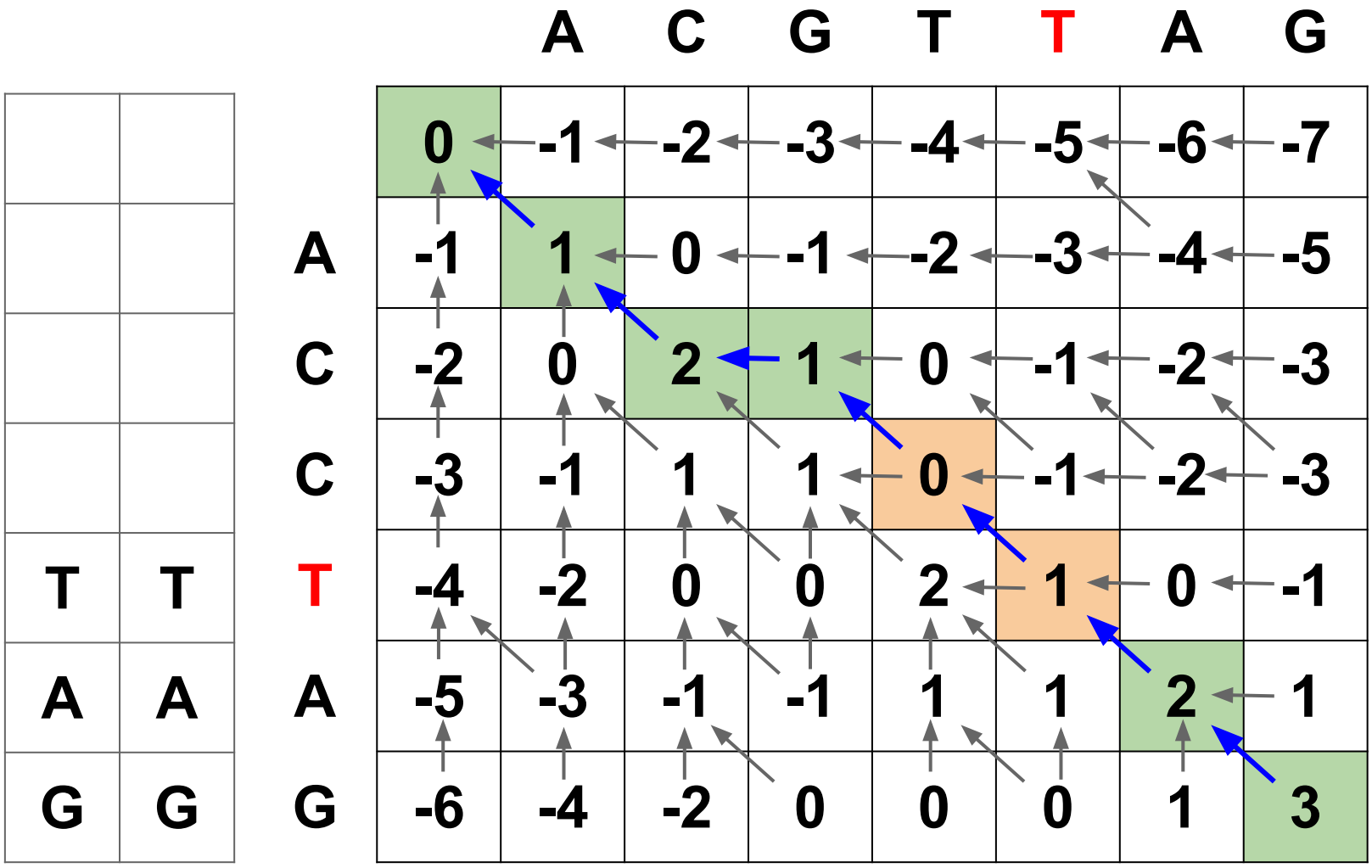
Needleman–Wunsch algorithm



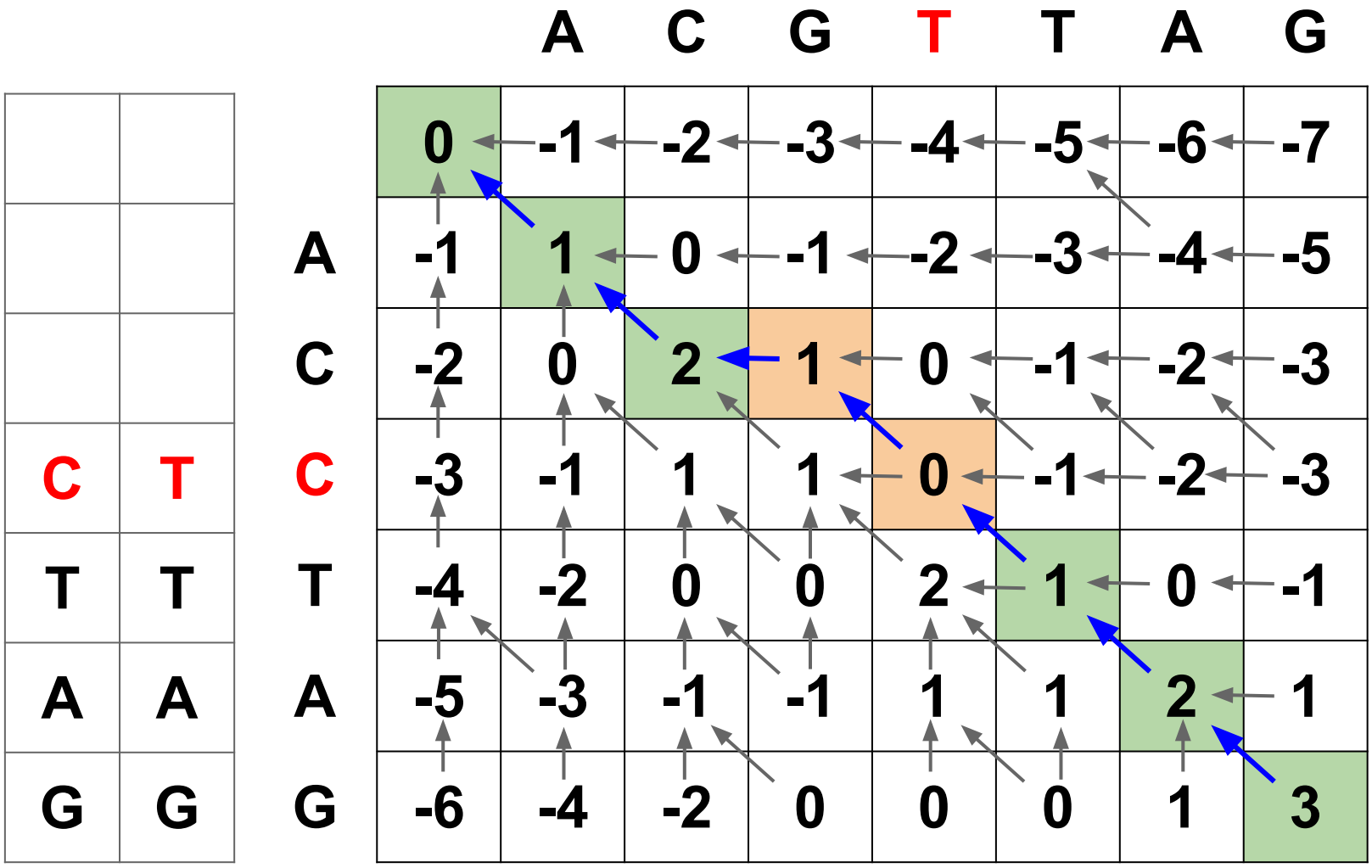
Needleman–Wunsch algorithm



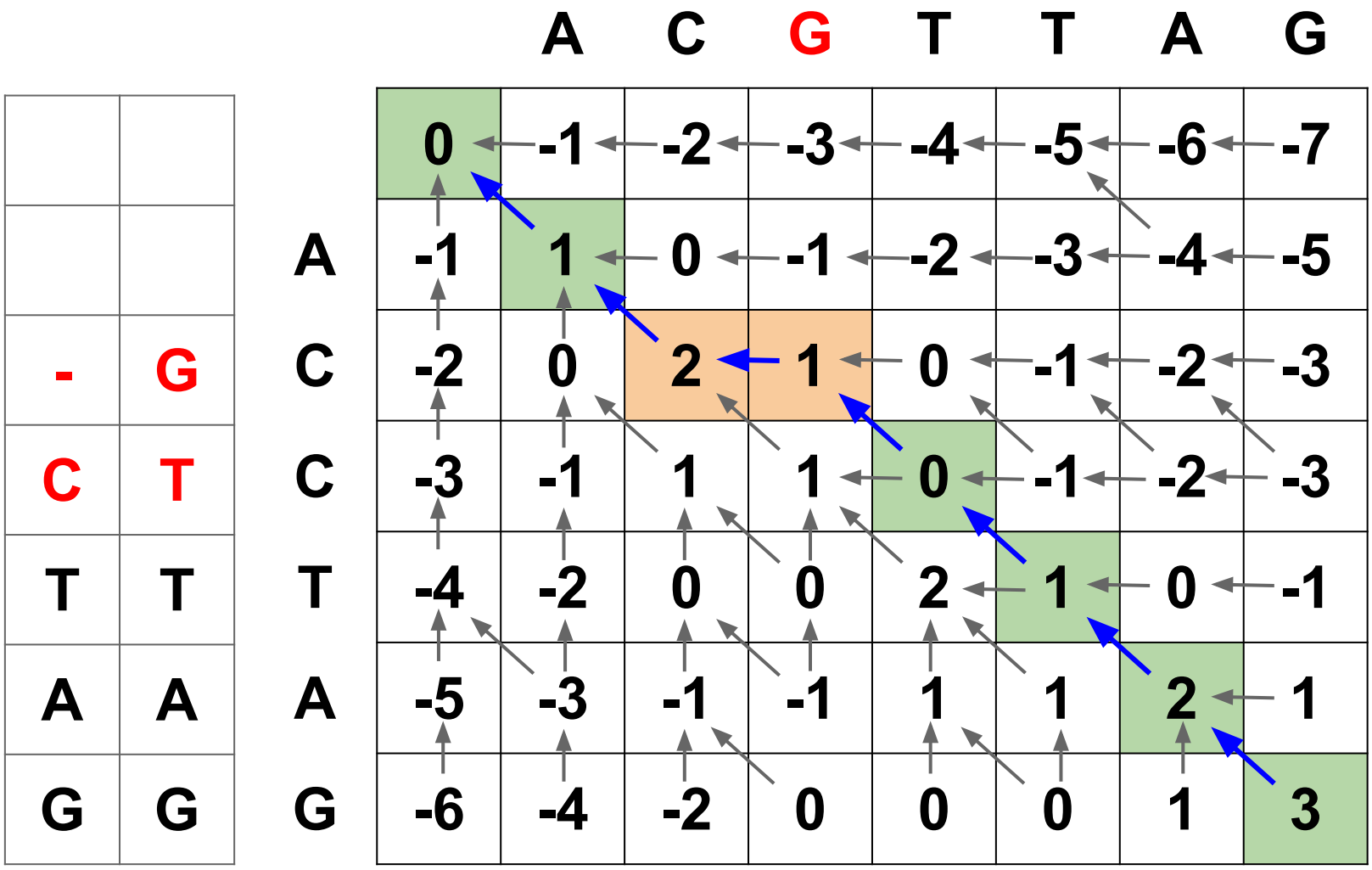
Needleman–Wunsch algorithm



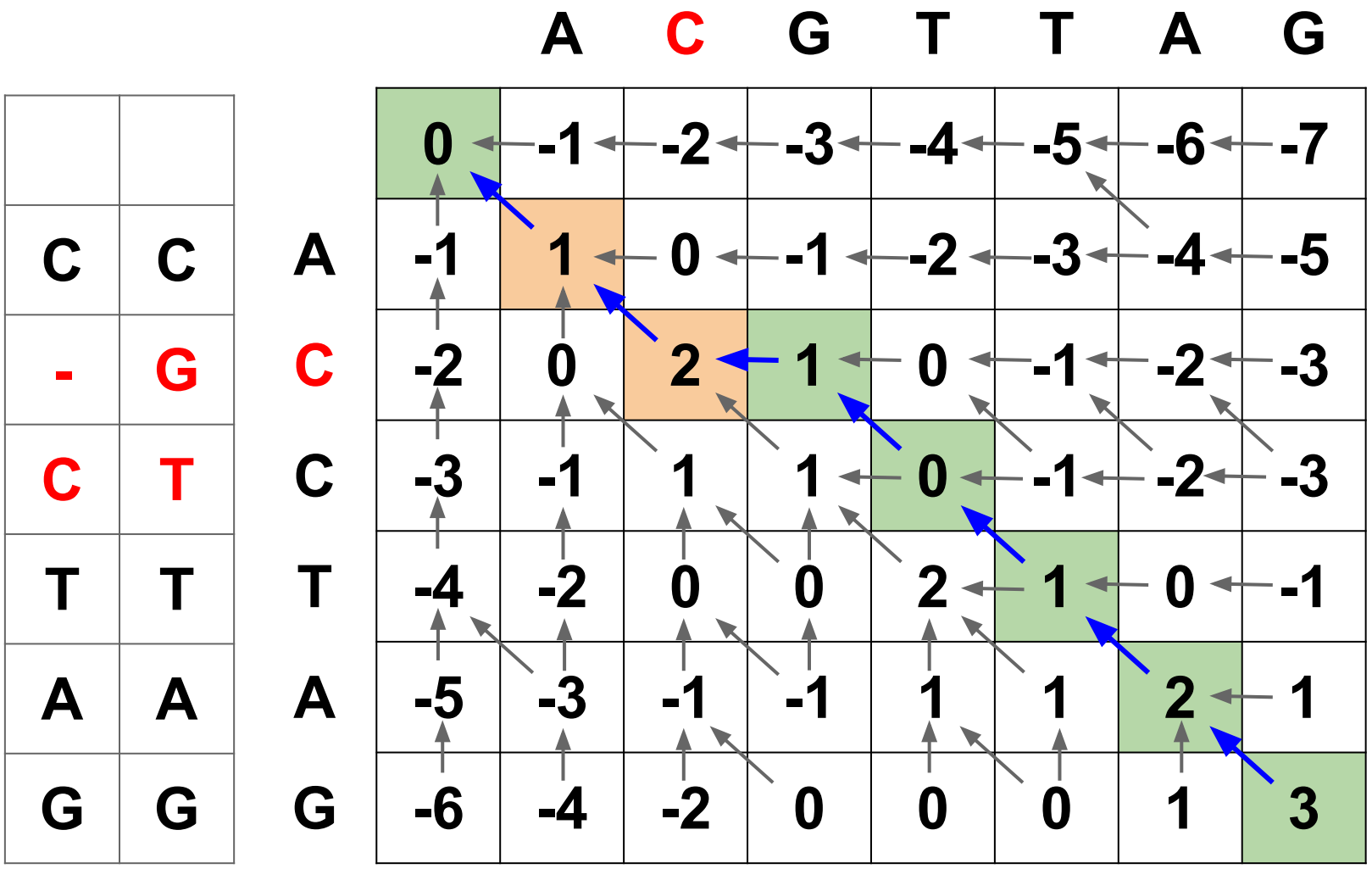
Needleman–Wunsch algorithm



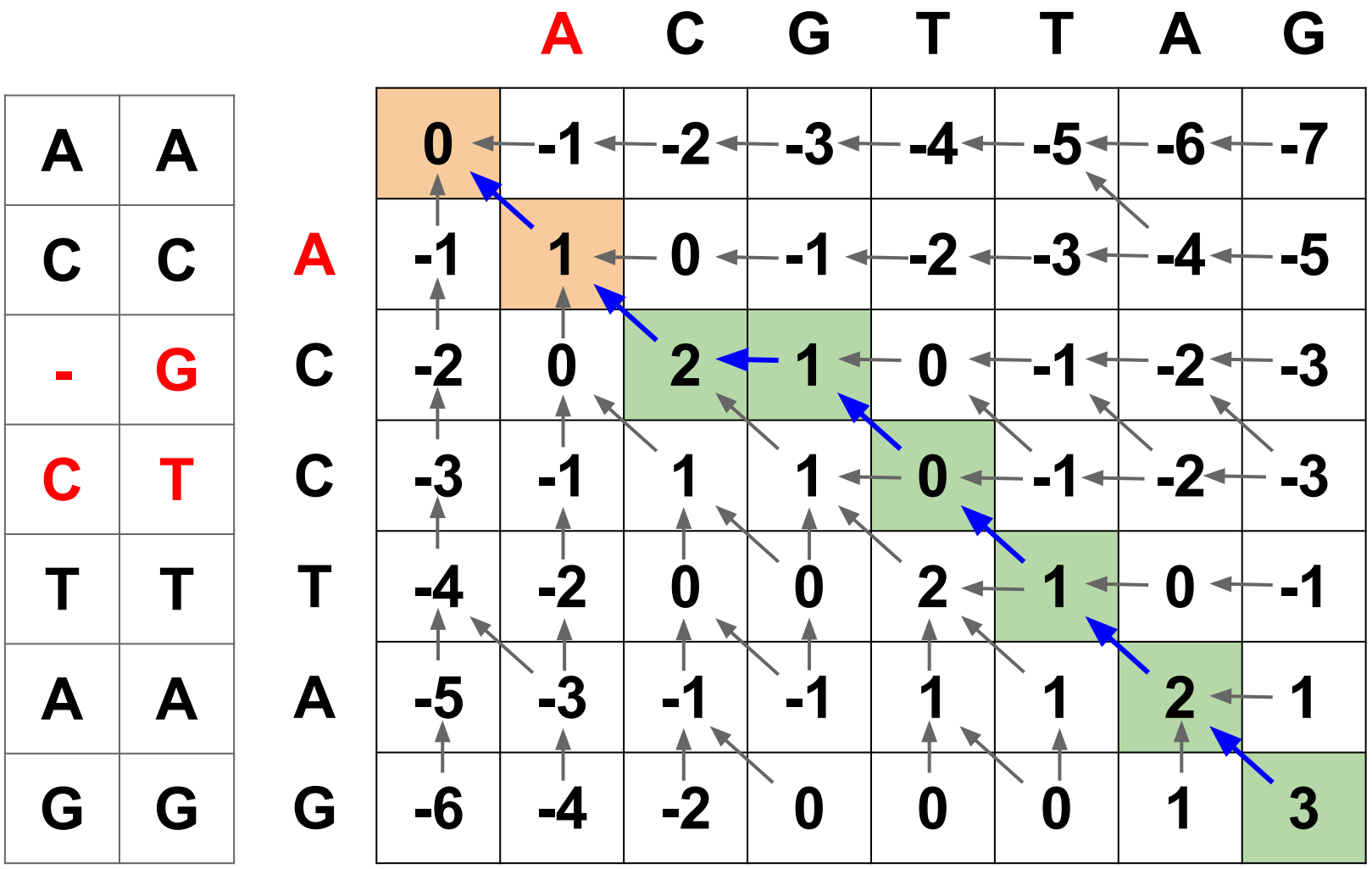
Needleman–Wunsch algorithm



Needleman–Wunsch algorithm



Needleman–Wunsch algorithm



Needleman–Wunsch algorithm

A C G T T A G

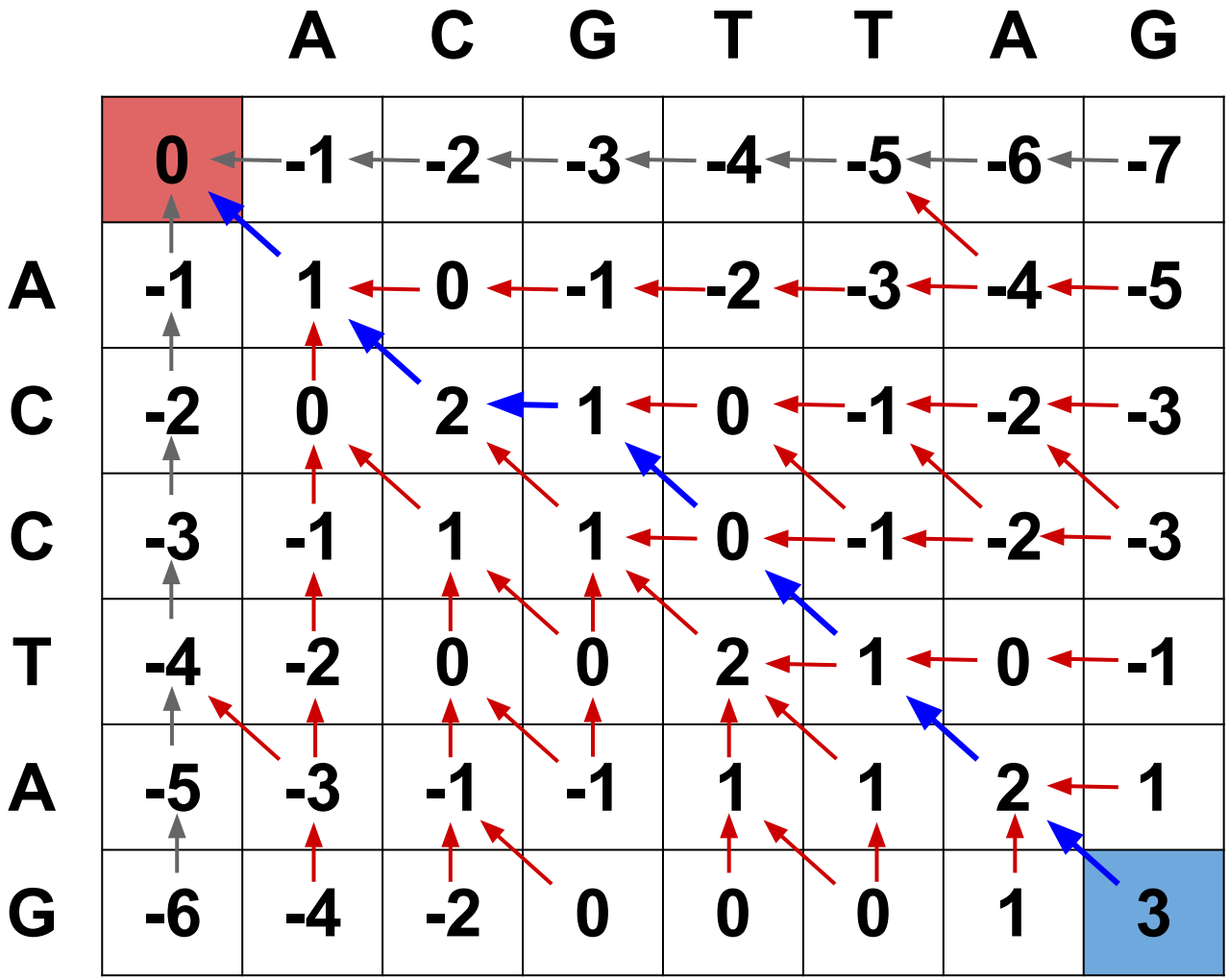
A C C T - A G

A C G T T A G

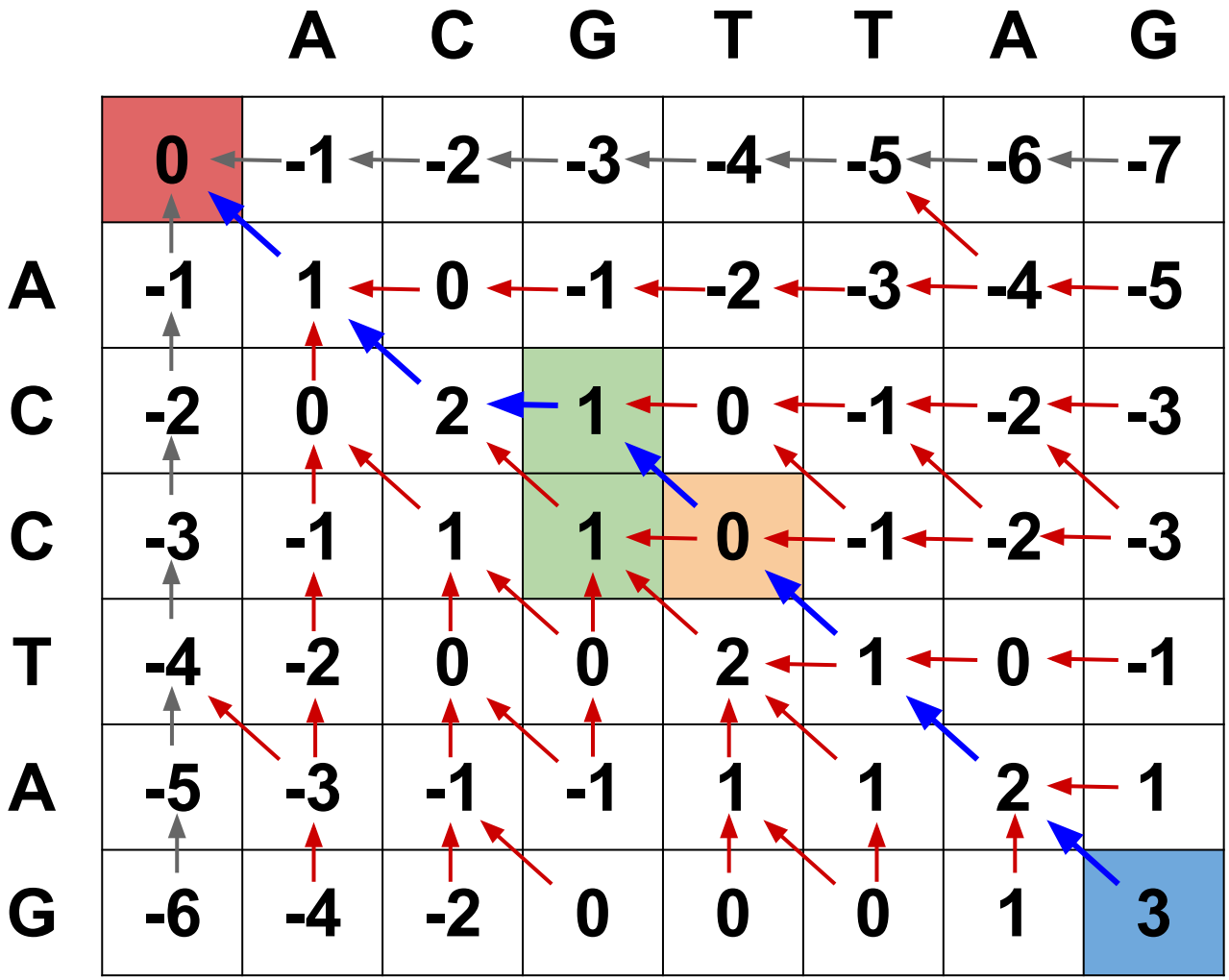
A C - C T A G

Score(ACGTTAG, ACCTAG) = 3

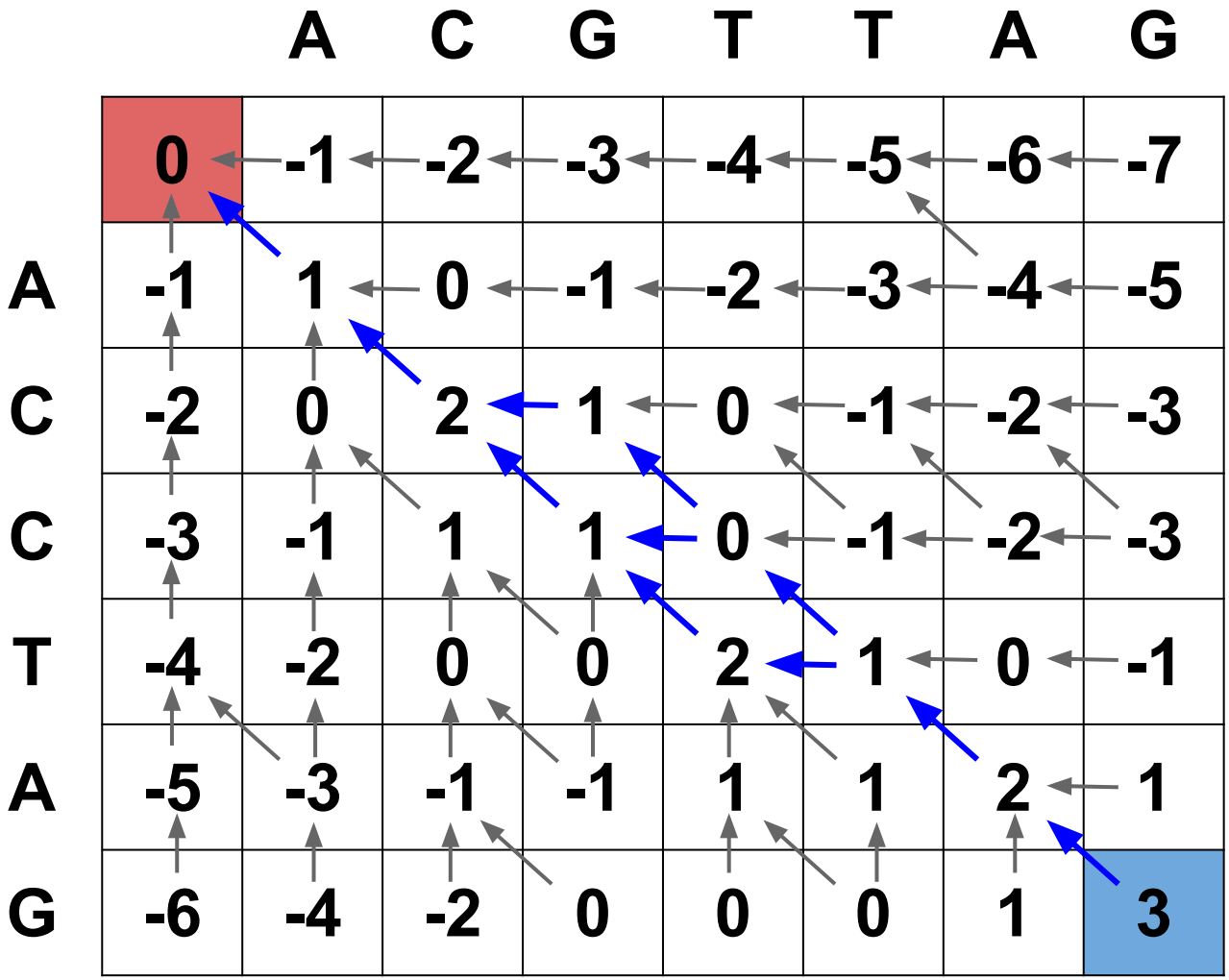
Needleman–Wunsch algorithm



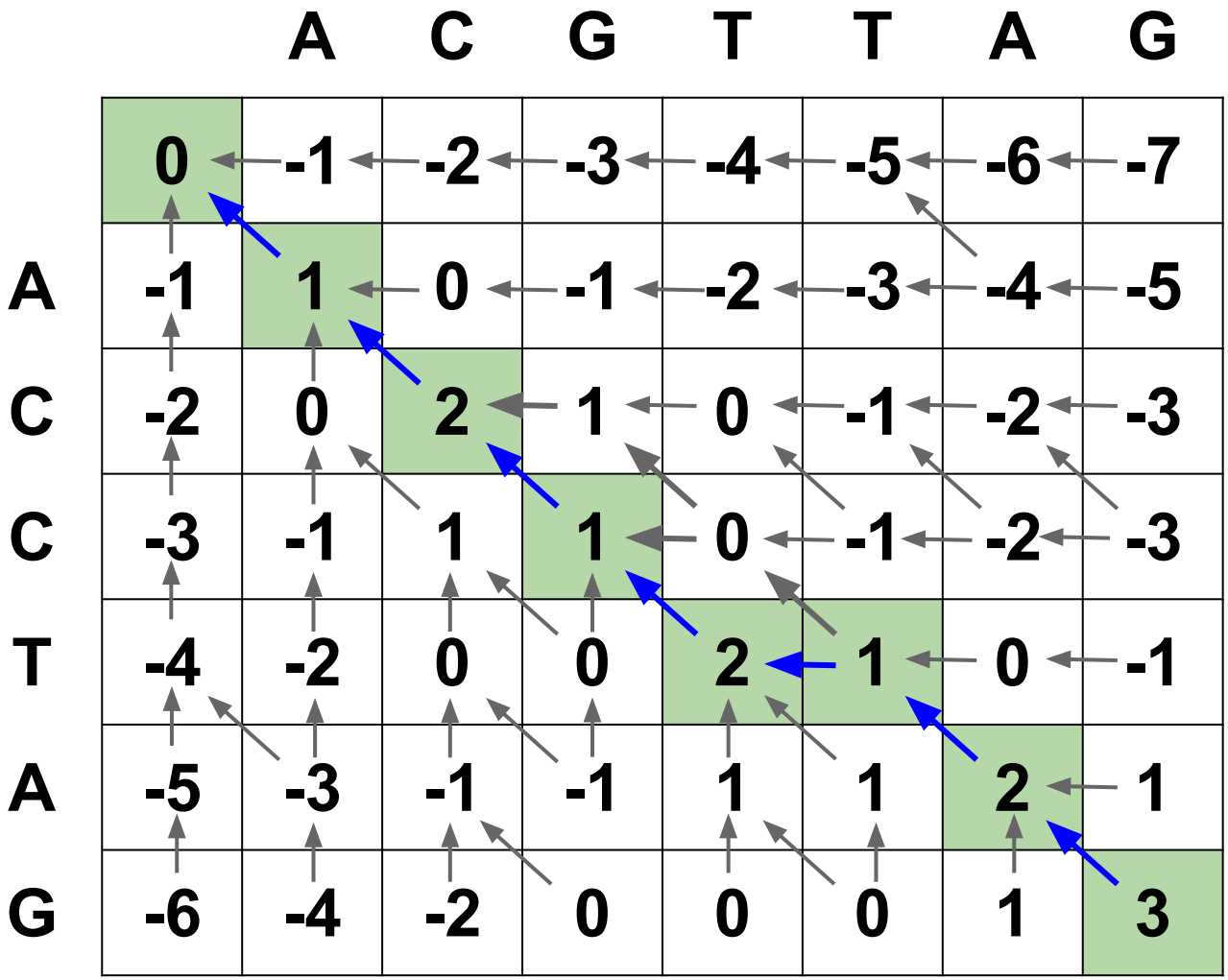
Needleman–Wunsch algorithm



Needleman–Wunsch algorithm



Needleman–Wunsch algorithm



Needleman–Wunsch algorithm

Time =

Memory =

Needleman–Wunsch algorithm

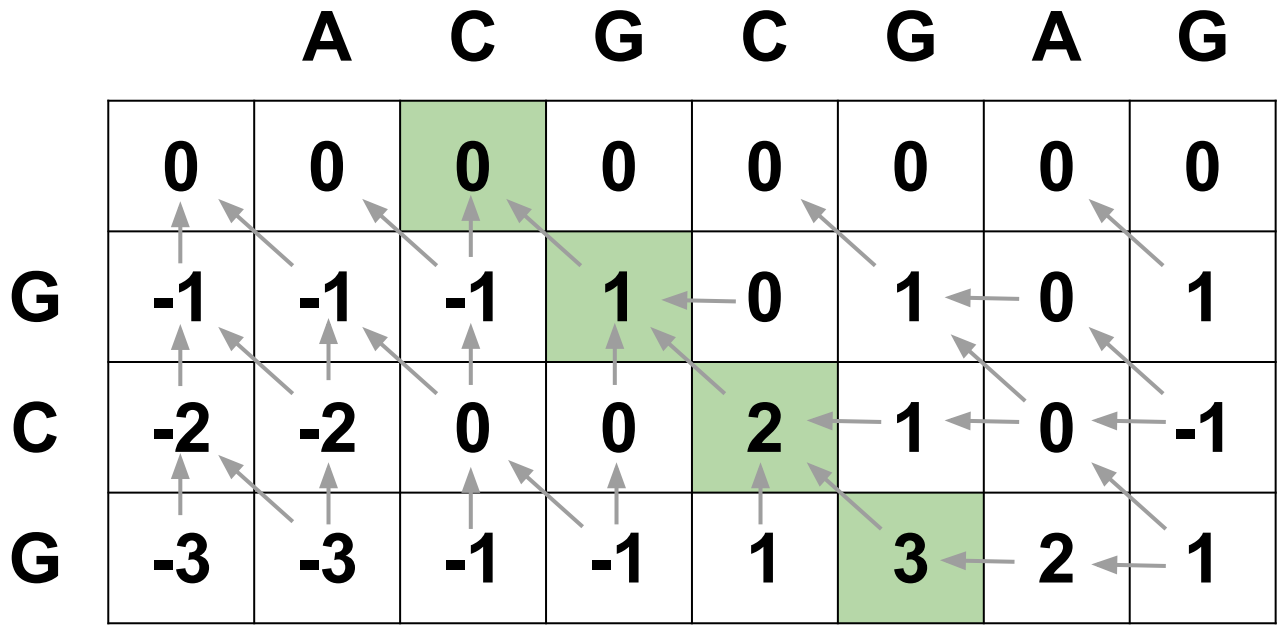
$$\text{Time} = O(n^2)$$

$$\text{Memory} = O(n^2)$$

How about the remaining questions?

- Are they similar?
- Is one of them similar to a part of another?
- What is similar in the sequences?

Needleman–Wunsch algorithm



- Is one of them similar to a part of another?

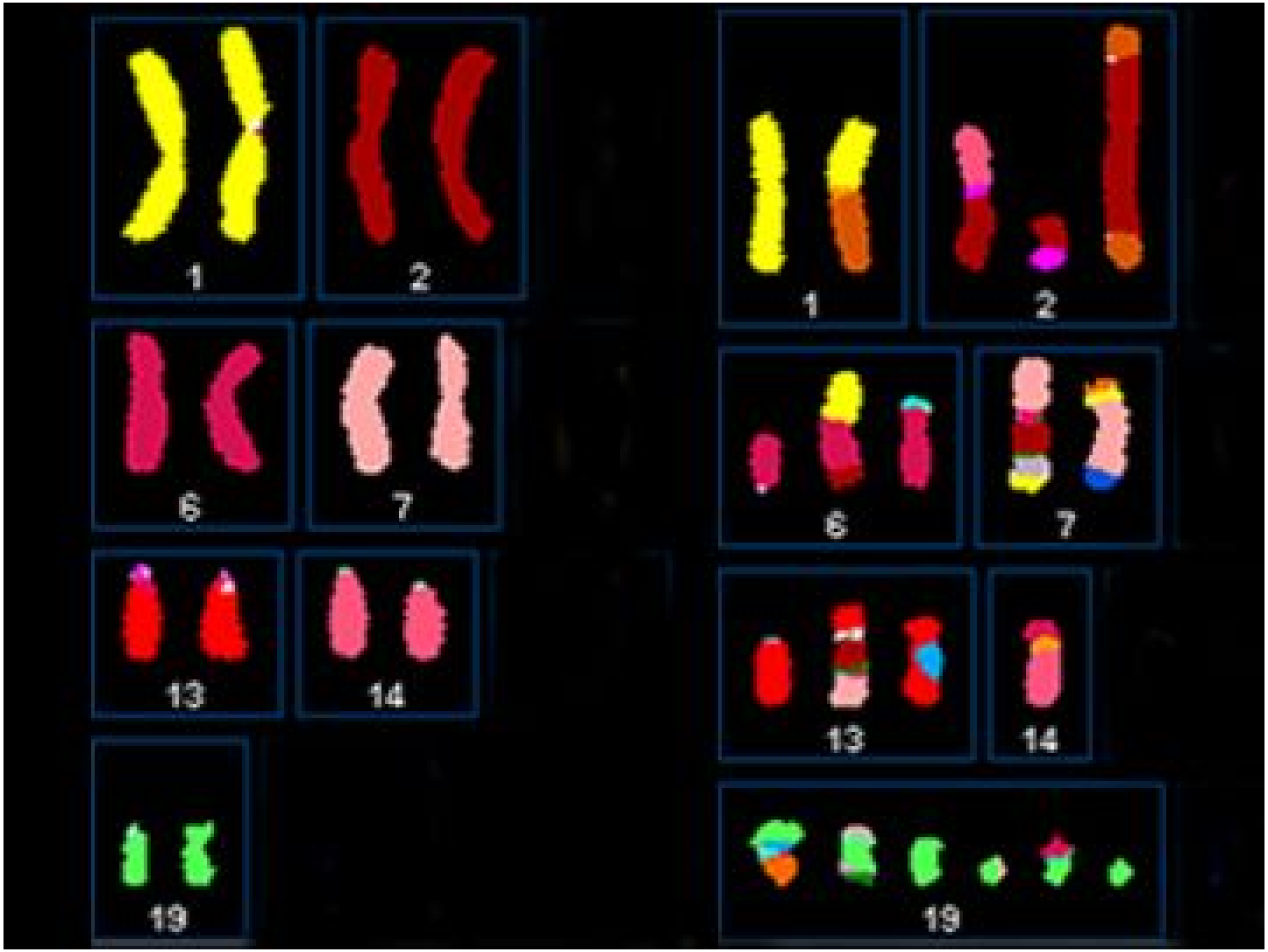
How about the remaining questions?

- Are they similar?
- Is one of them similar to a part of another?
- What is similar in the sequences?

Why is this not enough?

- We deal with genomes of extreme size
- We have extreme number of sequencing reads
- We need more careful alignment that can handle structural rearrangements

Rearrangements



Burrows-Wheeler transform

a c a a c g

Burrows-Wheeler transform

a c a a c g $

Burrows-Wheeler transform

a c a a c g $

$ a c a a c g

Burrows-Wheeler transform

a c a a c g $

$ a c a a c g

g $ a c a a c

Burrows-Wheeler transform

a c a a c g $

$ a c a a c g

g $ a c a a c

c g $ a c a a

a c g $ a c a

a a c g $ a c

c a a c g $ a

Burrows-Wheeler transform

$ a c a a c g

a a c g $ a c

a c a a c g $

a c g $ a c a

c a a c g $ a

c g $ a c a a

g $ a c a a c

Burrows-Wheeler transform

$ a c a a c g

a a c g $ a c

a c a a c g $

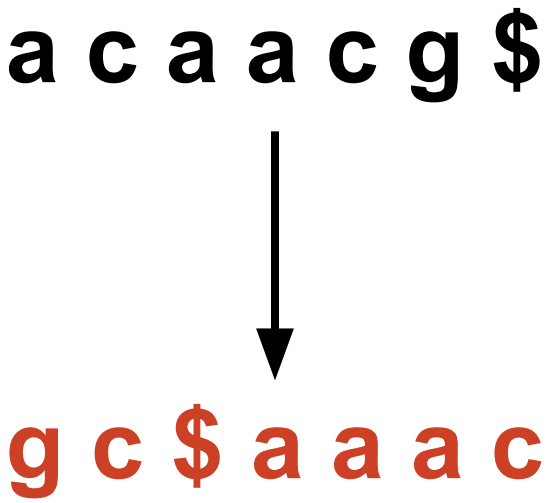
a c g $ a c a

c a a c g $ a

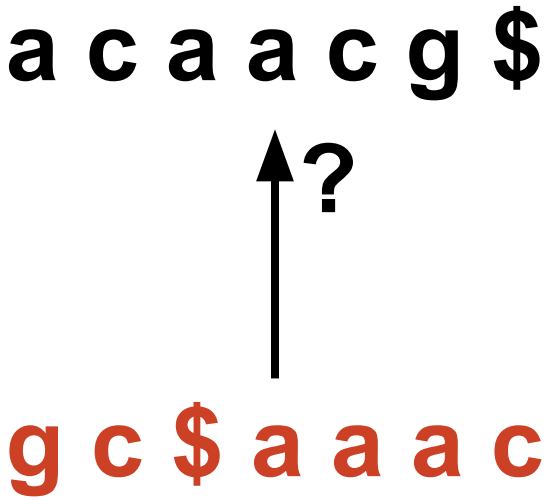
c g $ a c a a

g $ a c a a c

Burrows-Wheeler transform



Burrows-Wheeler transform



Burrows-Wheeler transform

$ a c a a c g

a a c g $ a c

a c a a c g $

a c g $ a c a

c a a c g $ a

c g $ a c a a

g $ a c a a c

Burrows-Wheeler transform

$ a c a a c g

a a c g $ a c

a c a a c g $

a c g $ a c a

c a a c g $ a

c g $ a c a a

g $ a c a a c

Burrows-Wheeler transform

g $ a c a a c

c a a c g $ a

$ a c a a c g

a a c g $ a c

a c a a c g $

a c g $ a c a

c g $ a c a a

Burrows-Wheeler transform

$ a c a a c g

a a c g $ a c

a c a a c g $

a c g $ a c a

c a a c g $ a

c g $ a c a a

g $ a c a a c

Burrows-Wheeler transform

g	$ a c a a c g
c	a a c g $ a c
$	a c a a c g $
a	a c g $ a c a
a	c a a c g $ a
a	c g $ a c a a
c	g $ a c a a c

Burrows-Wheeler transform

$	$ a c a a c g
a	a a c g $ a c
a	a c a a c g $
a	a c g $ a c a
c	c a a c g $ a
c	c g $ a c a a
g	g $ a c a a c

Burrows-Wheeler transform

$	$ a c a a c g
a	a a c g $ a c
a	a c a a c g $
a	a c g $ a c a
c	c a a c g $ a
c	c g $ a c a a
g	g $ a c a a c

Burrows-Wheeler transform

g $	$ a c a a c g
c a	a a c g $ a c
$ a	a c a a c g $
a a	a c g $ a c a
a c	c a a c g $ a
a c	c g $ a c a a
c g	g $ a c a a c

Burrows-Wheeler transform

$ a	$ a c a a c g
a a	a a c g $ a c
a c	a c a a c g $
a c	a c g $ a c a
c a	c a a c g $ a
c g	c g $ a c a a
g $	g $ a c a a c

Burrows-Wheeler transform

$ a	$ a c a a c g
a a	a a c g $ a c
a c	a c a a c g $
a c	a c g $ a c a
c a	c a a c g $ a
c g	c g $ a c a a
g $	g $ a c a a c

Burrows-Wheeler transform

g $ a	$ a c a a c g
c a a	a a c g $ a c
$ a c	a c a a c g $
a a c	a c g $ a c a
a c a	c a a c g $ a
a c g	c g $ a c a a
c g $	g $ a c a a c

Burrows-Wheeler transform

$ a c	$ a c a a c g
a a c	a a c g $ a c
a c a	a c a a c g $
a c g	a c g $ a c a
c a a	c a a c g $ a
c g $	c g $ a c a a
g $ a	g $ a c a a c

Burrows-Wheeler transform

$ a c	$ a c a a c g	
a a c	a a c g $ a c	
a c a	a c a a c g $	
a c g	a c g $ a c a	
c a a	c a a c g $ a	
c g $	c g $ a c a a	
g $ a	g $ a c a a c	

Burrows-Wheeler transform

g $ a c	$ a c a a c g
c a a c	a a c g $ a c
$ a c a	a c a a c g $
a a c g	a c g $ a c a
a c a a	c a a c g $ a
a c g $	c g $ a c a a
c g $ a	g $ a c a a c

Burrows-Wheeler transform

$ a c a	$ a c a	a c g
a a c g	a a c g	$ a c
a c a a	a c a a	c g $
a c g $	a c g $	a c a
c a a c	c a a c	g $ a
c g $ a	c g $ a	c a a
g $ a c	g $ a c	a a c

Burrows-Wheeler transform

g $ a c a	$ a c a a c g
c a a c g	a a c g $ a c
$ a c a a	a c a a c g $
a a c g $	a c g $ a c a
a c a a c	c a a c g $ a
a c g $ a	c g $ a c a a
c g $ a c	g $ a c a a c

Burrows-Wheeler transform

$ a c a a	$ a c a a c	g
a a c g $	a a c g $ a	c
a c a a c	a c a a c g	$
a c g $ a	a c g $ a c	a
c a a c g	c a a c g $	a
c g $ a c	c g $ a c a	a
g $ a c a	g $ a c a a	c

Burrows-Wheeler transform

g $ a c a a	$ a c a a c g
c a a c g $	a a c g $ a c
$ a c a a c	a c a a c g $
a a c g $ a	a c g $ a c a
a c a a c g	c a a c g $ a
a c g $ a c	c g $ a c a a
c g $ a c a	g $ a c a a c

Burrows-Wheeler transform

$ a c a a c	$ a c a a c g
a a c g $ a	a a c g $ a c
a c a a c g	a c a a c g $
a c g $ a c	a c g $ a c a
c a a c g $	c a a c g $ a
c g $ a c a	c g $ a c a a
g $ a c a a	g $ a c a a c

Burrows-Wheeler transform

g $ a c a a c	$ a c a a c g
c a a c g $ a	a a c g $ a c
$ a c a a c g	a c a a c g $
a a c g $ a c	a c g $ a c a
a c a a c g $	c a a c g $ a
a c g $ a c a	c g $ a c a a
c g $ a c a a	g $ a c a a c

Burrows-Wheeler transform

$ a c a a c g	$ a c a a c g
a a c g $ a c	a a c g $ a c
a c a a c g $	a c a a c g $
a c g $ a c a	a c g $ a c a
c a a c g $ a	c a a c g $ a
c g $ a c a a	c g $ a c a a
g $ a c a a c	g $ a c a a c

First-last property

$$\$_1\ a\ c\ a\ a\ c\ g$$
$$a_1\ a\ c\ g\ \$\ a\ c$$
$$a_2\ c\ a\ a\ c\ g\ \$$$
$$a_3\ c\ g\ \$\ a\ c\ a$$
$$c_1\ a\ a\ c\ g\ \$\ a$$
$$c_2\ g\ \$\ a\ c\ a\ a$$
$$g_1\ \$\ a\ c\ a\ a\ c$$

First-last property

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c

a$_2$ c a a c g $_1$

a$_3$ c g $ a c a

c$_1$ a a c g $ a

c$_2$ g $ a c a a

g$_1$ $ a c a a c

First-last property

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c

a$_2$ c a a c g $$_1$

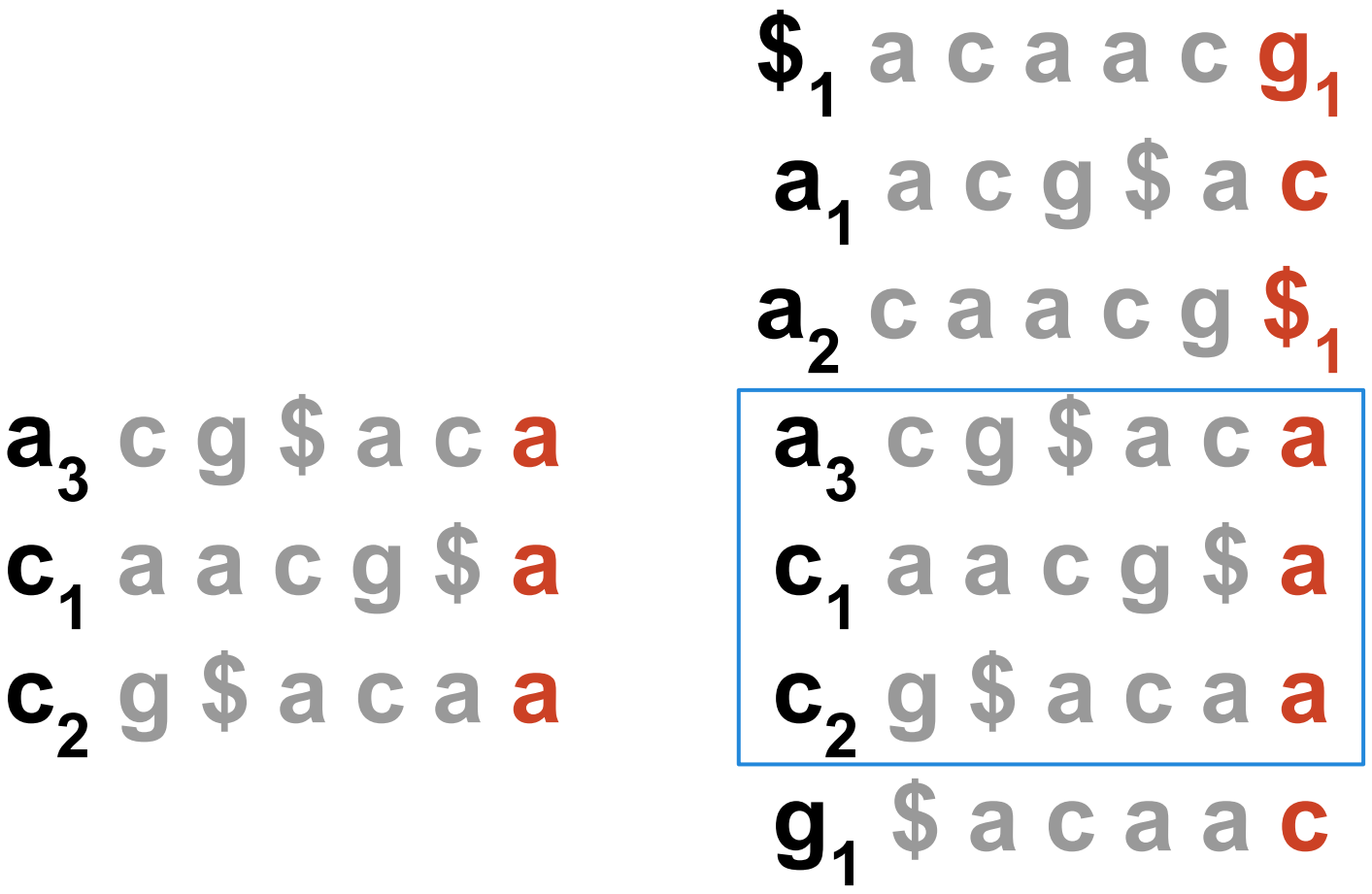
a$_3$ c g $ a c a

c$_1$ a a c g $ a

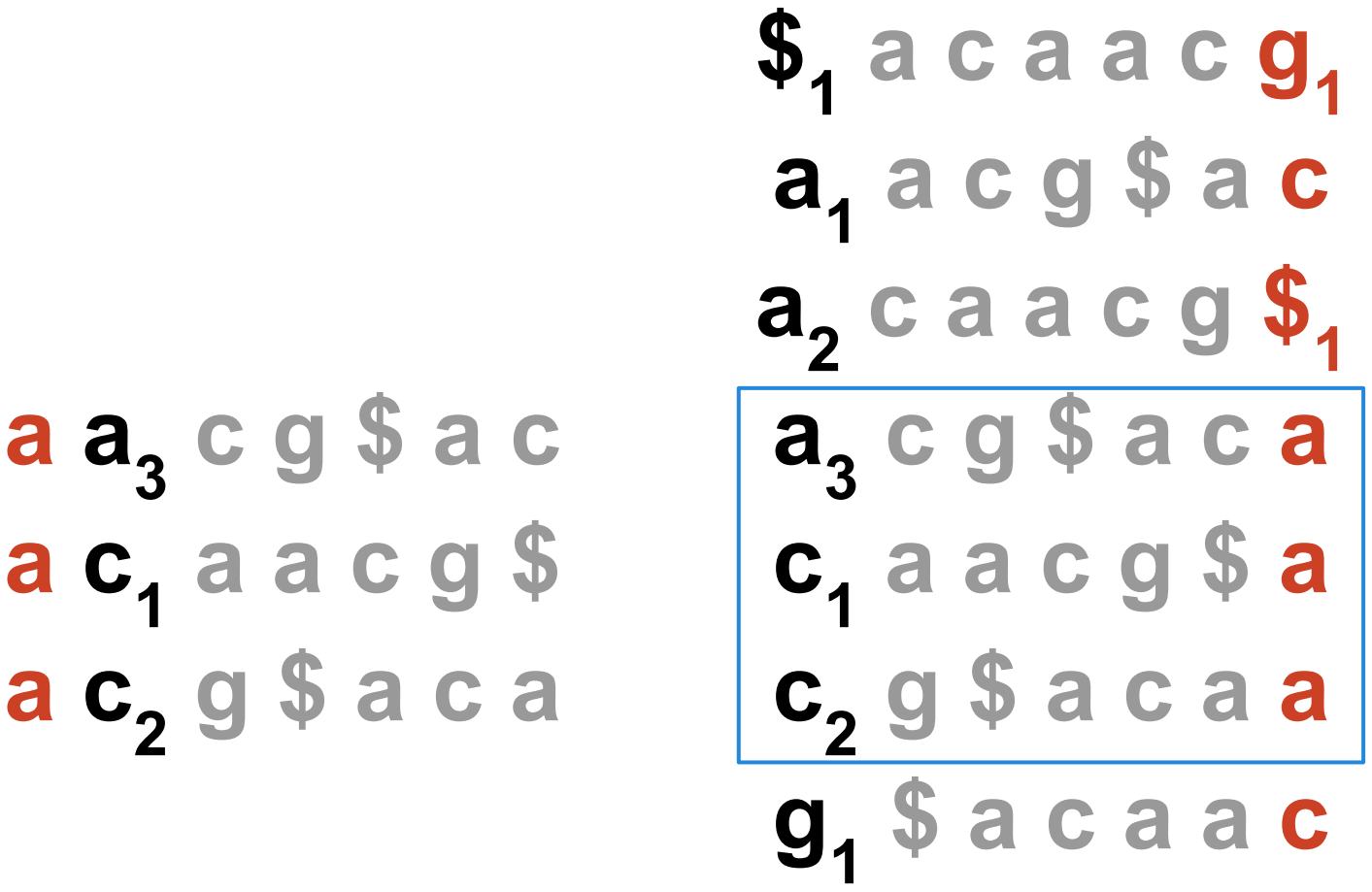
c$_2$ g $ a c a a

g$_1$ $ a c a a c

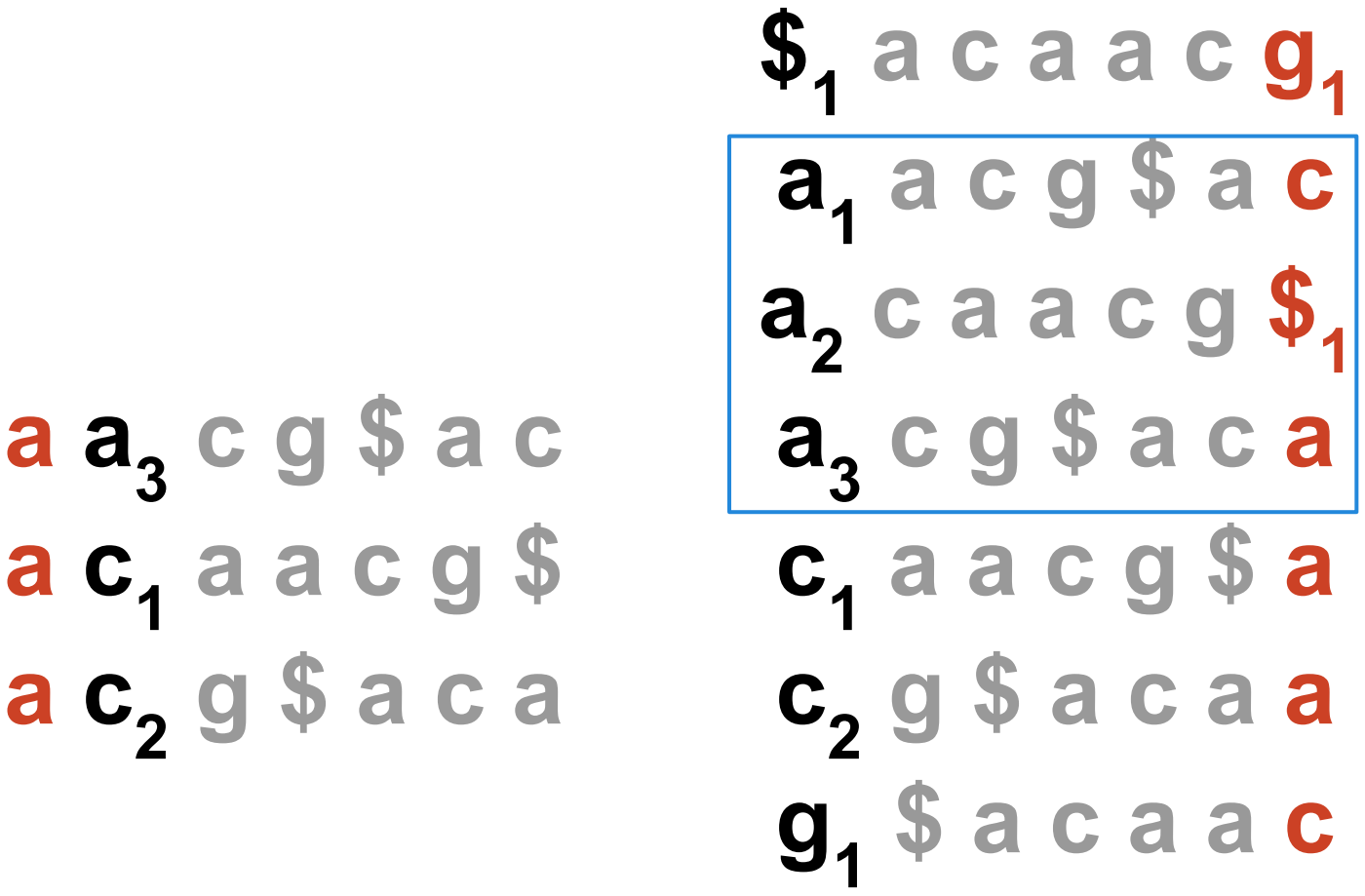
First-last property



First-last property



First-last property



First-last property

$$a_1\ a_3\ c\ g\ \$\ a\ c$$
$$a_2\ c_1\ a\ a\ c\ g\ \$$$
$$a_3\ c_2\ g\ \$\ a\ c\ a$$

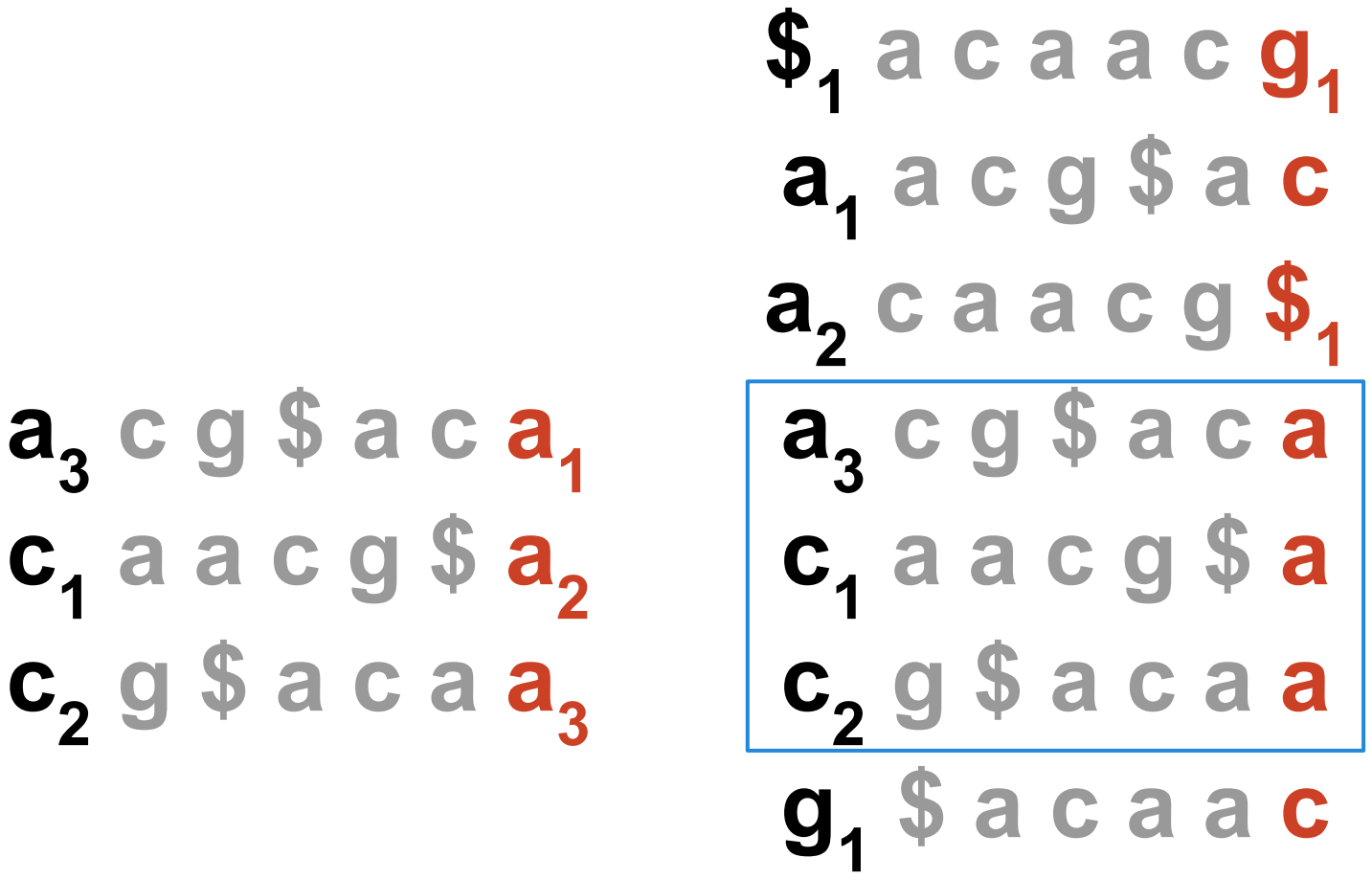
$$\$_1\ a\ c\ a\ a\ c\ g_1$$
$$a_1\ a\ c\ g\ \$\ a\ c$$
$$a_2\ c\ a\ a\ c\ g\ \$_1$$
$$a_3\ c\ g\ \$\ a\ c\ a$$
$$c_1\ a\ a\ c\ g\ \$\ a$$
$$c_2\ g\ \$\ a\ c\ a\ a$$
$$g_1\ \$\ a\ c\ a\ a\ c$$

First-last property

$$\$_1\ a\ c\ a\ a\ c\ g_1$$
$$a_1\ a\ c\ g\ \$\ a\ c$$
$$a_2\ c\ a\ a\ c\ g\ \$_1$$
$$a_3\ c\ g\ \$\ a\ c\ a_1$$
$$c_1\ a\ a\ c\ g\ \$\ a_2$$
$$c_2\ g\ \$\ a\ c\ a\ a_3$$

$$a_3\ c\ g\ \$\ a\ c\ a$$
$$c_1\ a\ a\ c\ g\ \$\ a$$
$$c_2\ g\ \$\ a\ c\ a\ a$$
$$g_1\ \$\ a\ c\ a\ a\ c$$

First-last property



First-last property

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c

a$_2$ c a a c g $_1$

a$_3$ c g $ a c a$_1$

c$_1$ a a c g $ a$_2$

c$_2$ g $ a c a a$_3$

g$_1$ $ a c a a c

First-last property

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c$_1$

a$_2$ c a a c g $_1$

a$_3$ c g $ a c a$_1$

c$_1$ a a c g $ a$_2$

c$_2$ g $ a c a a$_3$

g$_1$ $ a c a a c$_2$

Pattern search

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c$_1$

a$_2$ c a a c g $$_1$

a$_3$ c g $ a c a$_1$

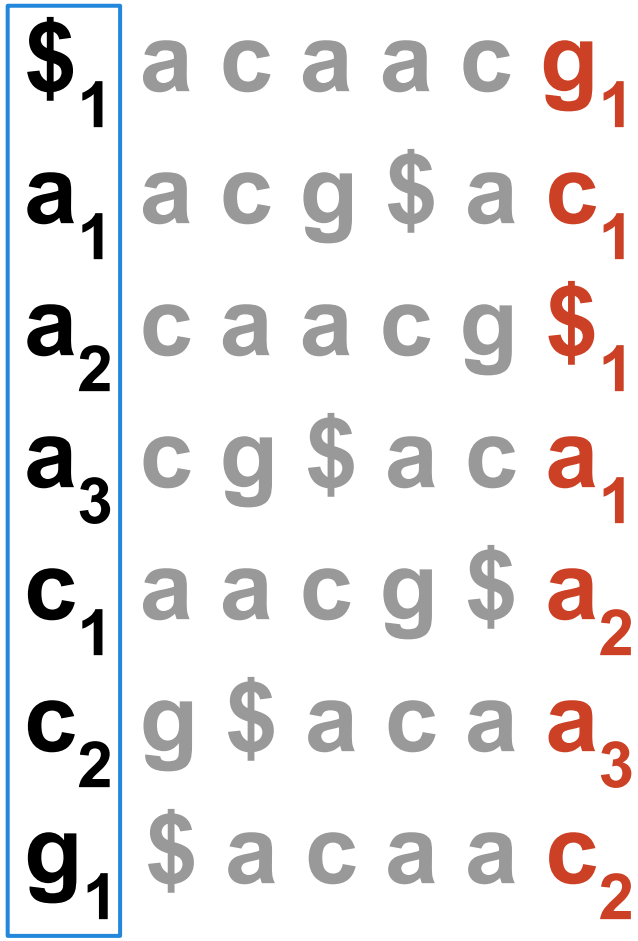
c$_1$ a a c g $ a$_2$

c$_2$ g $ a c a a$_3$

g$_1$ $ a c a a c$_2$

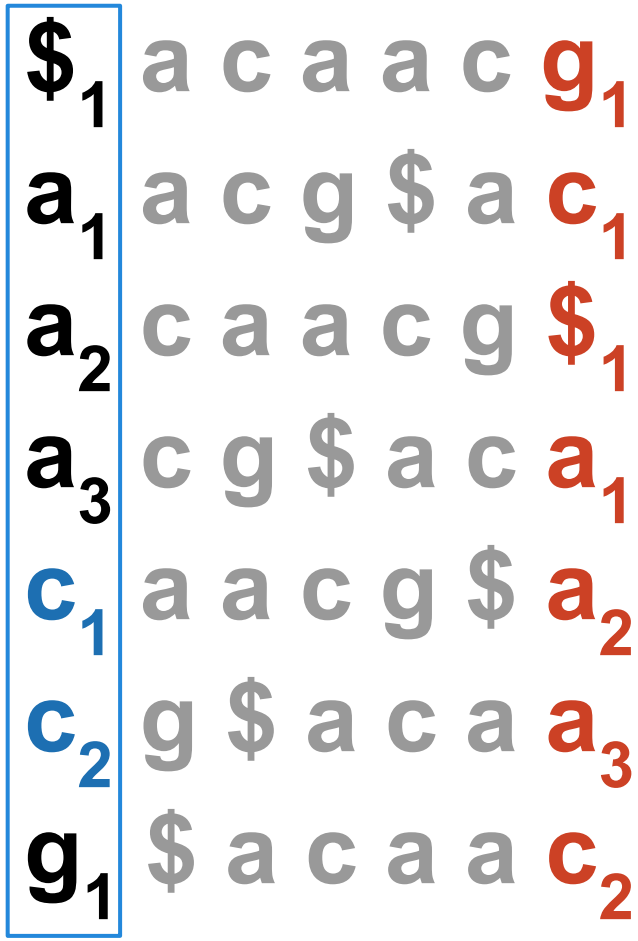
a a c

Pattern search



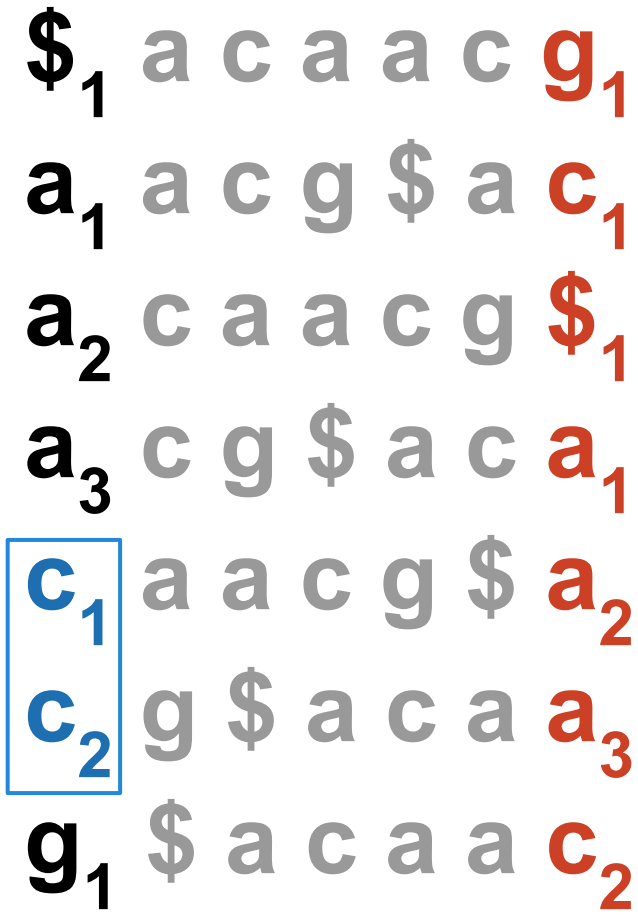
a a c

Pattern search



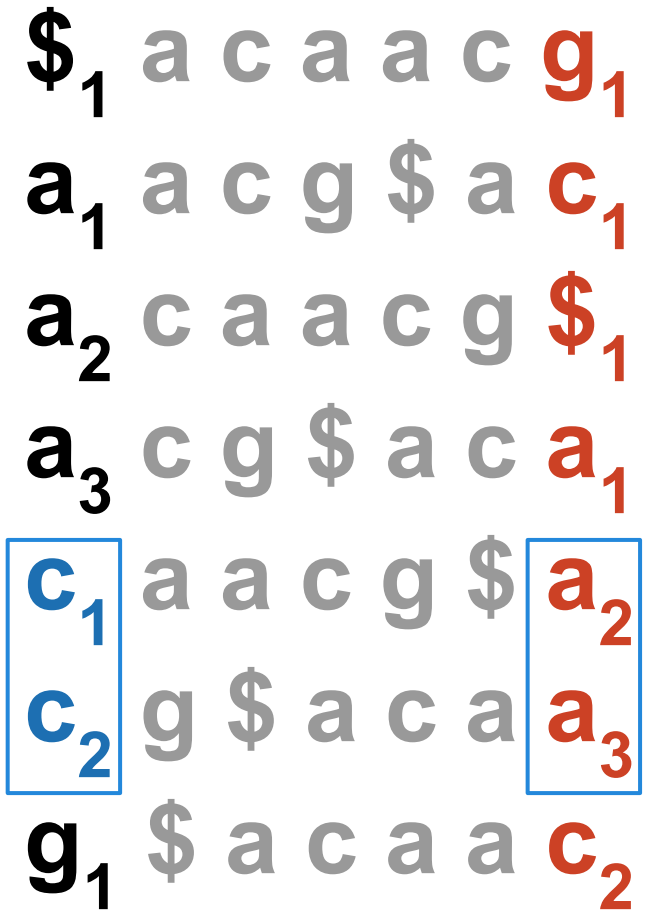
a a c

Pattern search



a a c

Pattern search



a a c

Pattern search

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c$_1$

a$_2$ c a a c g $$_1$

a$_3$ c g $ a c a$_1$

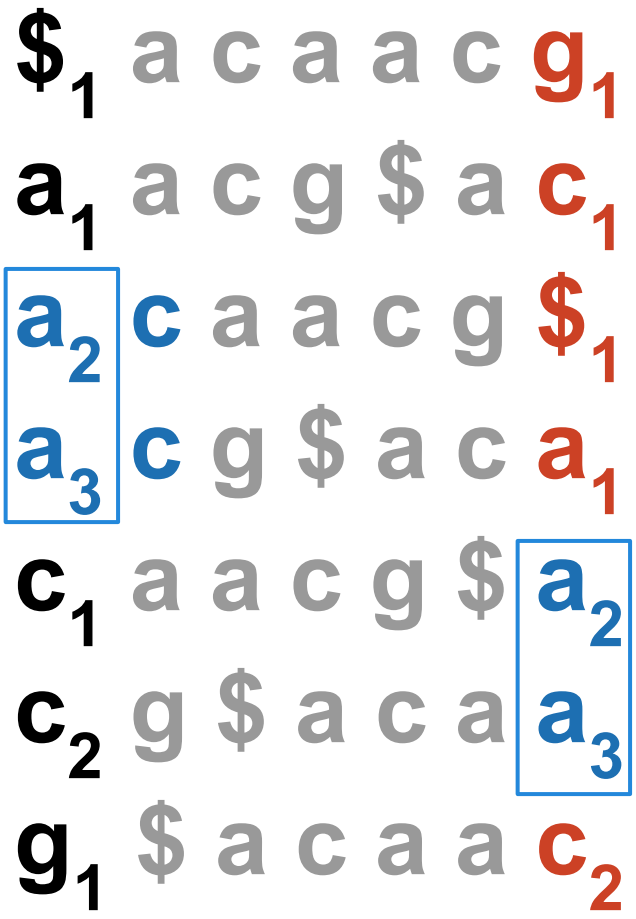
c$_1$ a a c g $ a$_2$

c$_2$ g $ a c a a$_3$

g$_1$ $ a c a a c$_2$

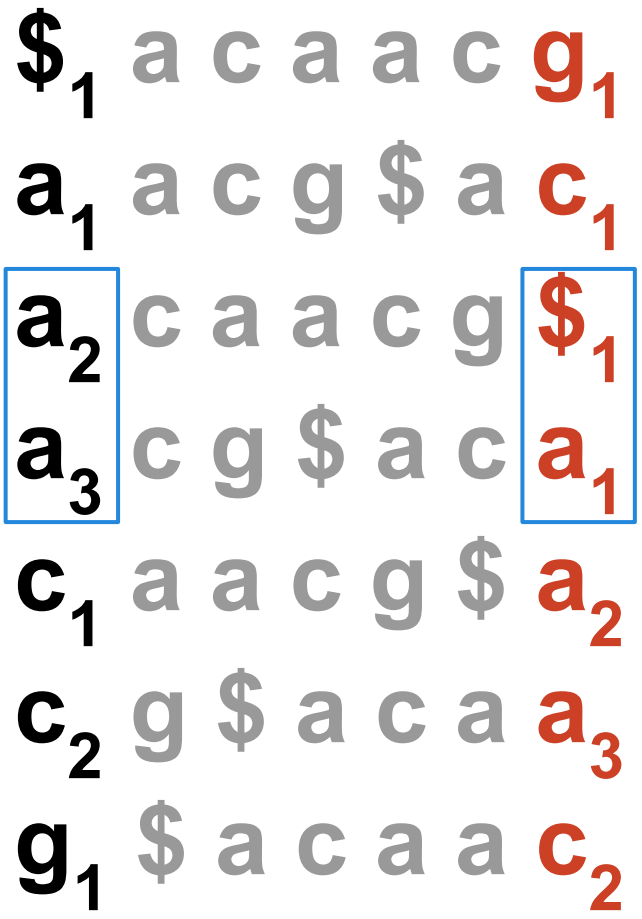
a a c

Pattern search



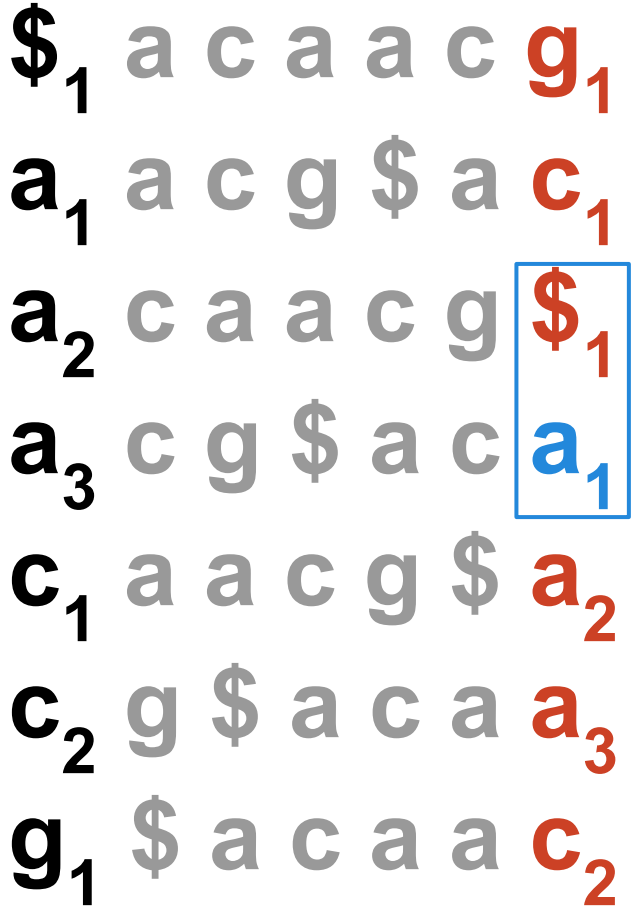
a a c

Pattern search



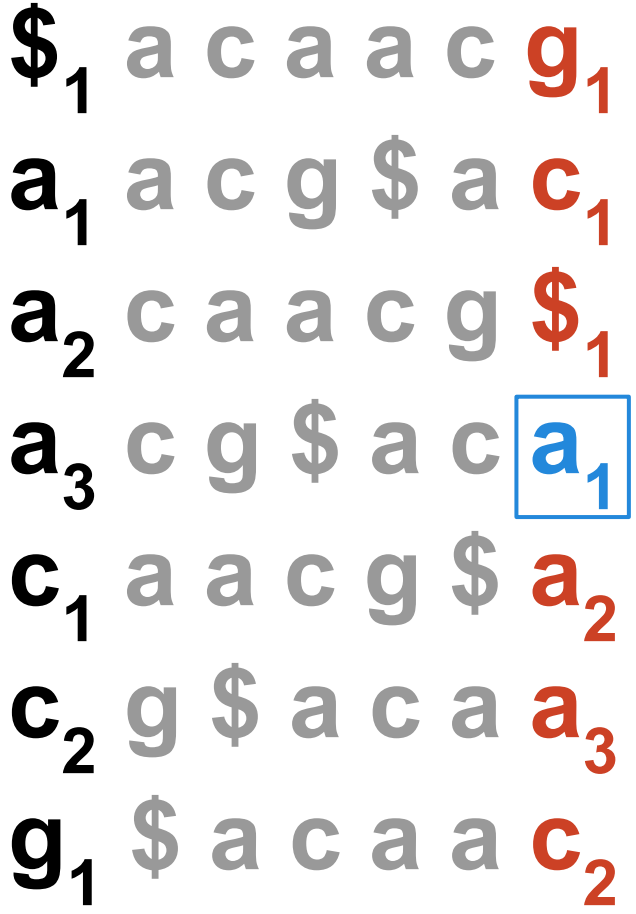
a a c

Pattern search



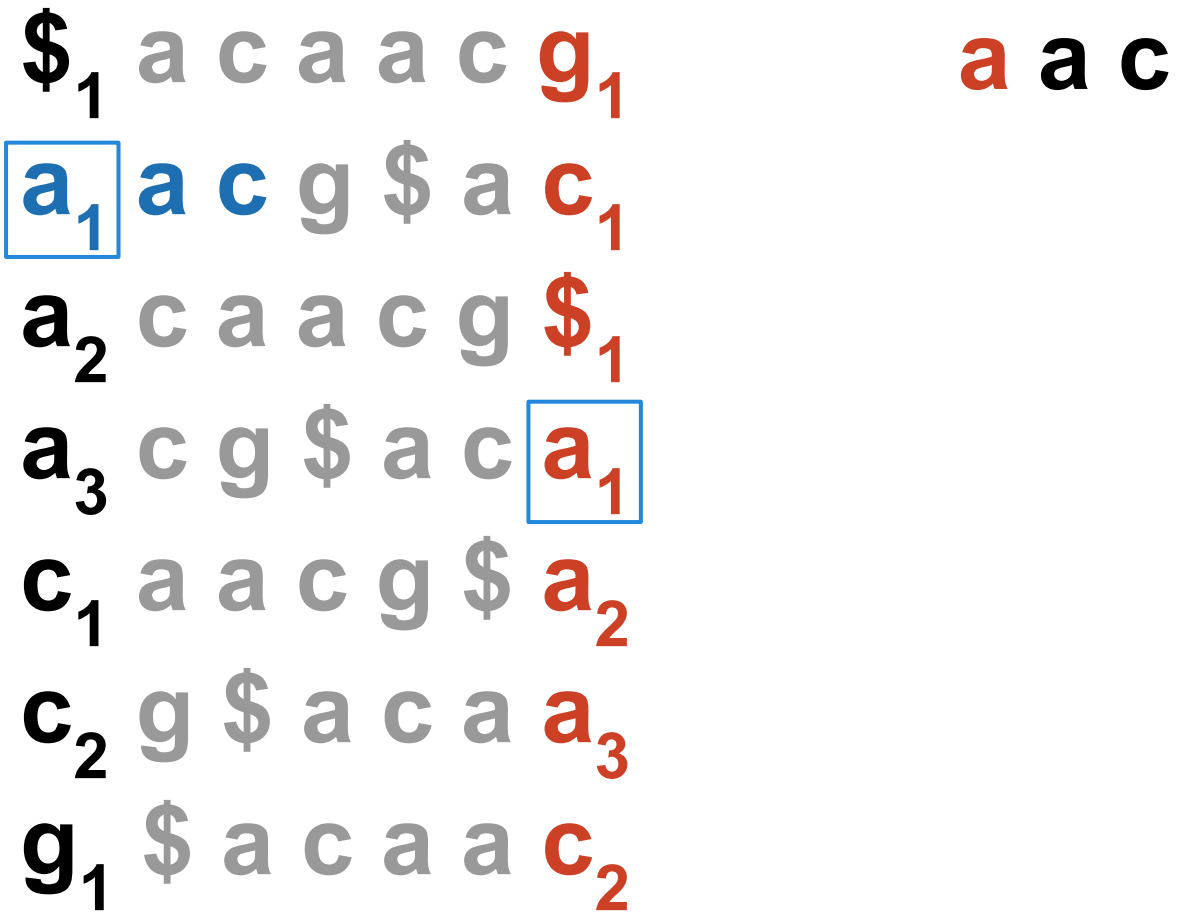
a a c

Pattern search



a a c

Pattern search



Pattern search

$_1$ a c a a c g$_1$
a$_1$ a c g $ a c$_1$
a$_2$ c a a c g $$_1$
a$_3$ c g $ a c a$_1$
c$_1$ a a c g $ a$_2$
c$_2$ g $ a c a a$_3$
g$_1$ $ a c a a c$_2$

a a c

Pattern search

For how long does it work?

Pattern search

In first column find index of a letter with given rank

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c$_1$

a$_2$ c a a c g $$_1$

a$_3$ c g $ a c a$_1$

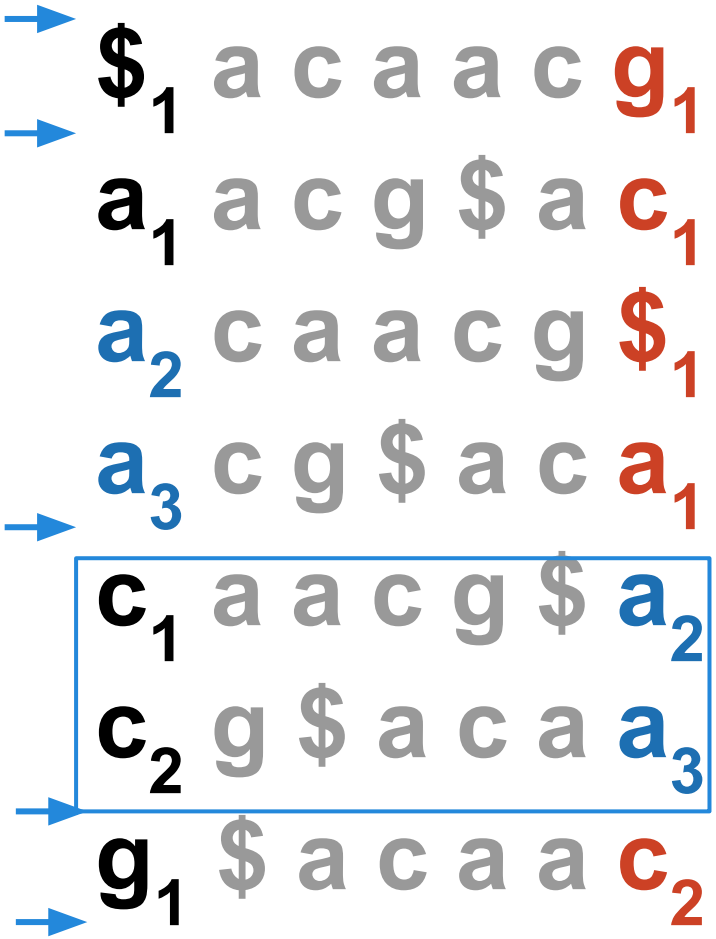
c$_1$ a a c g $ a$_2$

c$_2$ g $ a c a a$_3$

g$_1$ $ a c a a c$_2$

Pattern search

In first column find index of a letter with given rank



Pattern search

In first column find index of a letter with given rank — O(1)

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c$_1$

a$_2$ c a a c g $$_1$

a$_3$ c g $ a c a$_1$

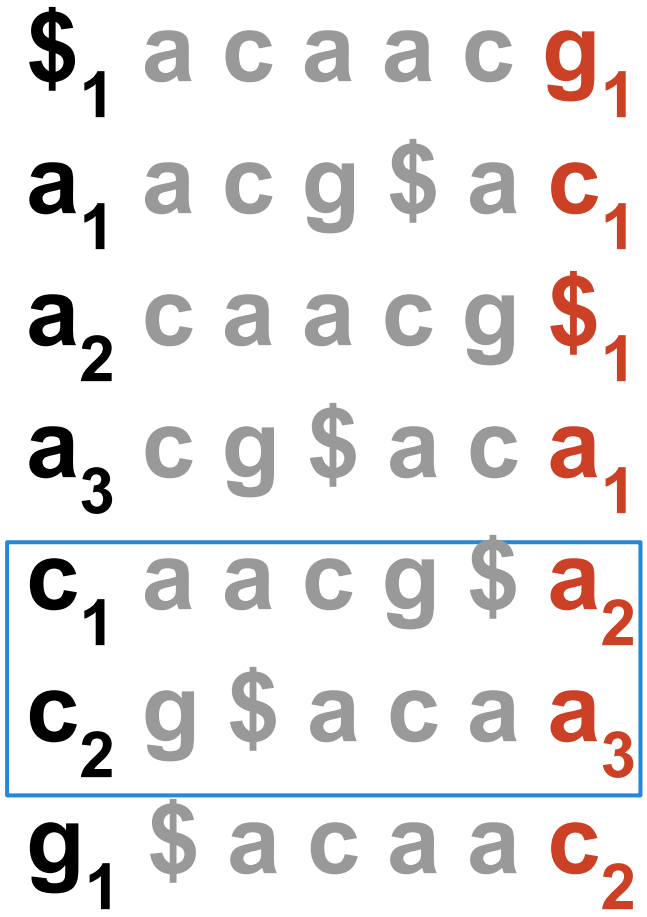
c$_1$ a a c g $ a$_2$

c$_2$ g $ a c a a$_3$

g$_1$ $ a c a a c$_2$

Pattern search

In BWT find all given letters in specified range



Pattern search

In BWT find a number of given letter above a certain row

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c$_1$

a$_2$ c a a c g $$_1$

a$_3$ c g $ a c a$_1$

c$_1$ a a c g $ a$_2$

c$_2$ g $ a c a a$_3$

g$_1$ $ a c a a c$_2$

Pattern search

In BWT find a number of given letter above a certain row

$\$_1$ a c a a c g_1	0
a_1 a c g $ a c_1	0
a_2 c a a c g $\$_1$	0
a_3 c g $ a c a_1	1
c_1 a a c g $ a_2	2
c_2 g $ a c a a_3	3
g_1 $ a c a a c_2	3

Pattern search

In BWT find a number of given letter above a certain row

$\$_1$ a c a a c g_1	0
a_1 a c g $ a c_1	0
a_2 c a a c g $\$_1$	0
a_3 c g $ a c a_1	1
c_1 a a c g $ a_2	2
c_2 g $ a c a a_3	3
g_1 $ a c a a c_2	3

Pattern search

In BWT find a number of given letter above a certain row

$\$_1$ a c a a c g_1	0
a_1 a c g $ a c_1	0
a_2 c a a c g $\$_1$	0
a_3 c g $ a c a_1	1
c_1 a a c g $ a_2	2
c_2 g $ a c a a_3	3
g_1 $ a c a a c_2	3

Pattern search

In BWT find a number of given letter above a certain row

$\$_1$ a c a a c g_1	0
a_1 a c g $ a c_1	1
a_2 c a a c g $\$_1$	1
a_3 c g $ a c a_1	1
c_1 a a c g $ a_2	1
c_2 g $ a c a a_3	1
g_1 $ a c a a c_2	2

Pattern search

In BWT find a number of given letter above a certain row

$$\$_1$$	a c a a c $$g_1$$	0
$$a_1$$	a c g $ a $$c_1$$	1
$$a_2$$	c a a c g $$\$_1$$	1
$$a_3$$	c g $ a c $$a_1$$	1
$$c_1$$	a a c g $ $$a_2$$	1
$$c_2$$	g $ a c a $$a_3$$	1
$$g_1$$	$ a c a a $$c_2$$	2

Pattern search

In BWT find a number of given letter above a certain row

$\$_1$ a c a a c g_1	0
a_1 a c g $ a c_1	1
a_2 c a a c g $\$_1$	1
a_3 c g $ a c a_1	1
c_1 a a c g $ a_2	1
c_2 g $ a c a a_3	1
g_1 $ a c a a c_2	2

Pattern search

In BWT find a number of given letter above a certain row — O(1)

$_1$ a c a a c g$_1$

a$_1$ a c g $ a c$_1$

a$_2$ c a a c g $$_1$

a$_3$ c g $ a c a$_1$

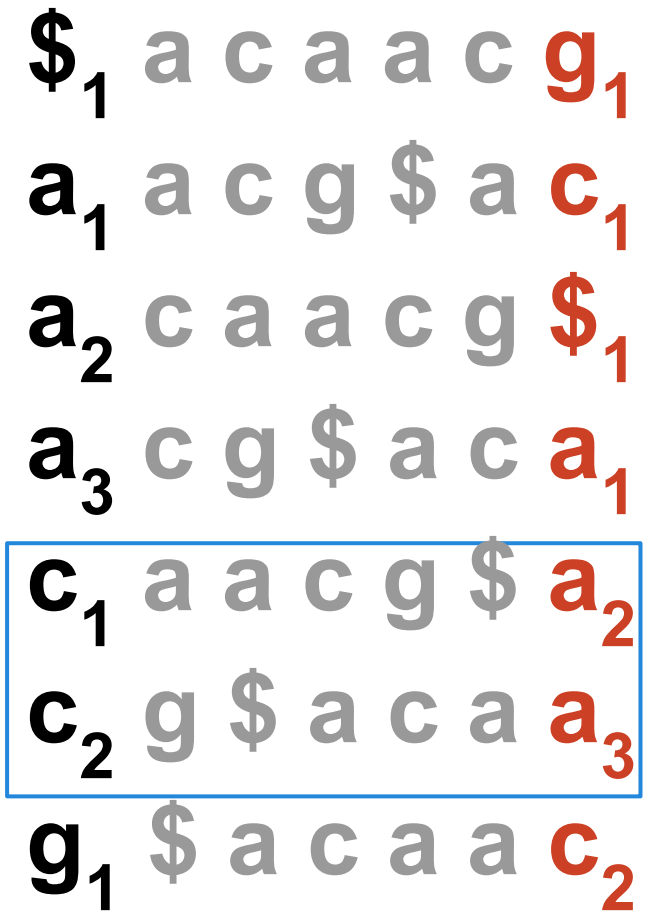
c$_1$ a a c g $ a$_2$

c$_2$ g $ a c a a$_3$

g$_1$ $ a c a a c$_2$

Pattern search

In BWT find all given letters in specified range — O(1)



Pattern search

Pattern search works in linear time

Thank you!

Questions?