ЛОКАЛНО ПОДРАВНЯВАНЕ НА БИОЛОГИЧНИ СЕКВЕНЦИИ

ПРОФ. ПЛАМЕНКА БОРОВСКА

ЛОКАЛНО ПОДРАВНЯВАНЕ НА БИОЛОГИЧНИ СЕКВЕНЦИИ

- ЦЕЛ – локалното подравняване на две секвенции от нуклеотиди или аминокиселини се прави с цел да се установи структурно и функционално сходство
- Най-често задаваният въпрос в молекулярната биология е дали две дадени секвенции са свързани или не, за да се идентифицира тяхната структура или функция. Най-простият начин да се отговори на този въпрос е да се сравнят техните секвенции.
- Подравняването на секвенции или сравняването на секвенции е една от фундаменталните области в биоинформатиката, която изследва начините за подреждане на генетичните (ДНК/РНК) секвенции или секвенциите от протеини, *с цел определяне на сходните участъци.*
- *Прилага се за да се установи съществува ли структурна, функционална и еволюционна връзка между секвенциите.*
- Когато се открие нова секвенция, структурата и функцията могат лесно да се предвидят, като се направи подравняване с избрани секвенции от биологичните бази данни.
- Когато подравняваните секвенции споделят общ прародител се приема, че новата секвенция би проявила подобна структура или функция.
- Колкото по-голямо е сходството между секвенциите, толкова по-голям е шансът те да споделят подобна структура или функция.

АЛГОРИТЪМ НА SMITH - WATERMAN

- Цел – да се открие нивото на сходство между две секвенции
- Изследваната секвенция се подава като заявка (query sequence) и тя се сравнява с различни секвенции от биологичните бази данни.
- **Локално подравняване (Local Alignment):** прилага се за да се открият сходствата или разликите между две секвенции - идентифицират се локални участъци с висока степен на сходство.
- Алгоритъмът на Smith Waterman за локално подравняване на две секвенции е предложен от Temple F. Smith и Michael S. Waterman през 1981 г.
- *Представлява вариант на алгоритъма на Needleman–Wunsch*, и аналогично се основава на алгоритъма на динамичното програмиране като гарантира откриването на оптималното локално подравняване по отношение на използваната оценъчна функция - матрица на заместване (substitution matrix) и схемата за оценка на празнините (gap-scoring scheme)

АЛГОРИТЪМ НА SMITH - WATERMAN

- *Основната разлика с алгоритъма на Needleman–Wunsch* е, че елементите на оценъчната матрица с отрицателни стойности се приравняват на 0, което откроява и прави видими локалните подравнявания с положителна оценка
- Процедурата за обратното проследяване (Traceback) стартира от елемента на оценъчната матрица с най-високата стойност и напредва следвайки указателите до достигането на елемент с нулева стойност, при което се получава оптималното локално подравняване.
- Поради своята квадратична сложност по отношение на времето и пространството, алгоритъмът често не може да бъде практически приложен към мащабни проблеми
- *Предимства на алгоритъмът на Smith-Watermen*:

1. дава възможност да се открият консервативни (запазени, непроменени) участъци в двете секвенции (conserved regions)
2. Може да се подравняват две частично припокриващи се секвенции, както и да се подравнява подсеквенция с цяла секвенция (да се провери дали даден участък от едната секвенция присъства и в другата секвенция)

СРАВНЕНИЕ НА АЛГОРИТМИТЕ NEEDLEMAN–WUNSCH И SMITH WATERMAN

- Алгоритъмът на Smith-Waterman намира сходните сегменти в две секвенции, докато алгоритъмът на Needleman – Wunsch подравнява две цялостни секвенции.

- И двата алгоритъма използват концепциите за матрица на заместване, функция за наказание за празнина, оценъчна матрица и процес за обратно проследяване.

	Алгоритъм на Smith–Waterman	Алгоритъм на Needleman–Wunsch
Инициализация на оценъчната матрица	Първият ред и първата колона на оценъчната матрица се запълват с нулеви елементи	Първият ред и първата колона на оценъчната матрица се запълват със съответните стойности на наказанията за празнини (gap penalty)
Схема за оценка	Елементите на оценъчната матрица с отрицателни стойности се нулират	Допускат се отрицателни стойности на елементите на оценъчната матрица
Процедура за обратно проследяване	Стартира от елемента на оценъчната матрица с максимална стойност и терминира при достигане на елемент с нулева стойност	Стартира от елемента в най-долния десен ъгъл на оценъчната матрица и терминира при достигане на елемента в най-горния ляв ъгъл на оценъчната матрица

СРАВНЕНИЕ НА АЛГОРИТМИТЕ NEEDLEMAN–WUNSCH И SMITH WATERMAN

- Едно от най-важните разлики е, че в системата за оценяване на алгоритъма на Smith–Waterman не се допуска отрицателна стойност на елемент на оценъчната матрица, което дава възможност за локално подравняване.
- Когато някой елемент на оценъчната матрица има отрицателна стойност, това означава, че секвенциите до тази позиция нямат сходства.
- С цел елиминиране на влиянието на предходното подравняване, при алгоритъма на Smith–Waterman елементите с отрицателни стойности се нулират и така се дава възможност да се откриват възможни подрежданията във всички посоки
- Спецификата на инициализацията на оценъчната матрица при алгоритъма на Smith–Waterman дава възможност за подравняване на всеки сегмент от една секвенция към произволна позиция в другата секвенция.
- В алгоритъма на Needleman – Wunsch, обаче, трябва да се взема предвид и наказанието за празнина в последната позиция на секвенцията при подравняването на целите секвенции

СИСТЕМИ ЗА ОЦЕНКА

- В оптимизиращите процедури за подравняване алгоритмите на Needleman-Wunsch и Smith-Waterman използват система за оценка.
- За подравняване на нуклеотидни (ДНК/РНК) секвенции използваните матрици за оценка са сравнително по-прости, тъй като честотата на мутацията за всички нуклеотиди е еднаква.
- Положителна или по-висока стойност се задава за съвпадение, а отрицателна или по-ниска стойност се присвоява за несъответствие. Тези резултати, базирани на предположения, могат да се използват за оценка на матриците.
- Има и други матрици за оценка, които са предефинирани, и се използват в случай на подравняване на секвенции от аминокиселини.
- PAM Matrices: Margaret Dayhoff е първата, разработила PAM матрицата, като PAM означава “приети точкови мутации” (Point Accepted Mutations).
- PAM матриците се изчисляват, като се отчитатат разликите в тясно свързани протеини. Една PAM единица (PAM1) указва една приета точкова мутация на 100 аминокиселинни остатъка, т.е. 1% промяна и 99% остава като такава.

СИСТЕМИ ЗА ОЦЕНКА

- BLOSUM: BLOcks SUbstitution Matrix, предложена от Henikoff и Henikoff през 1992 г., базирана на консервативни (запазени) участъци (conserved regions). Тези матрици представят действителни стойности на идентичност в проценти или казано просто, *представят сходствата в %.* Blosum 62 означава, че сходството между двете секвенции е 62 %.

- Наказание за празнини (gap penalty): алгоритмите на динамичното програмиране прилагат наказания за празнини с цел да се максимизира биологичният смисъл. Наказанието за всяка празнина се изважда от оценката, която е възприета.

- Оценката за празнините определя наказание за подравняване, в което има вмъкване на нуклеотид (insertion) или изтриване на нуклеотид (deletion).

- В хода на еволюцията, има случаи, в които се наблюдават дълги поредици от празнини в рамките на секвенцията, и следователно, в случая линейното наказание за празнини не е подходящо. В тези случаи се прилага *наказание за начална празнина (gap open) на поредица от празнини и наказание за разтегляне на секвенцията (gap extension) за последващите празнини – при 5 празнини или повече*.

НАКАЗАНИЯ ЗА ПРАЗНИНИ

- Наказанието *gap open* винаги се прилага за първата празнина от поредицата празнини, като последващите празнини се наказват за разтегляне на секвенцията (gap extension).
- Наказанието за разтегляне е винаги по-малко от наказанието за първата празнина от поредицата.
- Типични стойности –12 за gap opening, и –4 за gap extension.
- Целта е да се подпомогне избягването на разпръснати малки празнини
- В биологичен аспект резултатът от подравняването трябва да се анализира по различен начин от практически съображения. От една страна, частичното сходство между две секвенции е често срещано явление; от друга страна, възникването на мутацията в един ген може да доведе до дълга поредица от празнини.
- Следователно, свързаните празнини, образуващи дълга поредица от празнини, обикновено са по-благоприятни от наличието на множеството разпръснати празнини.
- GAPpenalty = v + ku, където v е наказанието open gap, а u е наказанието за разтегляне (удължаване) на секвенцията

EMBOSS: THE EUROPEAN MOLECULAR BIOLOGY OPEN SOFTWARE SUITE HTTP://EMBOSS.OPEN-BIO.ORG/

```
TACGGGCCCGCTAC

TAGCCCTATCGGTCA
```

Оценъчна матрица за нуклеотиди DNAfull

EMBOSS Matcher

```
TACGGGCCCGCTA-C
||   | || | || |
TA---G-CC-CTATC
```

Линейно наказание за празнини

```
TACGGGCCCGCTA
||   |  |  |||
TA---G--CCCTA
```

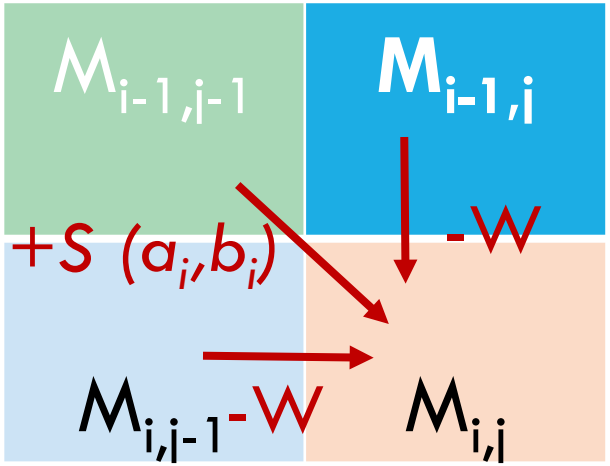
Gap opening
Gap extension
End gap penalty
End Gap Open penalty
End gap extend penalty

Помага за избягването на разпръснати малки празнини

https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/EMBOSS+Matcher+Help+and+Documentation

АЛГОРИТЪМ НА SMITH - WATERMAN

$$M_{i,j} = MAX \begin{cases} M_{i-1,j-1} + S_{i,j} \\ M_{i,j-1} + W \\ M_{i-1,j} + W \\ 0 \end{cases}$$



Където: i,j определят реда и колоната

$M_{i,j}$ - елемент на оценъчната матрица

$S_{i,j}$ е оценката

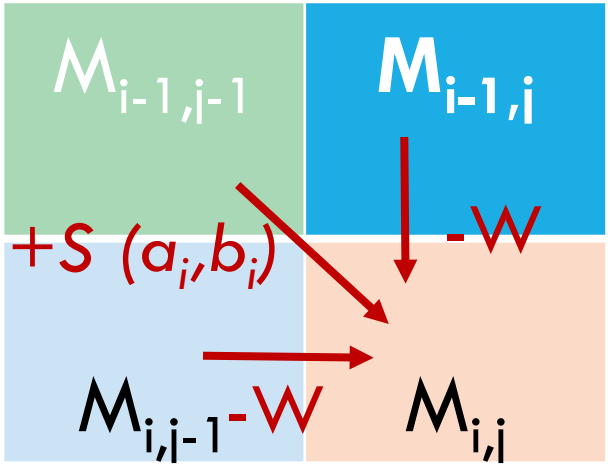
W – подравняване с празнина (gap alignment)

АЛГОРИТЪМ НА SMITH WATERMAN

1. Инициализация на оценъчната матрица

2. Запълване на оценъчната матрица с оценките

$$M_{i,j} = Maximum\left[M_{i-1,j-1} + S_{i,j},\ M_{i,j-1} + W,\ M_{i-1,j} + W,\ 0\right]$$



Където: i,j определят реда и колоната

$M_{i,j}$ - елемент на оценъчната матрица

$S_{i,j}$ е оценката

W – подравняване с празнина (gap alignment)

3. Обратно проследяване в оценъчната матрица за откриване на оптималното подравняване

Пример: CGTGAATTCAT (sequence#1 - A)

GACTTAC (sequence #2 - B)

Match score +5

Mismatch score -3

Gap penalty W=-4

1. Инициализация на оценъчната матрица

2. Запълване на оценъчната матрица с оценките и маркиране на указателите за връщане (1 или повече) сочещ/и клетката/клетките, от която/които е получена MAX стойност

$M_{0,0} = 0$; $M_{1,0} = 0$; $M_{0,1} = 0$;

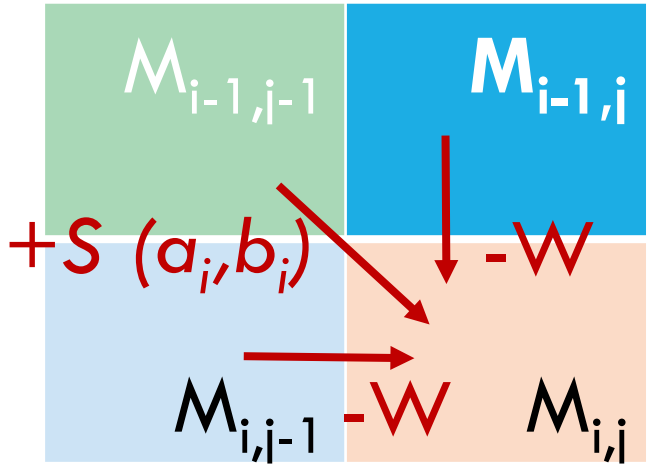
$S_{1,1}$ (mismatch) = -3

W=-4

$M_{1,1} = MAX [(M_{0,0} + S_{1,1}), (M_{1,0}+W), (M_{0,1}+W), 0]$

$M_{1,1} = MAX [(0-3), (0-4), (0-4), 0]$

$M_{1,1} = MAX [-3, -4, -4, 0] = 0$



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
A	0											
C	0											
T	0											
T	0											
A	0											
C	0											

ОЦЕНЪЧНАТА МАТРИЦА Е ЗАПЪЛНЕНА С ОЦЕНКИТЕ И УКАЗАТЕЛИТЕ ЗА ОБРАТНО ПРОСЛЕДЯВАНЕ (TRACEBACK POINTERS)

	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	10	6	2	0	0	5	1
C	0	5	1	0	0	6	7	3	0	5	1	2
T	0	1	2	6	2	2	3	12	8	4	2	6
T	0	0	0	7	3	0	0	8	17	13	9	7
A	0	0	0	3	4	8	5	4	13	14	18	14
C	0	5	1	0	0	4	5	2	9	18	14	15

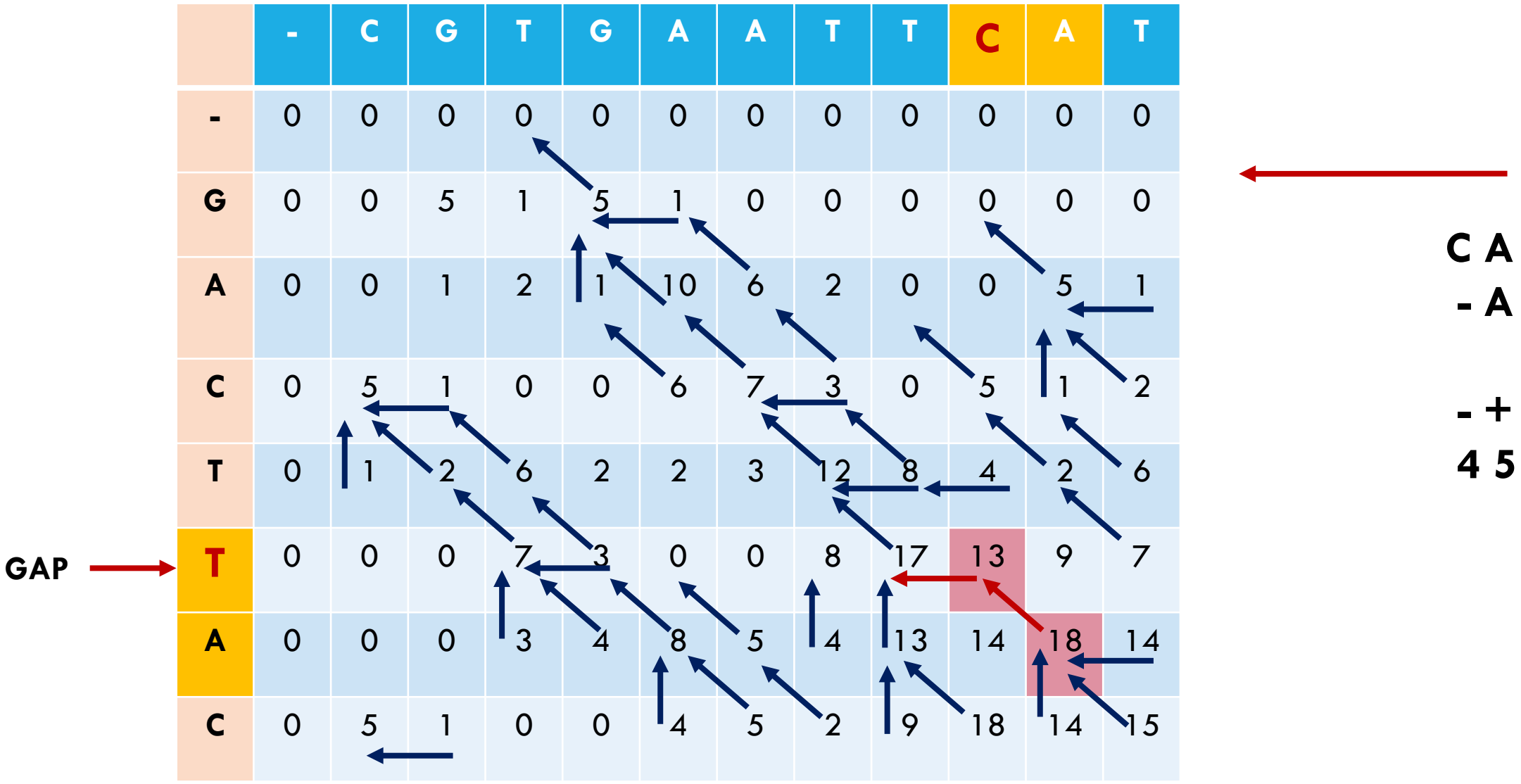
ОБРАТНО ПРОСЛЕДЯВАНЕ (TRACEBACK)

	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	10	6	2	0	0	5	1
C	0	5	1	0	0	6	7	3	0	5	1	2
T	0	1	2	6	2	2	3	12	8	4	2	6
T	0	0	0	7	3	0	0	8	17	13	9	7
A	0	0	0	3	4	8	5	4	13	14	**18**	14
C	0	5	1	0	0	4	5	2	9	18	14	15

A
A

+
5

ОБРАТНО ПРОСЛЕДЯВАНЕ (TRACEBACK)



ОБРАТНО ПРОСЛЕДЯВАНЕ (TRACEBACK)

	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	10	6	2	0	0	5	1
C	0	5	1	0	0	6	7	3	0	5	1	2
T	0	1	2	6	2	2	3	12	8	4	2	6
T GAP	0	0	0	7	3	0	0	8	**17**	**13**	9	7
A	0	0	0	3	4	8	5	4	13	14	**18**	14
C	0	5	1	0	0	4	5	2	9	18	14	15

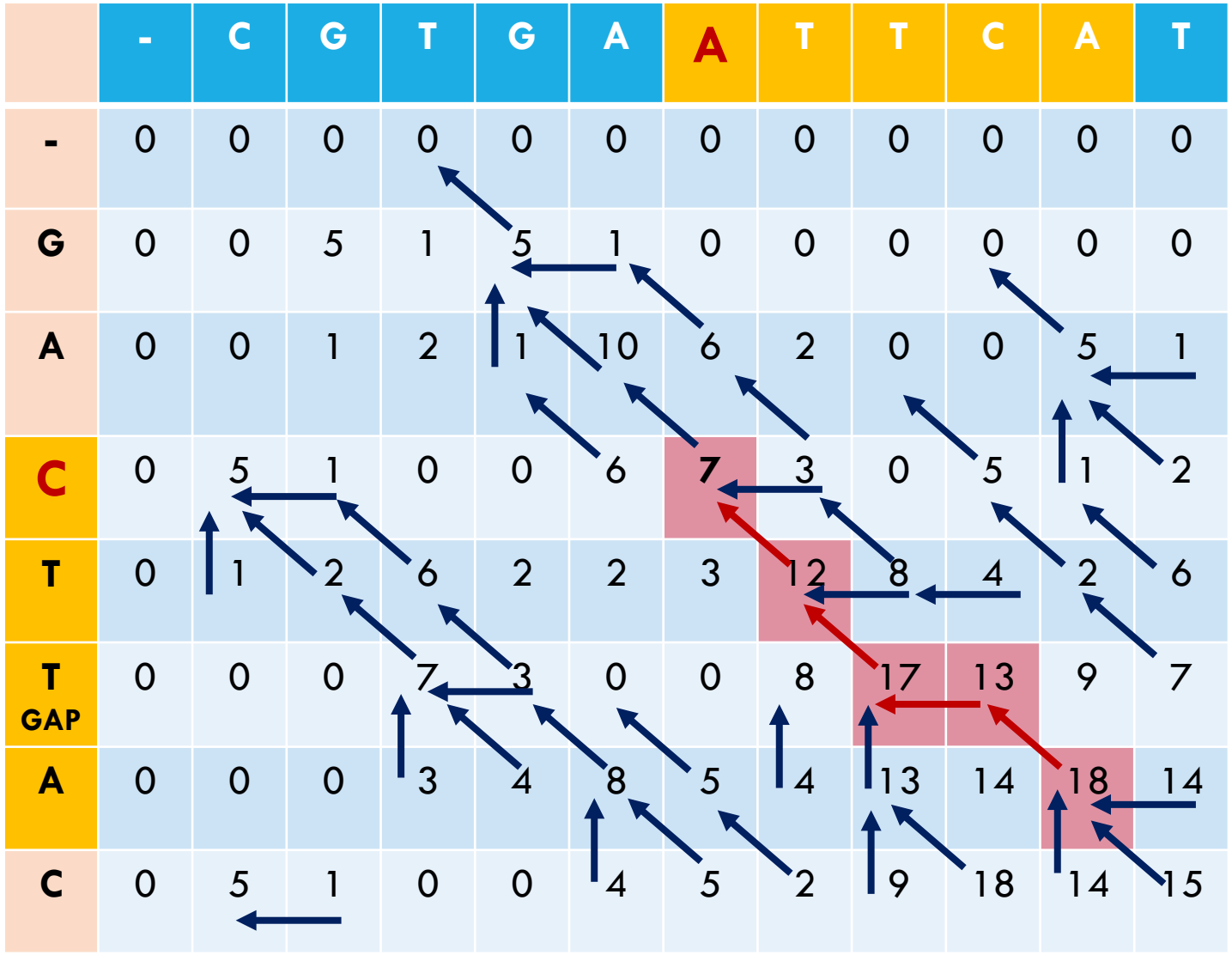
T C A
T - A
+ - +
5 4 5

ОБРАТНО ПРОСЛЕДЯВАНЕ (TRACEBACK)

	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	10	6	2	0	0	5	1
C	0	5	1	0	0	6	7	3	0	5	1	2
T	0	1	2	6	2	2	3	12	8	4	2	6
T GAP	0	0	0	7	3	0	0	8	17	13	9	7
A	0	0	0	3	4	8	5	4	13	14	18	14
C	0	5	1	0	0	4	5	2	9	18	14	15

T T C A
T T - A
+ + - +
5 5 4 5

ОБРАТНО ПРОСЛЕДЯВАНЕ (TRACEBACK)



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	10	6	2	0	0	5	1
C	0	5	1	0	0	6	7	3	0	5	1	2
T	0	1	2	6	2	2	3	12	8	4	2	6
T GAP	0	0	0	7	3	0	0	8	17	13	9	7
A	0	0	0	3	4	8	5	4	13	14	18	14
C	0	5	1	0	0	4	5	2	9	18	14	15

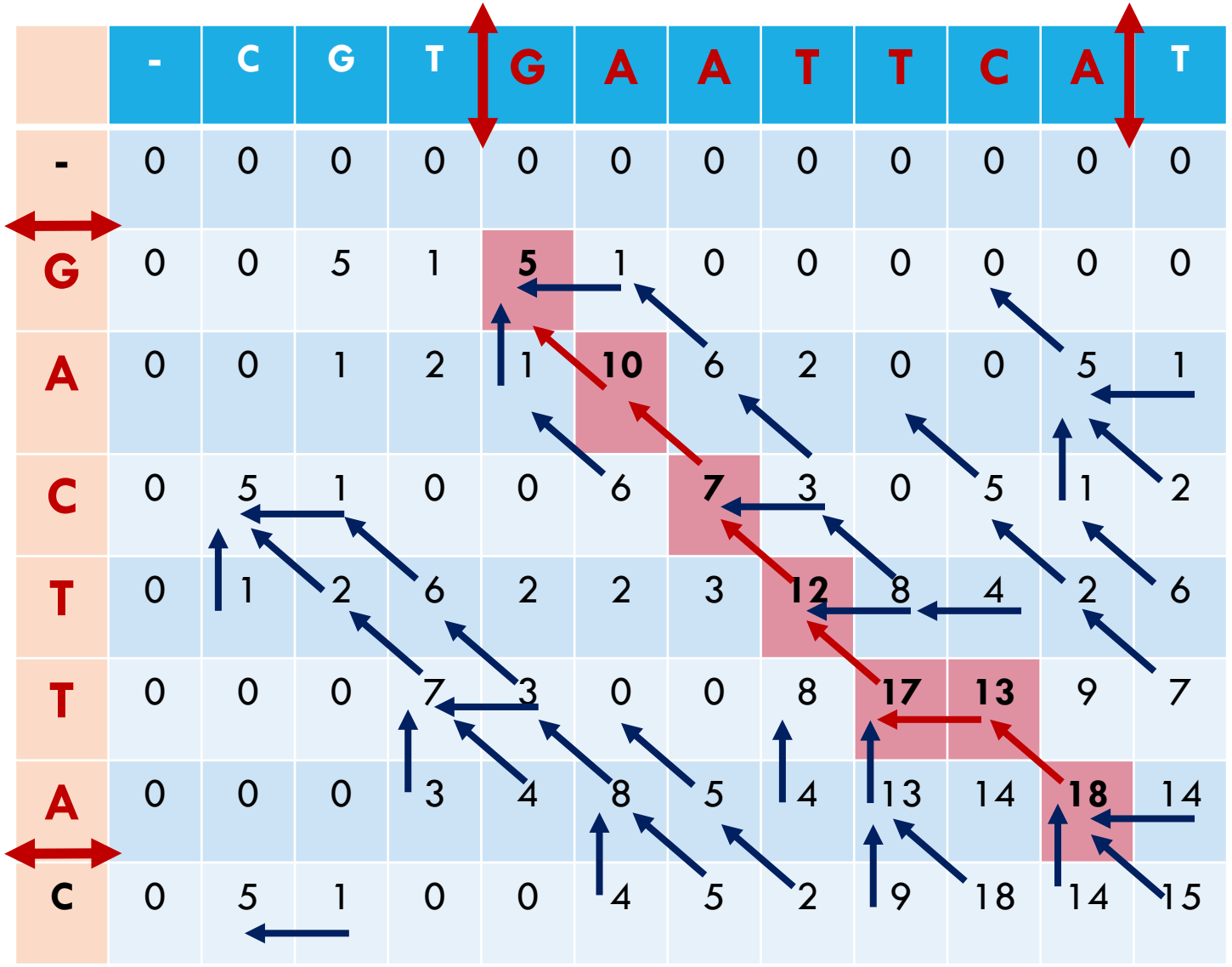
A T T C A
C T T - A
- + + - +
3 5 5 4 5

ОБРАТНО ПРОСЛЕДЯВАНЕ (TRACEBACK)

	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	**10**	6	2	0	0	5	1
C	0	5	1	0	0	6	**7**	3	0	5	1	2
T	0	1	2	6	2	2	3	**12**	8	4	2	6
T GAP	0	0	0	7	3	0	0	8	**17**	**13**	9	7
A	0	0	0	3	4	8	5	4	13	14	**18**	14
C	0	5	1	0	0	4	5	2	9	18	14	15

A A T T C A
A C T T - A
+ - + + - +
5 3 5 5 4 5

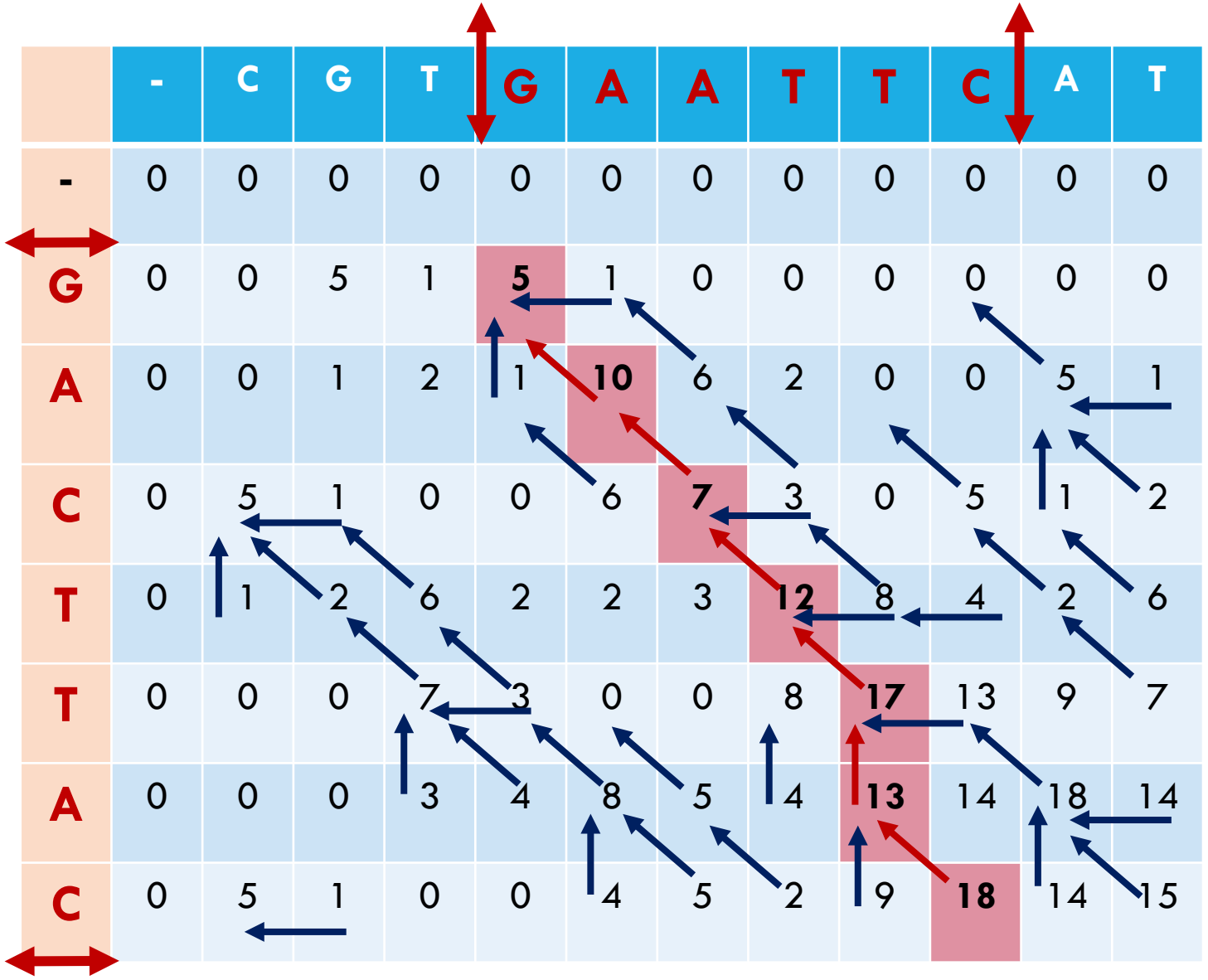
Обратното проследяване стартира от елемента с най-висока оценка, следвайки указателите до достигането на елемент с оценка 0



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	**5**	1	0	0	0	0	0	0
A	0	0	1	2	1	**10**	6	2	0	0	5	1
C	0	5	1	0	0	6	**7**	3	0	5	1	2
T	0	1	2	6	2	2	3	**12**	8	4	2	6
T	0	0	0	7	3	0	0	8	**17**	**13**	9	7
A	0	0	0	3	4	8	5	4	13	14	**18**	14
C	0	5	1	0	0	4	5	2	9	18	14	15

G A A T T C A
G A C T T - A
+ + - + + - +
5 5 3 5 5 4 5
Обща оценка 18

ВЪЗМОЖНО Е ОТ ДАДЕН ЕЛЕМЕНТ ДА ИМА 2 УКАЗАТЕЛЯ. В ТОЗИ СЛУЧАЙ МОГАТ ДА СЕ РАЗГЛЕДАТ ДВЕТЕ ПОДРАВНЯВАНИЯ



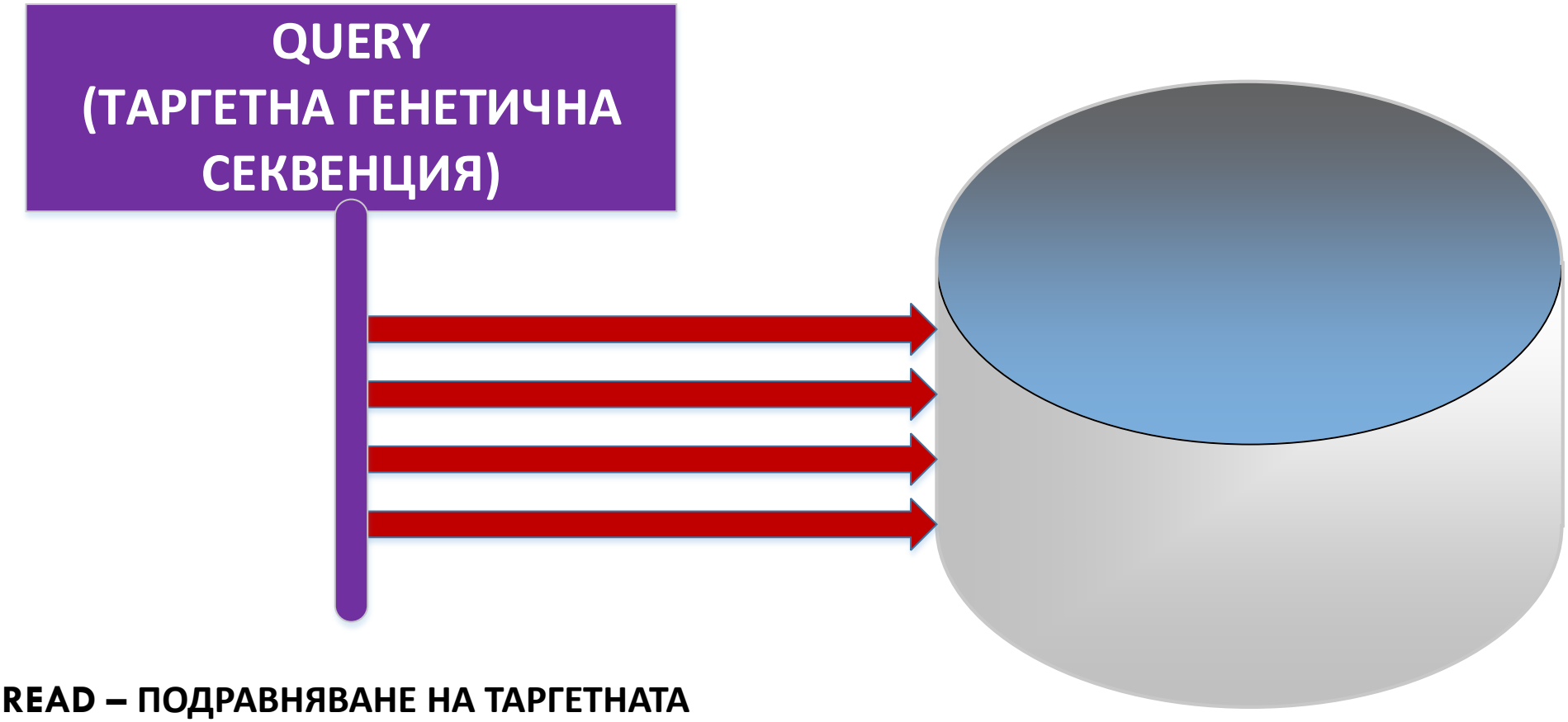
	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	**5**	1	0	0	0	0	0	0
A	0	0	1	2	1	**10**	6	2	0	0	5	1
C	0	5	1	0	0	6	**7**	3	0	5	1	2
T	0	1	2	6	2	2	3	**12**	8	4	2	6
T	0	0	0	7	3	0	0	8	**17**	13	9	7
A	0	0	0	3	4	8	5	4	**13**	14	18	14
C	0	5	1	0	0	4	5	2	9	**18**	14	15

G A A T T - C
G A C T T A C
+ + - + + - +
5 5 3 5 5 4 5
Обща оценка 18

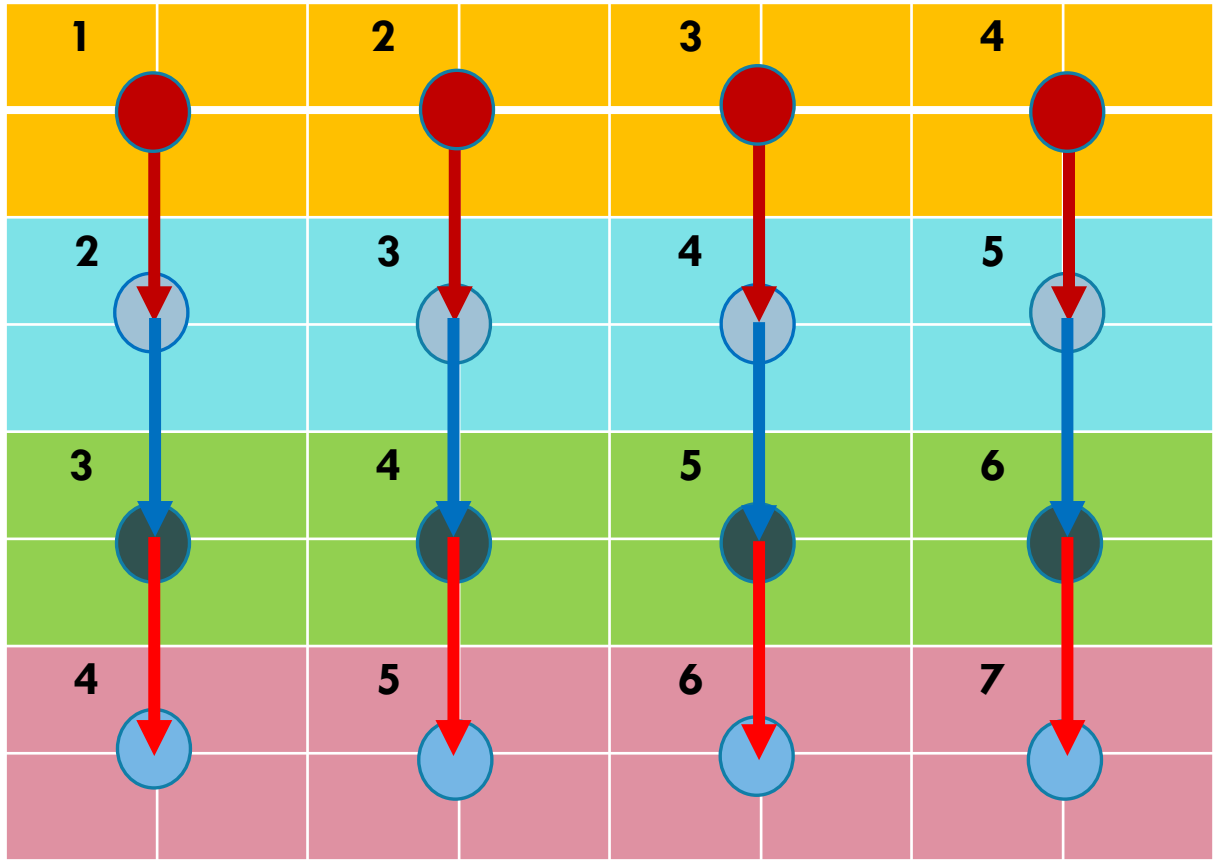
ПРАКТИЧЕСКИ АСПЕКТИ

- Търсене в биологични бази данни
- Заявка (Query) - Открита нова генетична секвенция (ДНК/РНК) или избрана от биологичните бази данни такава
- Глобално подравняване със секвенции в биологичните бази – алгоритъм на Needlemen – Wunch – по две секвенции
- Локално подравняване със секвенции в биологичните бази – алгоритъм на Smith – Waterman

ЛОКАЛНО ПОДРАВНЯВАНЕ СЪС СЕКВЕНЦИИ В БИОЛОГИЧНИТЕ БАЗИ – АЛГОРИТЪМ НА SMITH – WATERMAN

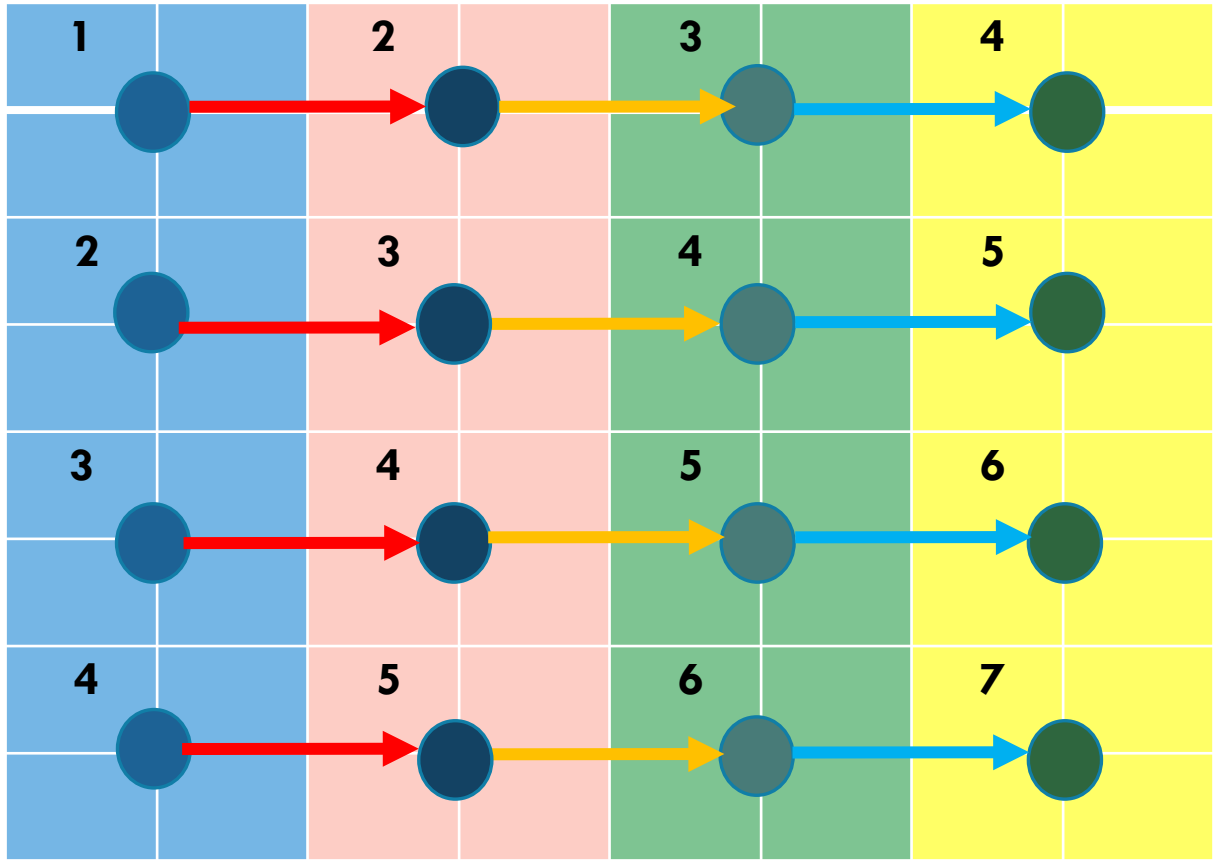


1 THREAD – ПОДРАВНЯВАНЕ НА ТАРГЕТНАТА СЕКВЕНЦИЯ С НЯКОЛКО СЕКВЕНЦИИ ОТ БАЗАТА ДАННИ



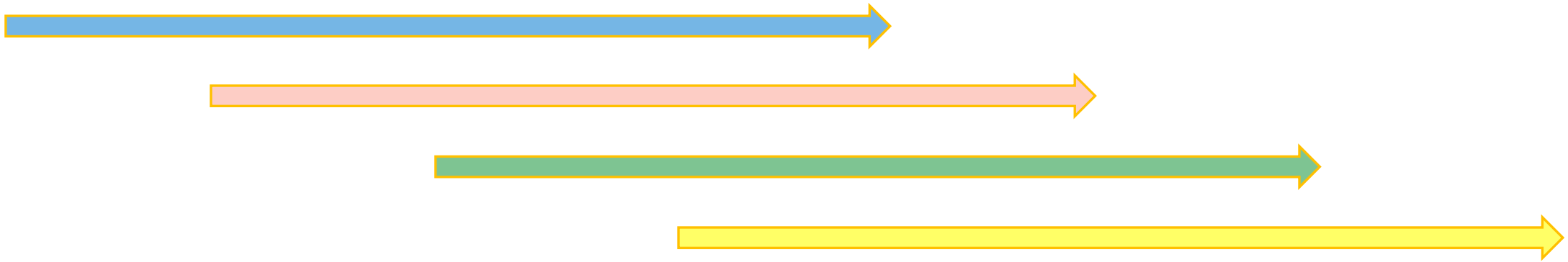
ПАРАЛЕЛИЗАЦИЯ ПРИ ПОДРАВНЯВАНЕТО НА ДВЕ СЕКВЕНЦИИ: ПО РЕДОВЕ С НИШКИ КОНВЕЙЕРИЗАЦИЯ (PIPELINE)

ОБРАБОТКАТА НА АНТИДИАГОНАЛНИТЕ ЕЛЕМЕНТИ НА МАТРИЦАТА Е НЕЗАВИСИМО



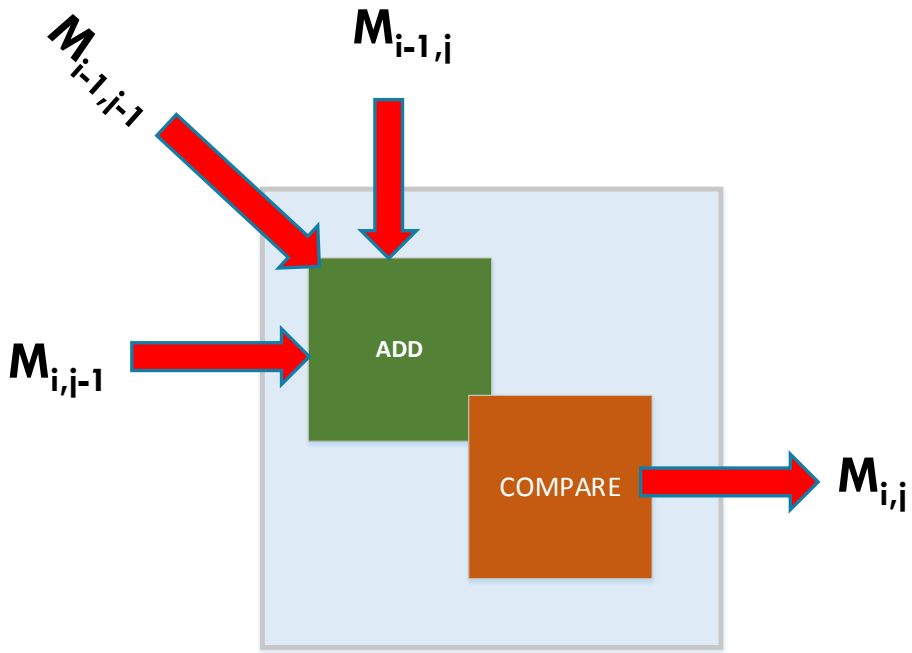
ПАРАЛЕЛИЗАЦИЯ ПРИ ПОДРАВНЯВАНЕТО НА ДВЕ СЕКВЕНЦИИ: ПО КОЛОНИ С НИШКИ КОНВЕЙЕРИЗАЦИЯ (PIPELINE)

ОБРАБОТКАТА НА АНТИДИАГОНАЛНИТЕ ЕЛЕМЕНТИ НА МАТРИЦАТА Е НЕЗАВИСИМО



ИМПЛЕМЕНТАЦИЯ С FPGA ЗА ОБЛАЧНИ УСЛУГИ – **SYSTOLIC ARRAY**

	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0		5	1	5	1	0	0	0	0	0	0
A	0	0			1	10	6	2	0	0	5	1
C	0	5	1	0			7	3	0	5	1	2
T	0	1	2	6	2	2		12	8	4	2	6
T	0	0	0	7	3	0	0		13	9		7
A	0	0	0	3	4	8	5	4	13		18	14
C	0	5	1	0	0	4	5	2	9	18		15



ФРОНТ НА ВЪЛНАТА (WAVEFRONT)

ПОСОКА НА АКТИВИРАНЕ

ПРАКТИЧЕСКИ АСПЕКТИ

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is the name given to the 2019 novel coronavirus.

COVID-19 is the name given to the disease associated with the virus.

SARS-CoV-2 is a new strain of coronavirus that has not been previously identified in humans.

29,882 bp linear RNA

https://www.ncbi.nlm.nih.gov/nuccore/?term=txid2697049%5bOrganism:noexp%5d

https://www.ncbi.nlm.nih.gov/nuccore/MT276326.1?report=fasta