

Uczenie maszynowe w bioinformatyce

Wykład 4: sekwencjonowanie de novo genomu organizmu (postać danych)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Sekwencjonowanie genomu organizmu

Przez sekwencjonowanie genomu organizmu rozumie się ustalenie kolejności nukleotydów całości jego materiału genetycznego. W przypadku człowieka oznacza to poznanie sekwencji DNA tworzącego genomu: jądrowy (ponad 3,2 mld par zasad w 22 autosomach i obu chromosomach płci) oraz mitochondrialny (około 16,5 kbp).

Główną przeszkodą na drodze do poznania pełnej sekwencji nieznanego genomu (czyli wykonania tzw. sekwencjonowania *de novo*) jest dysproporcja między jego rozmiarem a możliwościami technik sekwencjonowania.

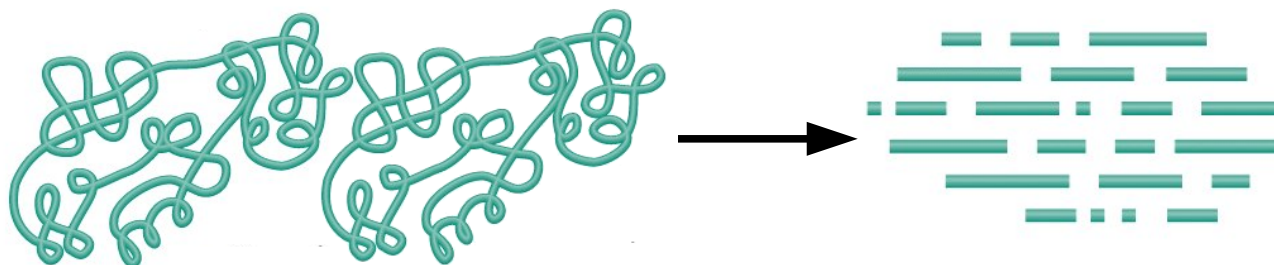
Nawet w przypadku prostych organizmów nie jest możliwe zsekwencjonowanie całego genomu w pojedynczej reakcji: najmniejsze genomy bakteryjne mają ponad 500 tysięcy par zasad (kbp) podczas gdy dostępne techniki pozwalają jednorazowo określić sekwencje cząstek DNA o zdecydowanie mniejszej długości (do 1 kbp dla metody Sangera lub od kilkudziesięciu do kilkuset bp dla technik nowej generacji).

Aby ominąć ten problem podczas sekwencjonowania genomów wykorzystuje się różne odmiany tzw. **strategii *shotgun***, w której **całkowita sekwencja składana jest z dużego zbioru, częściowo zachodzących na siebie fragmentów.**

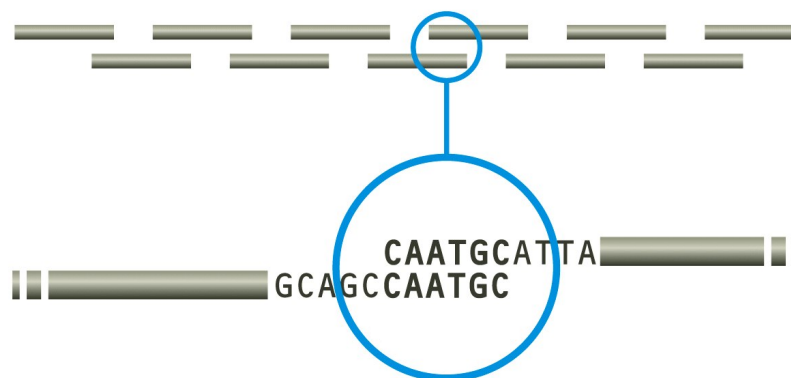
Sekwencjonowanie genomu organizmu: strategia *shotgun*

W ogólnym schemacie strategii *shotgun* można wyróżnić trzy podstawowe etapy:

- **fragmentację wielu kopii całego genomu lub osobnych chromosomów na odcinki o losowej długości** poprzez sonikację (fragmentację za pomocą ultradźwięków).



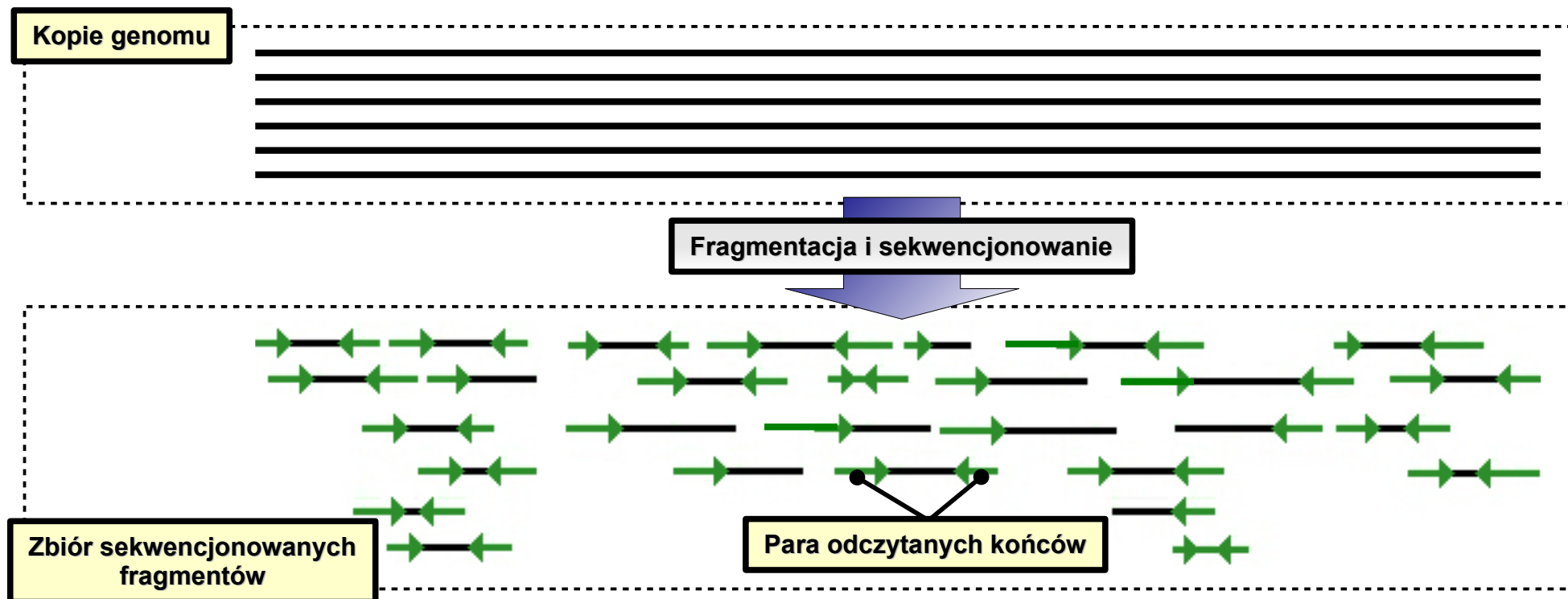
- **sekwencjonowanie** poszczególnych fragmentów;
- **składanie sekwencji genomu (*genome sequence assembly*)** ze zbioru zachodzących na siebie sekwencji fragmentów. Korzysta się przy tym z założenia, że sekwencje o jednakowej kolejności nukleotydów pochodzą z tego samego miejsca w genomie.



Sekwencjonowanie genomu organizmu: strategia *shotgun*

Charakterystyka danych z sekwencjonowania typu *shotgun*:

- **dla większości fragmentów nie znamy pełnej sekwencji**: metody sekwencjonowania pozwalają jedynie określić sekwencje ich początkowych i końcowych odcinków. Sekwencje w takich **parach odczytanych końców** mają długości mniejsze niż 1kbp. Ponadto sekwencje mogą zawierać błędy i mutacje;
- **składanie pełnej sekwencji genomu na podstawie zbioru takich odczytów wymaga wprowadzenia redundancji**: sekwencjonowana jest nadmiarowa liczba fragmentów, o całkowitej długości wielokrotnie przekraczającej rozmiar genomu.



Sekwencjonowanie genomu organizmu: strategia *shotgun*



[Drew Sheneman]

Uczenie maszynowe w bioinformatyce

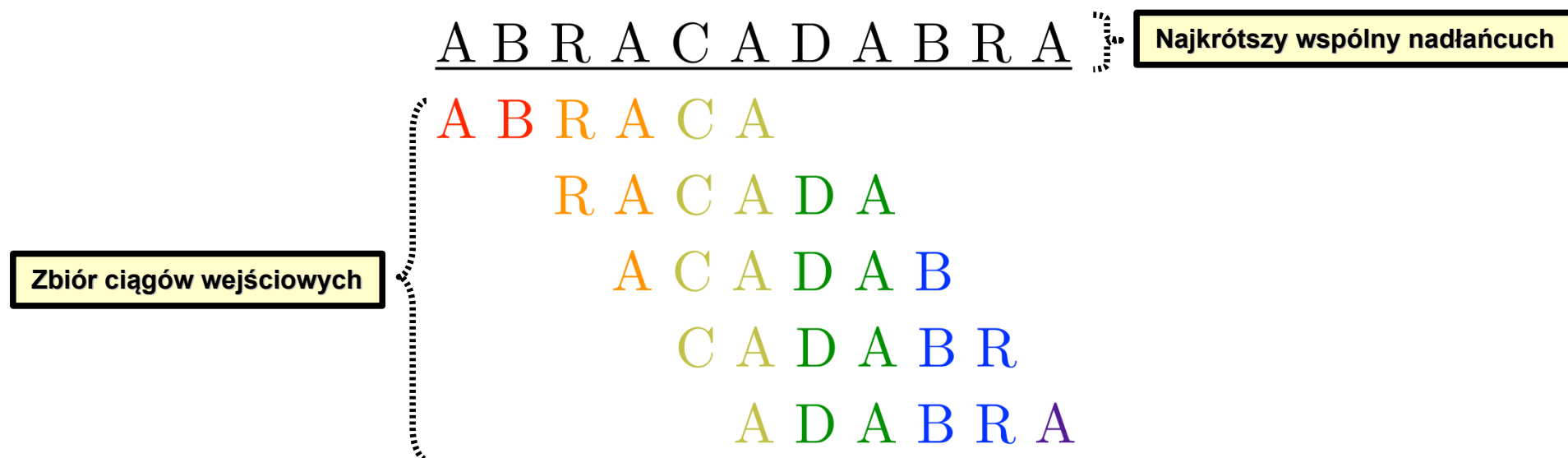
Wykład 4: sekwencjonowanie de novo genomu organizmu (składanie sekwencji)

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Składanie sekwencji genomu w sekwencjonowaniu *shotgun*

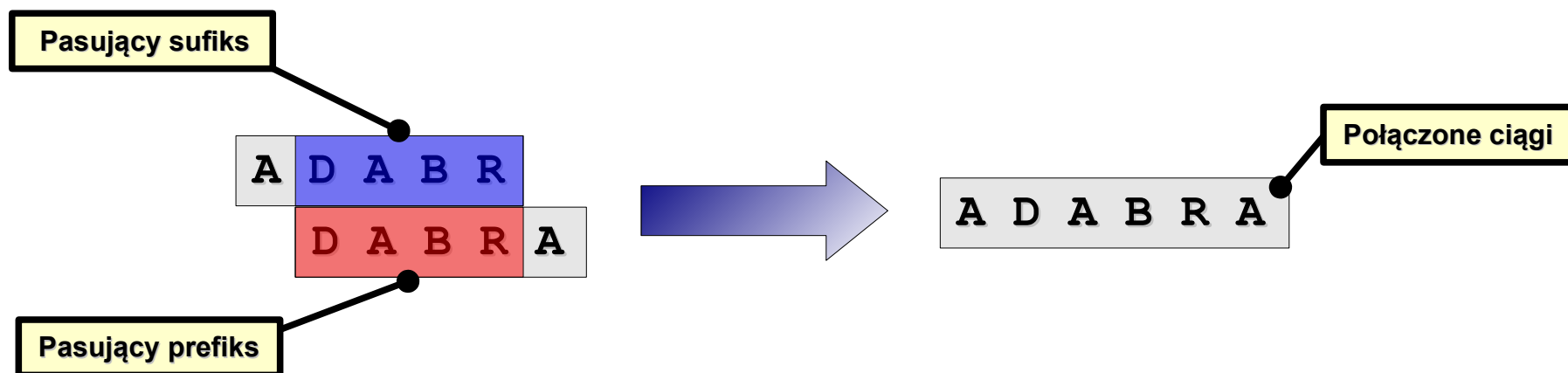
W idealnym przypadku składanie sekwencji genomu na podstawie zbioru odczytów mogłoby być traktowane jako szczególny przypadek problemu poszukiwania **najkrótszego wspólnego nadłańcucha** (*Shortest Common Superstring – SCS*), czyli wyznaczania najkrótszego możliwego ciągu znakowego zawierającego wszystkie ciągi z danego zbioru $\{S_1, S_2, \dots, S_N\}$.



Składanie sekwencji genomu w sekwencjonowaniu *shotgun*

Pomimo pozornej prostoty, **poszukiwanie najkrótszego wspólnego nadłańcucha jest problemem klasy NP (*Nondeterministic Polynomial*)**, którego złożoność obliczeniowa rośnie bardzo szybko wraz ze wzrostem liczby sekwencji wejściowych. W ogólnym przypadku tego typu problemy mogą być rozwiązywane tylko w sposób przybliżony.

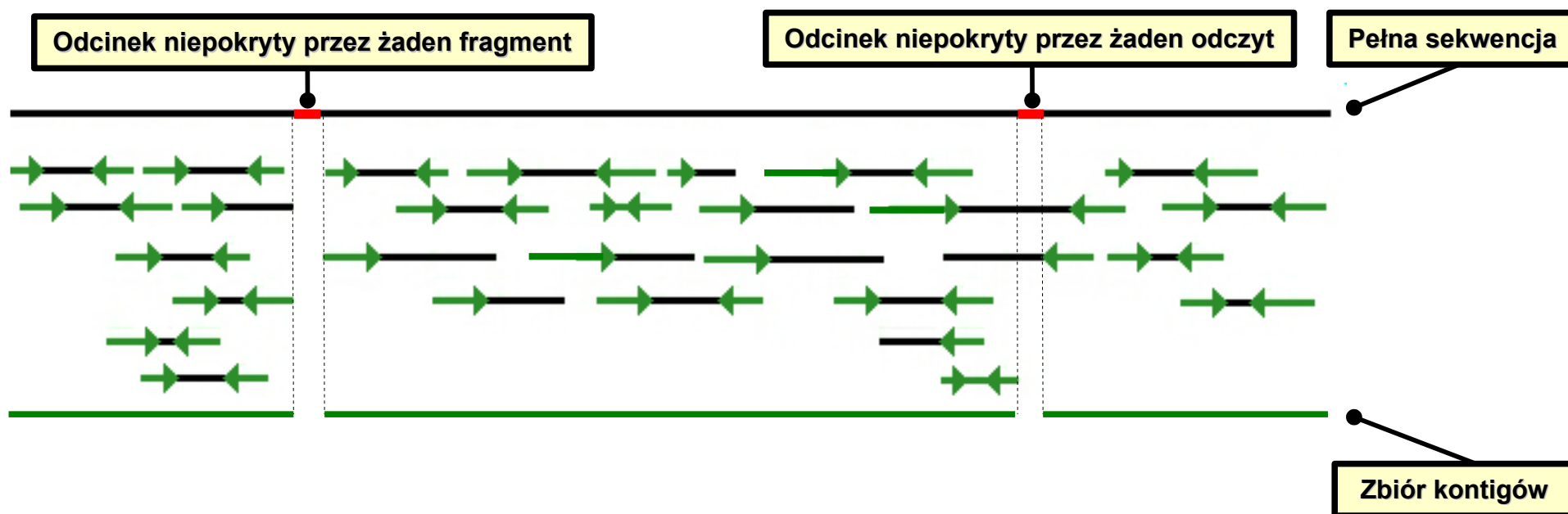
Najprostszym koncepcyjnie podejściem do jego przybliżonego rozwiązania jest zastosowanie **algorytmu typu zachłannego (*greedy algorithm*)**, który w kolejnych krokach łączy ciągi S_i i S_j ($i \neq j$) o najdłuższych pasujących sufiksach i prefiksach.



Składanie sekwencji genomu w sekwencjonowaniu *shotgun*

Dla rzeczywistych danych pomiędzy składaniem sekwencji genomu a problemem najkrótszego wspólnego nadłańcucha istnieją istotne różnice:

- odtworzenie całkowitej sekwencji genomu może okazać się niemożliwe. Zamiast tego uzyskuje się zbiór ciągłych podsekwencji zwanych **kontigami**. Ich liczba zależy od długości odczytów oraz stopnia redundancji (oba te czynniki można wyrazić wspólnie przez **pokrycie**, czyli średnią liczbę odczytów zawierających daną pozycję w genomie). Nieciągłość sekwencji uzyskanej w procesie składania powoduje, że konieczne staje się wprowadzenia dodatkowego **kroku łączenia kontigów**;



Składanie sekwencji genomu w sekwencjonowaniu *shotgun*

- ze względu na możliwość występowania błędów i mutacji nie możemy zakładać idealnej zgodności sekwencji odczytów. Dlatego też poszukując nakładających się odczytów trzeba użyć algorytmu wyznaczania **lokalnego dopasowania sekwencji**;

```
...GACACTGAAGTCGGGTTCCCACAGGAAT-GGAAGG
                        || ||||| |||||
                        ACTGGAATCCGAAGGGGGTATAGTGCGACCCGGCTAA...
```

- w miejscach nakładania się dwóch lub większej liczby odczytów konieczne jest wyznaczenie **sekwencji konsensusowej**, czyli określenie nukleotydów najczęściej występujących na poszczególnych pozycjach.

```
Dopasowane sekwencje nakładających się obszarów: ...ACTAG...
                                                    ||
                                                    ...AC-GT...
                                                    || |
                                                    ...ACT-T...
                                                    |||
                                                    ...AGTGA...
```

```
Sekwencja konsensusowa: ...ACTGT...
```

Dopasowanie sekwencji

Globalne dopasowanie (*global alignment*) sekwencji S_1 i S_2 o alfabecie $\{A, G, C, T\}$ polega na znalezieniu ciągów A_1 i A_2 o alfabecie $\{A, G, C, T, -\}$ spełniających warunki:

- S_1 i S_2 otrzymuje się z A_1 i A_2 poprzez usunięcie wszystkich znaków przerwy $\{-\}$;
- długość A_1 jest równa długości A_2 ;
- znaki przerwy nie mogą występować w A_1 i A_2 na tych samych pozycjach.

S_1 :	FTFTALILLAVAV
S_2 :	FTALLLA AV
A_1 :	FTFTALILLAVAV
A_2 :	F--TALL-L A-AV

Obszar **dopasowania lokalnego (*local alignment*)** nie musi obejmować całej długości sekwencji S_1 i S_2 . Tym samym dopasowanie lokalne pozwala zidentyfikować podobne regiony w obrębie sekwencji, które – rozpatrywane w całości – nie muszą wykazywać znacznego podobieństwa.

S_1 :	FTFTALILLAVAV
S_2 :	FTALLLA AV
A_1 :	FT FTALILL -AVAV
A_2 :	-- FTAL-LLAAV --

Dopasowanie sekwencji: system oceny (scoring)

Obliczenie **optymalnego** dopasowania sekwencji (globalnego lub lokalnego) wymaga ustalenia **systemu oceniania**, w skład którego wchodzi trzy elementy:

- **macierz substytucji**, której element $e(\alpha, \beta)$ określa wpływ na ocenę wystąpienia pary nukleotydów α i β na tej samej pozycji ciągów A_1 i A_2 . Prosta macierz tego typu zawiera jednakowe wartości: dodatnie, gdy $\alpha = \beta$ oraz ujemne, gdy $\alpha \neq \beta$. Bardziej złożone macierze uwzględniają różnice w tempie mutacji nukleotydów;

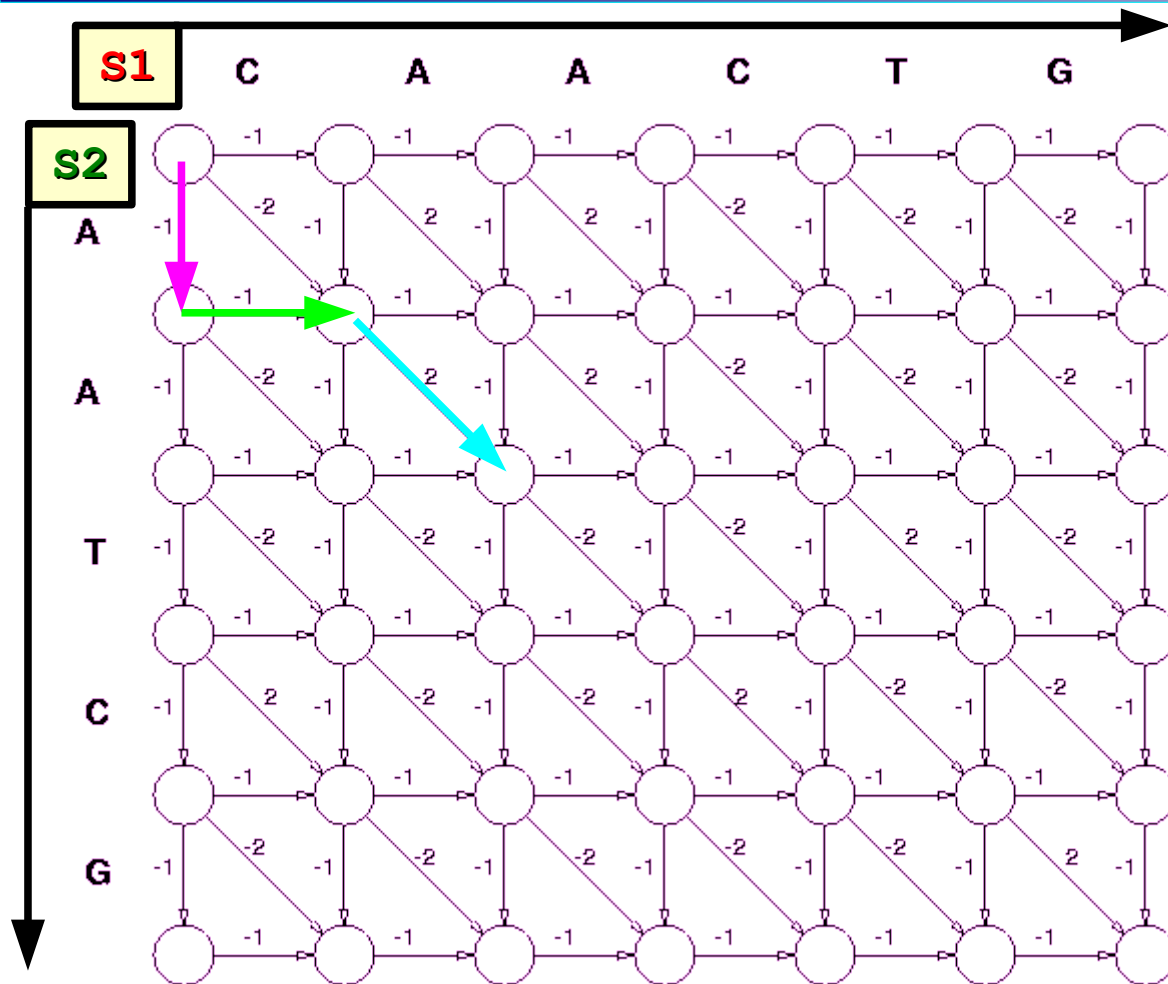
	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

- **kara za występowanie znaku przerwy** w dopasowanych ciągach. W najprostszym przypadku jest to pojedyncza wartość w (np. równa -1). W bardziej złożonych systemach oceny występują osobne kary za otwarcie i kontynuację przerwy;
- sposób liczenia **oceny (score)** w oparciu o elementy macierzy substytucji i wartość kary za przerwy. W najprostszym przypadku $score$ wyznaczany jest jako suma:

$$score = \sum_{\text{dopasowane pary}} e(\alpha, \beta) + \sum_{\text{przerwy}} w$$

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: -CA

A2: A-A

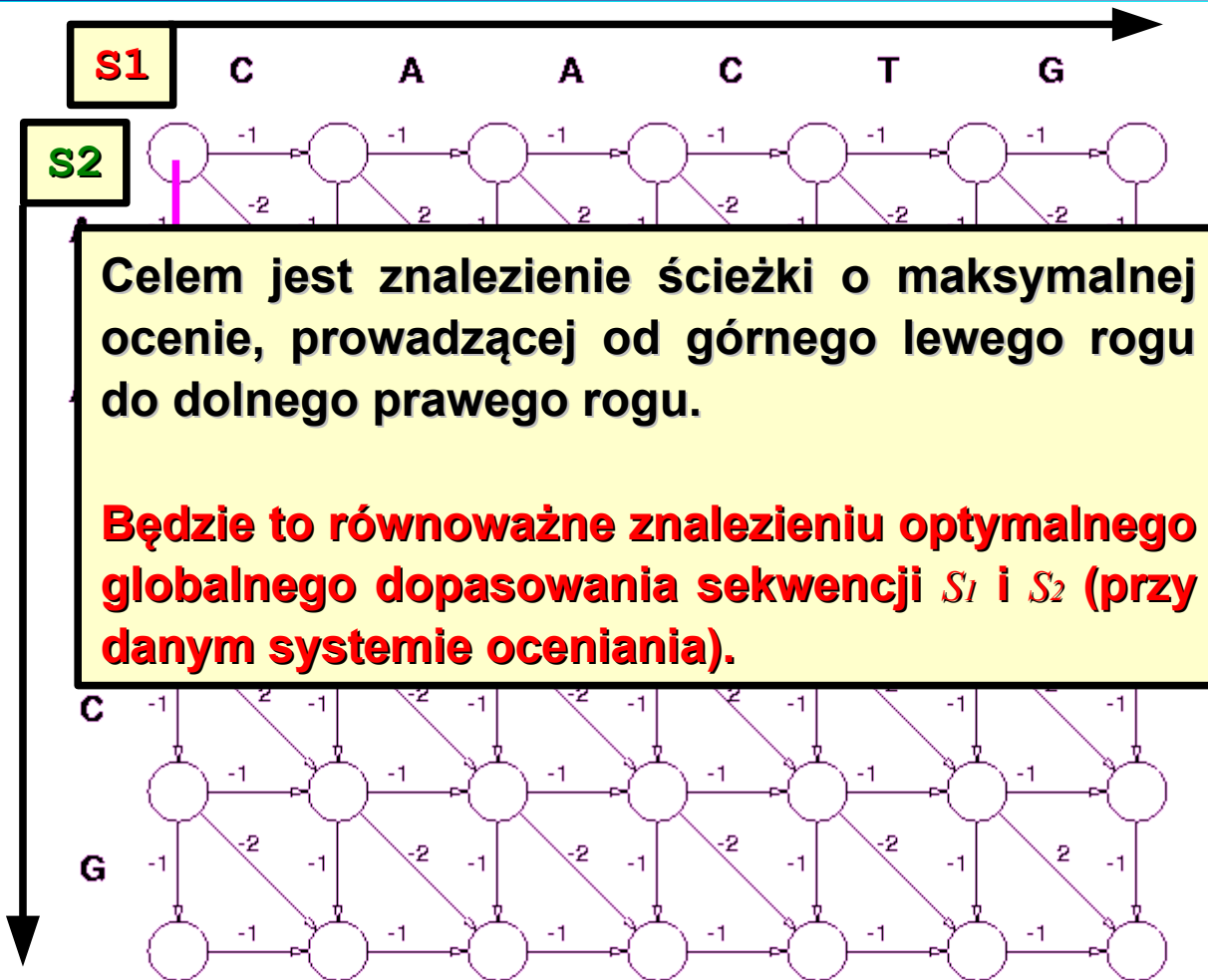
Ocena

score = (-1) + (-1) + (+2)

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji *S1* i *S2*. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji *A1* i nukleotydu do *A2* („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji *A2* i nukleotydu do *A1* („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do *A1* i *A2* („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji *S1* i *S2* różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: -CA

A2: A-A

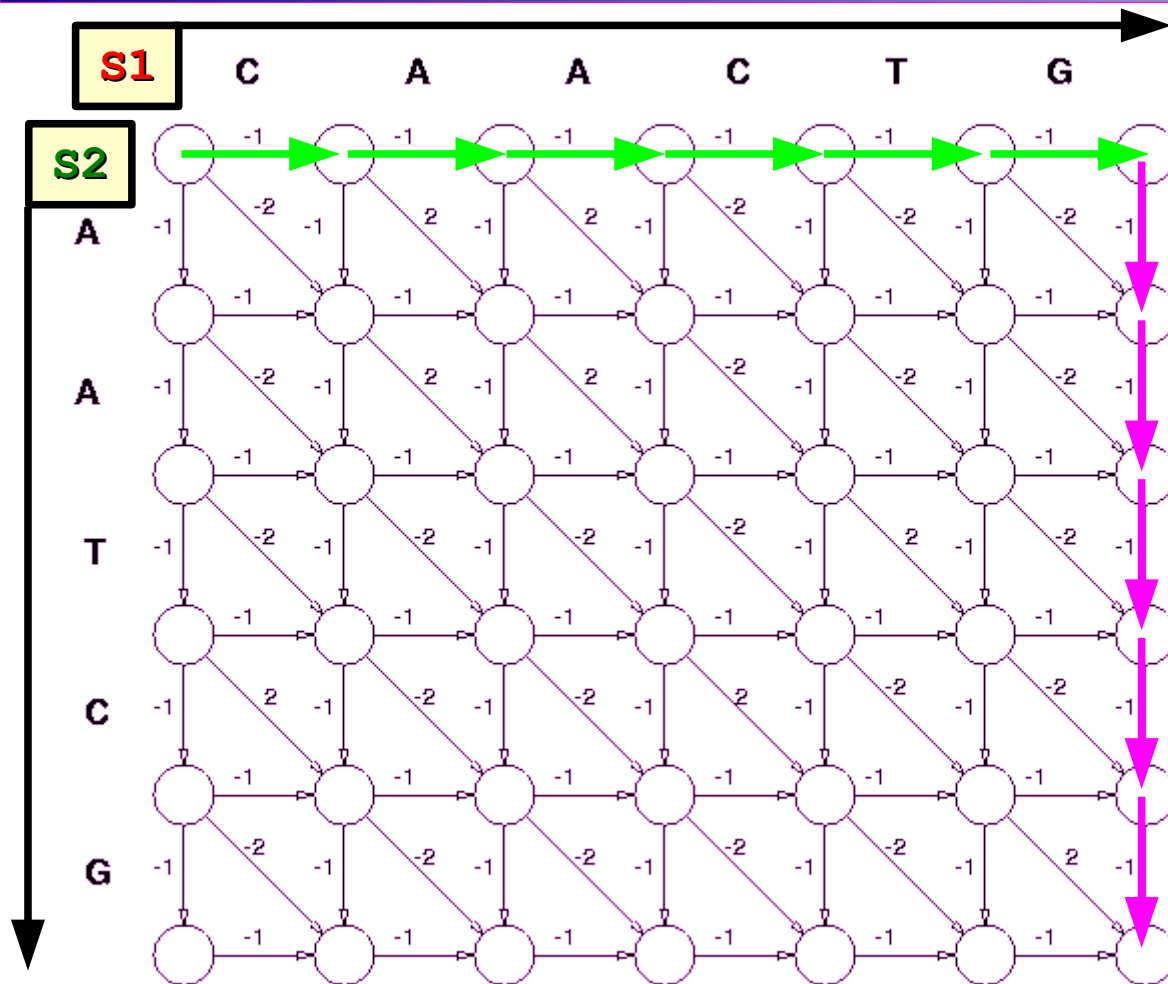
Ocena

score = (-1) + (-1) + (+2)

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji S_1 i S_2 . O jakości dopasowania świadczy jego ocena będąca sumą „kara” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji A_1 i nukleotydu do A_2 („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji A_2 i nukleotydu do A_1 („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do A_1 i A_2 („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji S_1 i S_2 różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: CAACTG-----
A2: -----AATCG

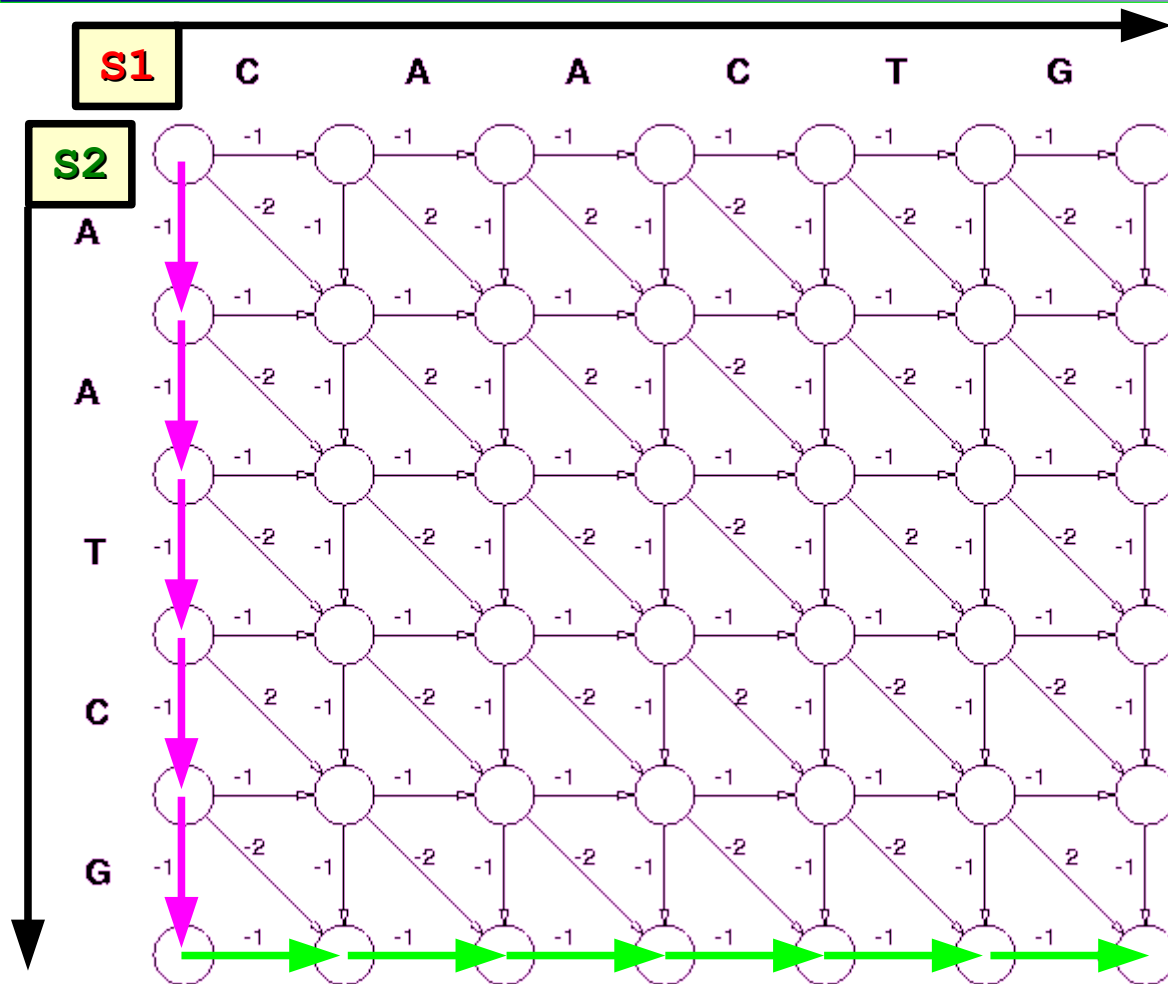
Ocena

score = -11

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji *S1* i *S2*. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji *A1* i nukleotydu do *A2* („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji *A2* i nukleotydu do *A1* („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do *A1* i *A2* („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji *S1* i *S2* różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: -----CAACTG
A2: AATCG-----

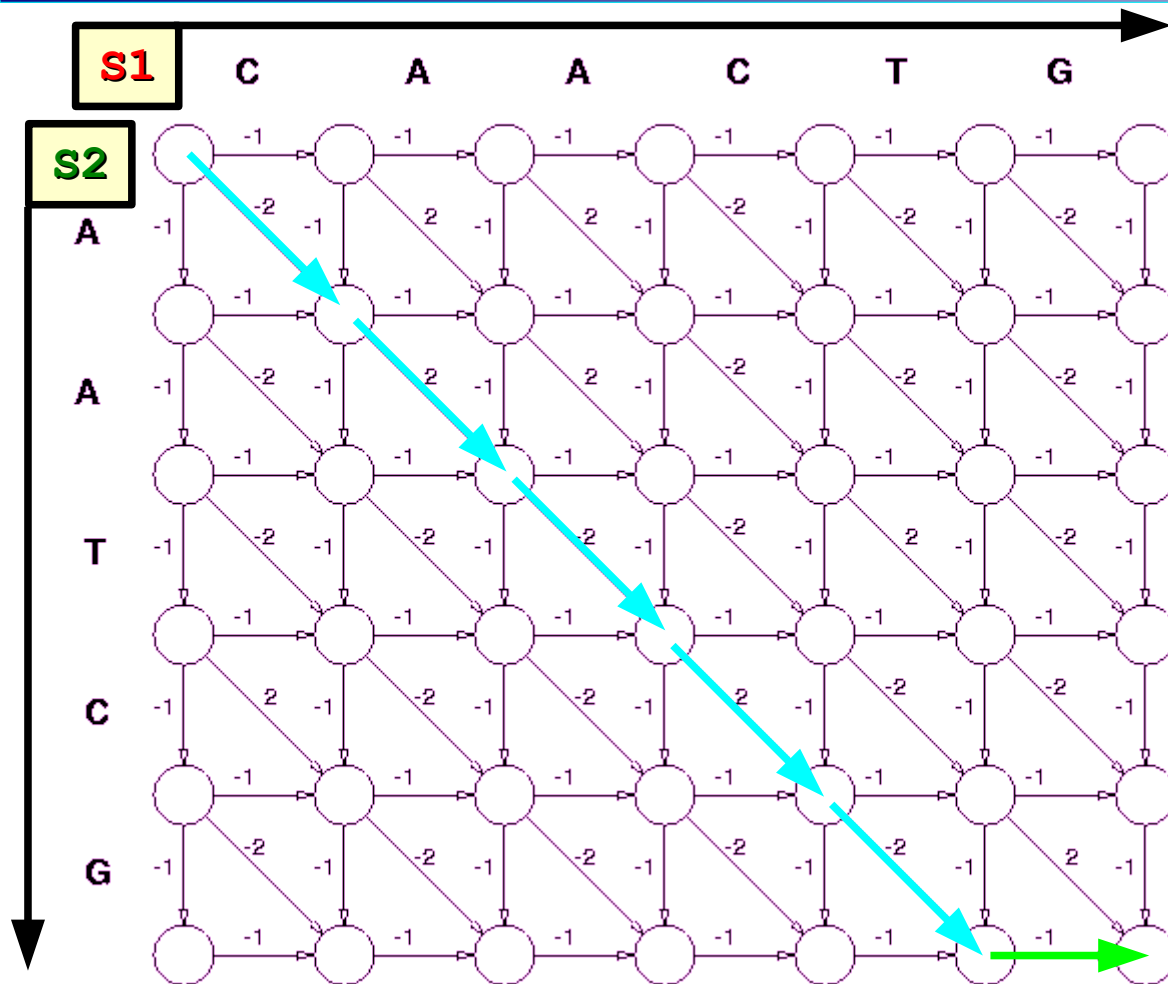
Ocena

score = -11

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: CAACTG

A2: AATCG-

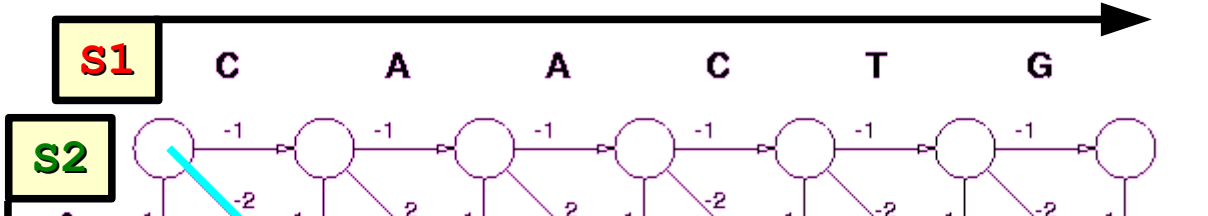
Ocena

score = -3

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kara” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

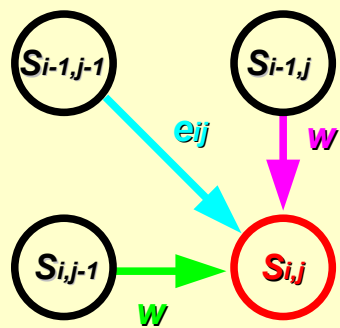
- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



Znając zasady „gry” możemy z góry wypełnić planszę w taki sposób aby z węzłami związane były najwyższe możliwe oceny (wynikające z optymalnych ścieżek do nich prowadzących).

Zaczynając od górnego lewego rogu (ocena 0) węzłom przypisujemy ocenę jako maksymalną wartością ocen trzech wskazujących na niego węzłów (uzupełnionych o wagi krawędzi):



$$S_{i,j} = \max(S_{i-1,j} + w, S_{i,j-1} + w, S_{i-1,j-1} + e_{ij});$$

- gdzie:
- $S_{i,j}$ - ocena węzła w i -tej kolumnie i j -tym wierszu
 - e_{ij} - waga krawędzi
 - w - wartość kary za przerwę

System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1:

A2:

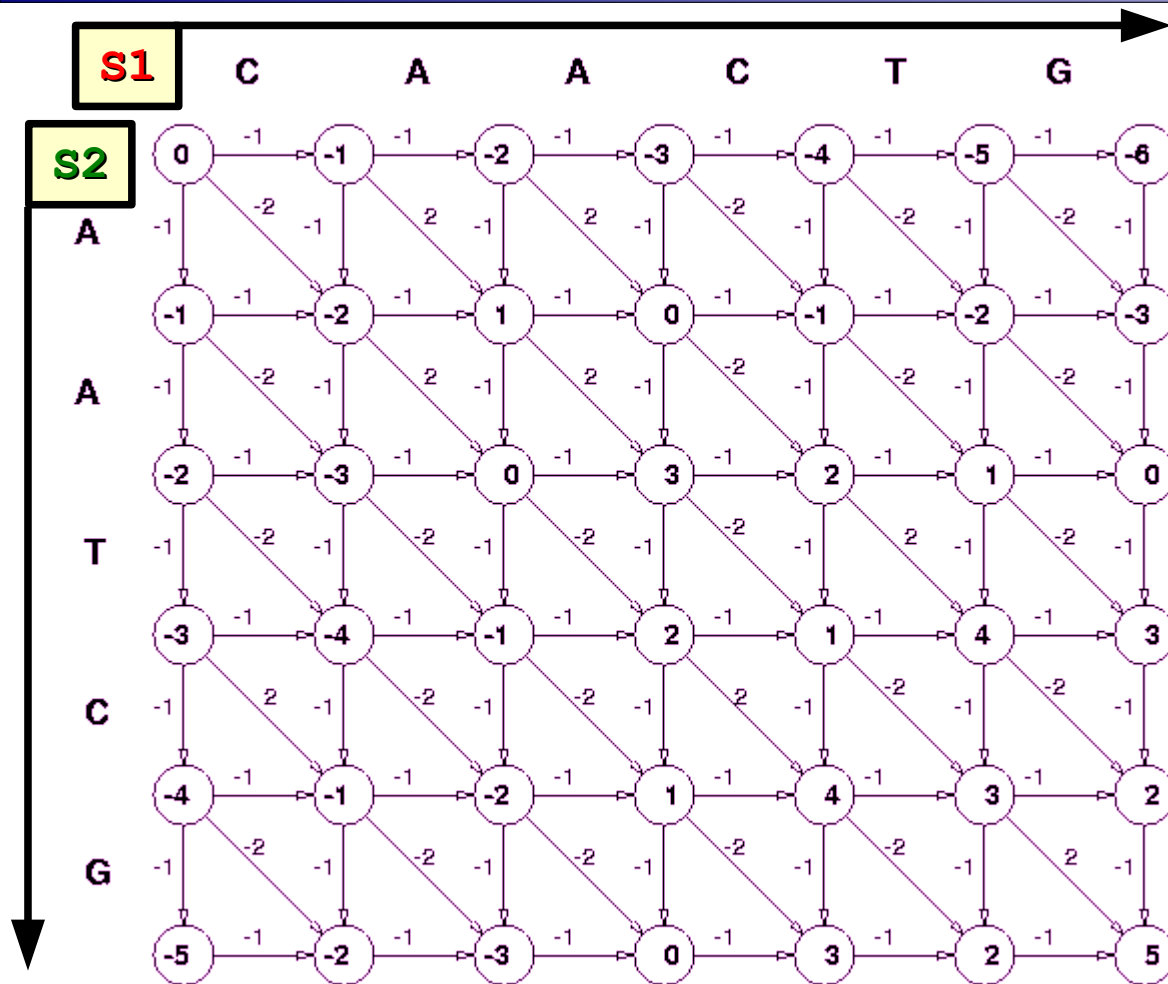
Ocena

score = 0

... dolnego prawego dolnego reprezentuje ... i S2. O jakości dopasowania świadczy ... iach zależnych od systemu punktacji)

... kleotydu do A2 („kara” -1)
... nukleotydu do A1 („kara” -1)
... („kara” -2 jeżeli nukleotydy na danych ...
... jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1:

A2:

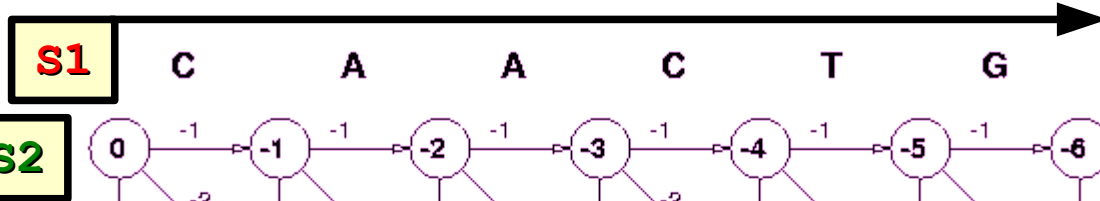
Ocena

score = 0

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

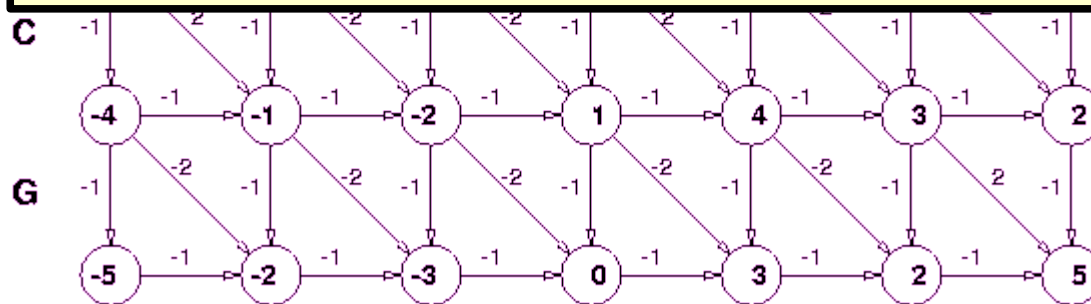
- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



Wypełniona plansza pozwala nam budować optymalną ścieżkę „od tyłu”.

Zaczynając od dolnego prawego rogu w każdym ruchu wybieramy jako następny ten węzeł, który na etapie wypełniania planszy doprowadził nas do obecnego.



Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kara” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

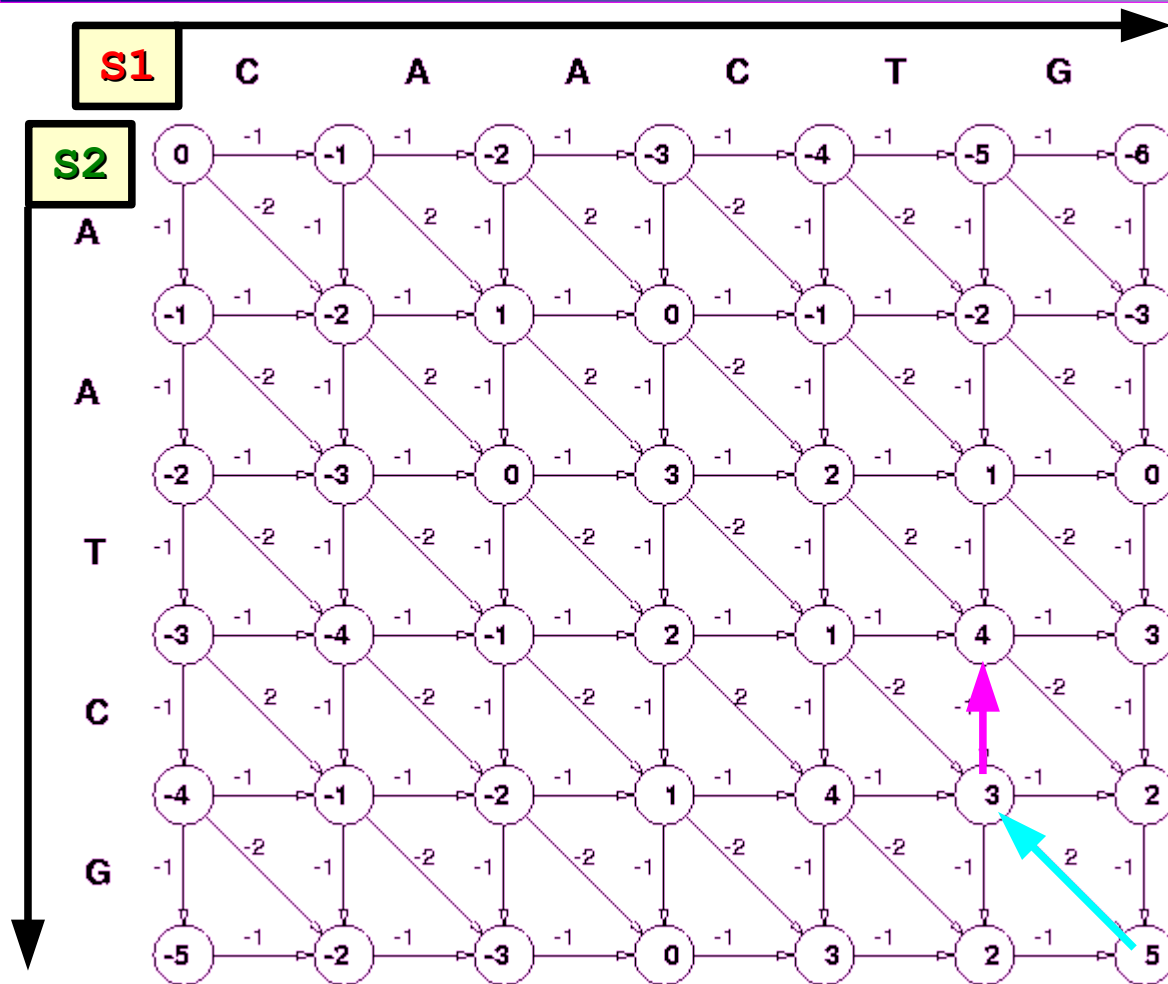
A1:

A2:

Ocena

score = 0

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: -G
A2: CG

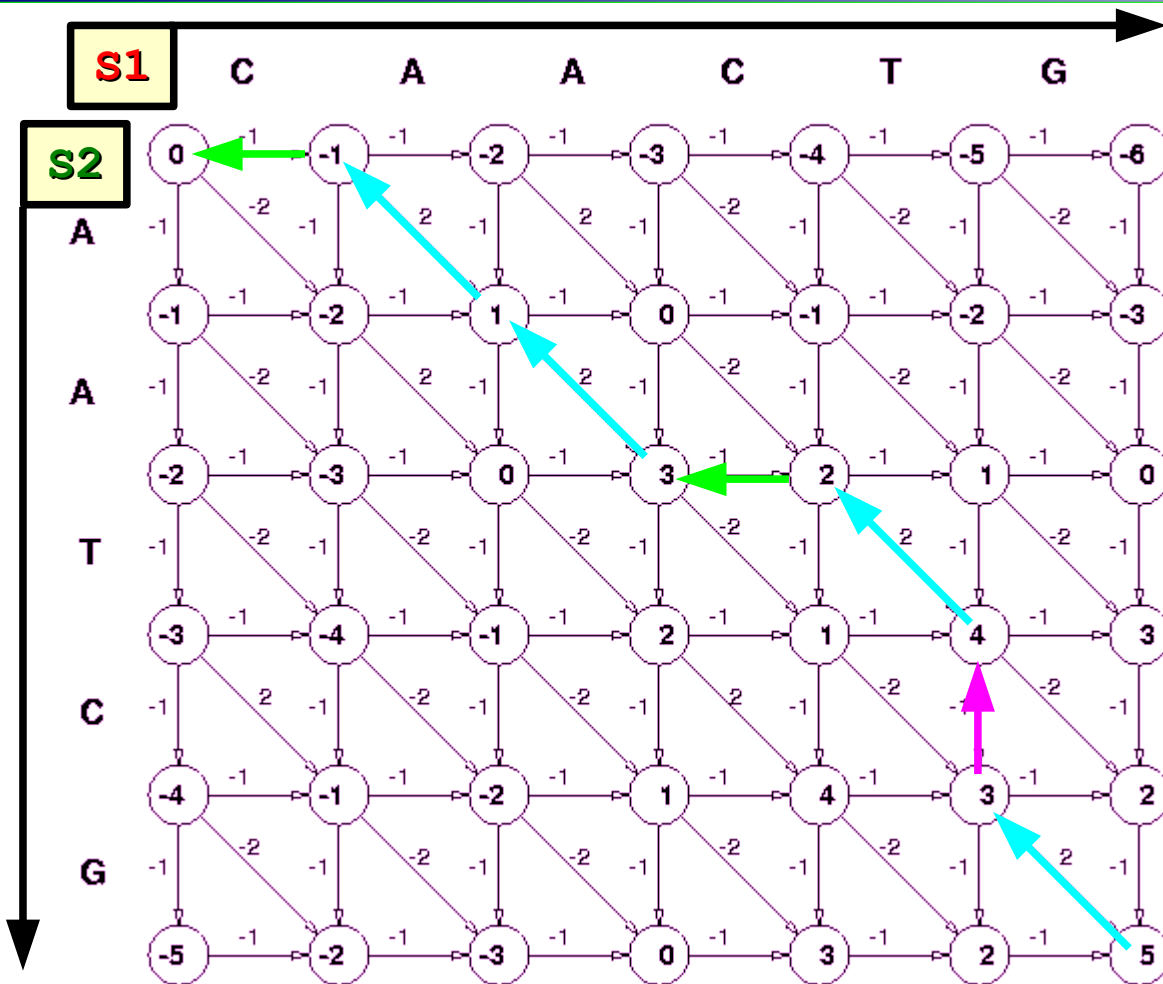
Ocena

score = 2 - 1

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: CAACT-G

A2: -AA-TCG

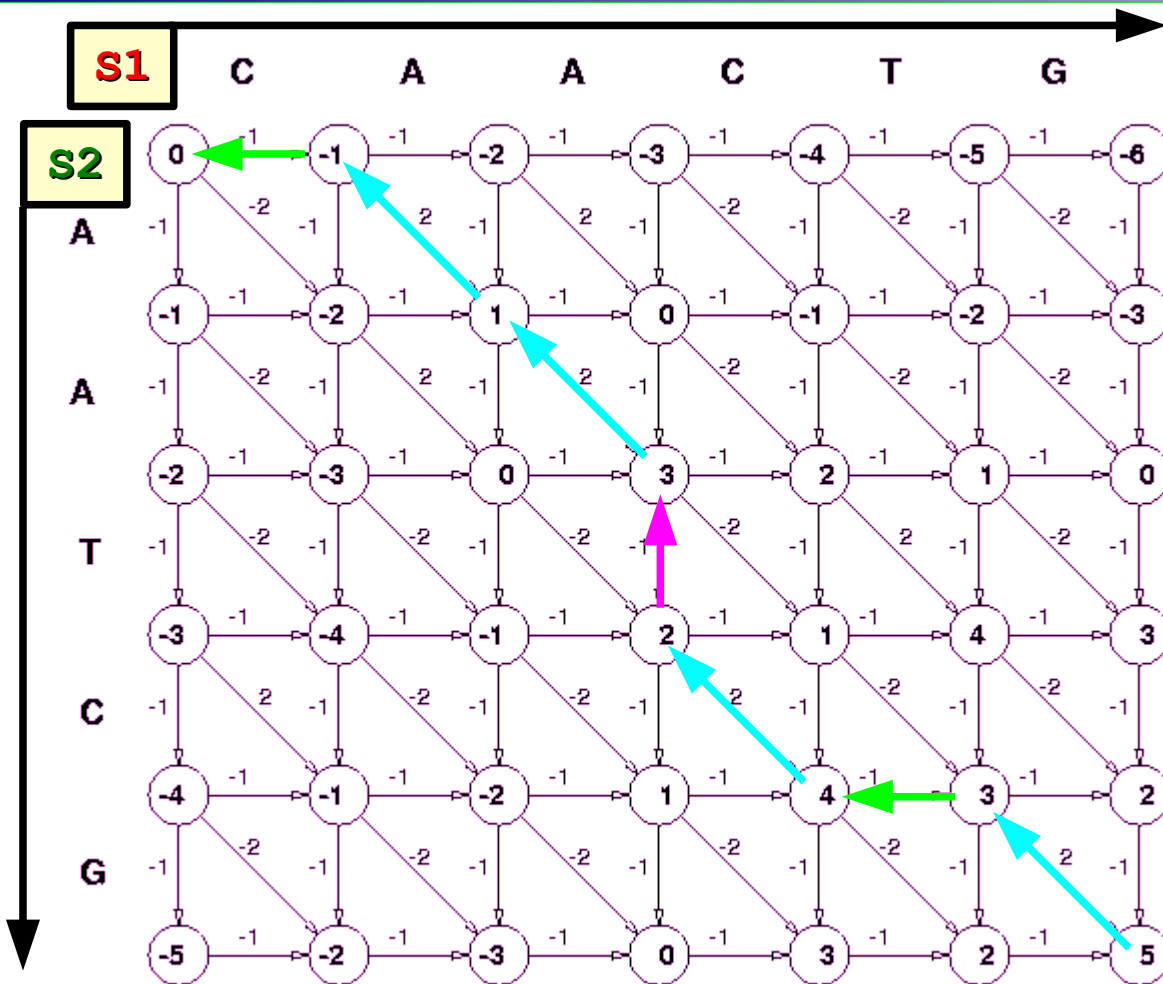
Ocena

score = 2 - 1 + 2 - 1 + 2 + 2 - 1 = 5

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: CAA-CTG
A2: -AATC-G

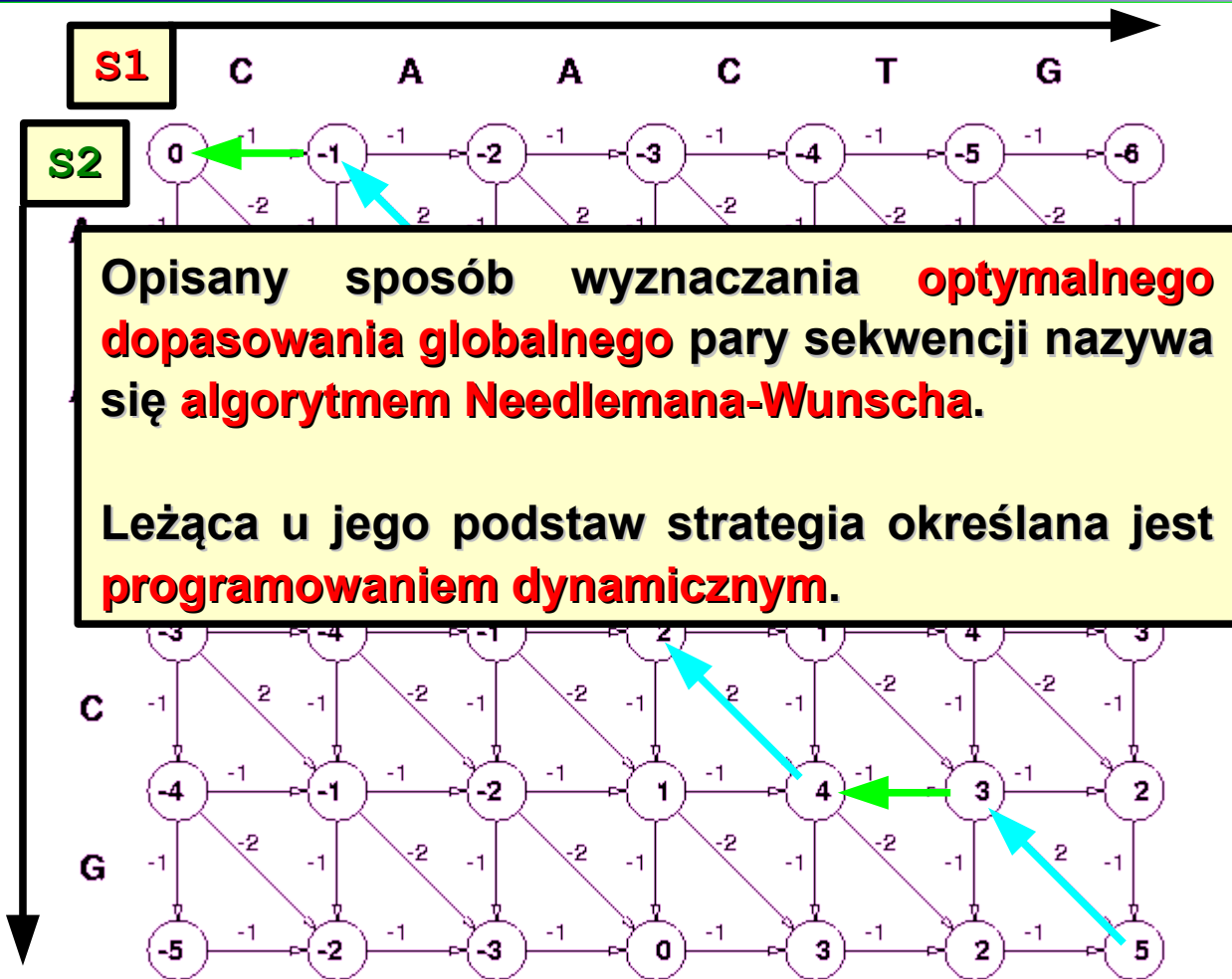
Ocena

score = 2 - 1 + 2 - 1 + 2 + 2 - 1 = 5

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie globalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: CAA-CTG
A2: -AATC-G

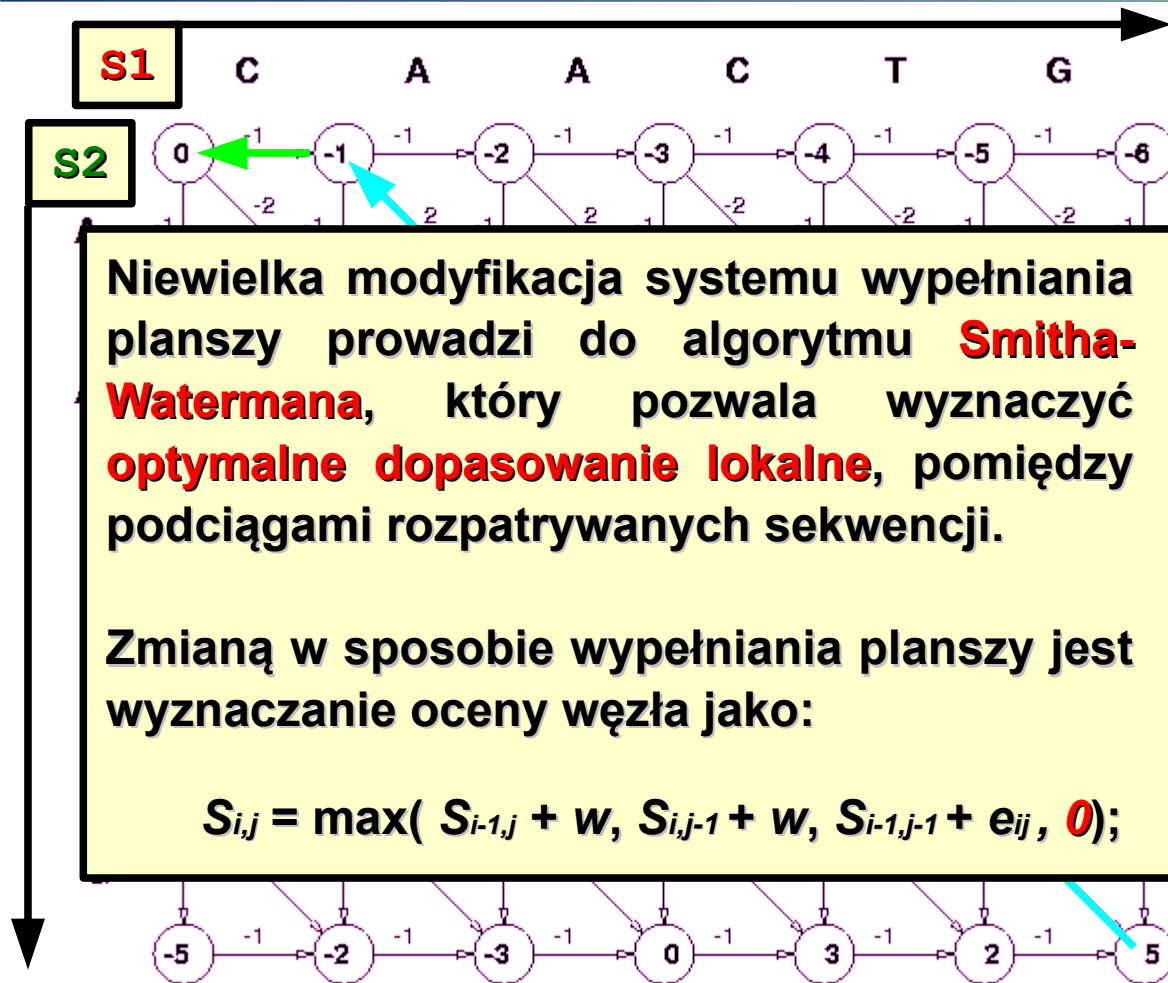
Ocena

score = 2 - 1 + 2 - 1 + 2 + 2 - 1 = 5

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kara” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie lokalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: CAA-CTG

A2: -AATC-G

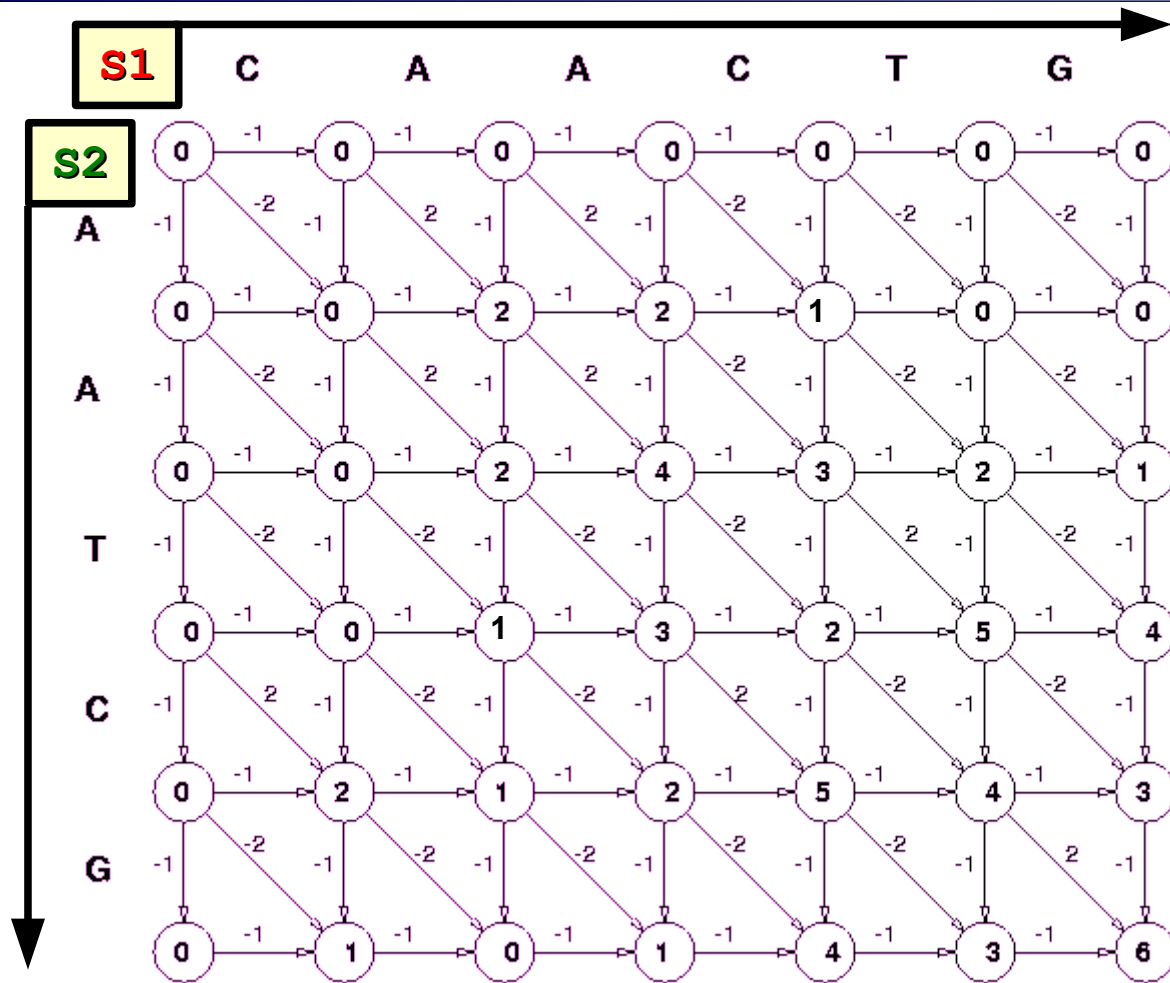
Ocena

score = 2 - 1 + 2 - 1 + 2 + 2 - 1 = 5

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kara” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie lokalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1:

A2:

Ocena

score = 0

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie lokalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1:

A2:

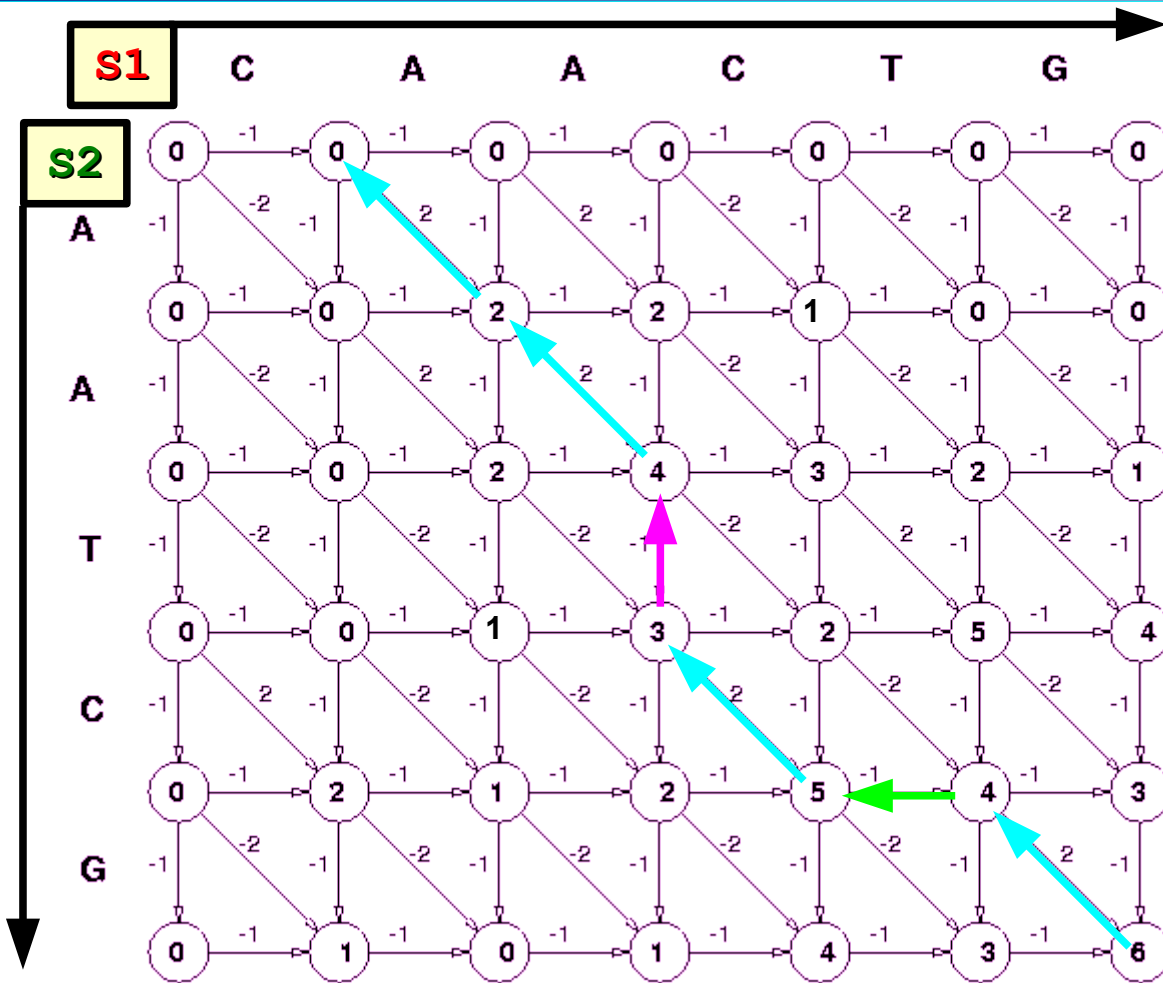
Ocena

score = 0

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie lokalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: AA-CTG
A2: AATC-G

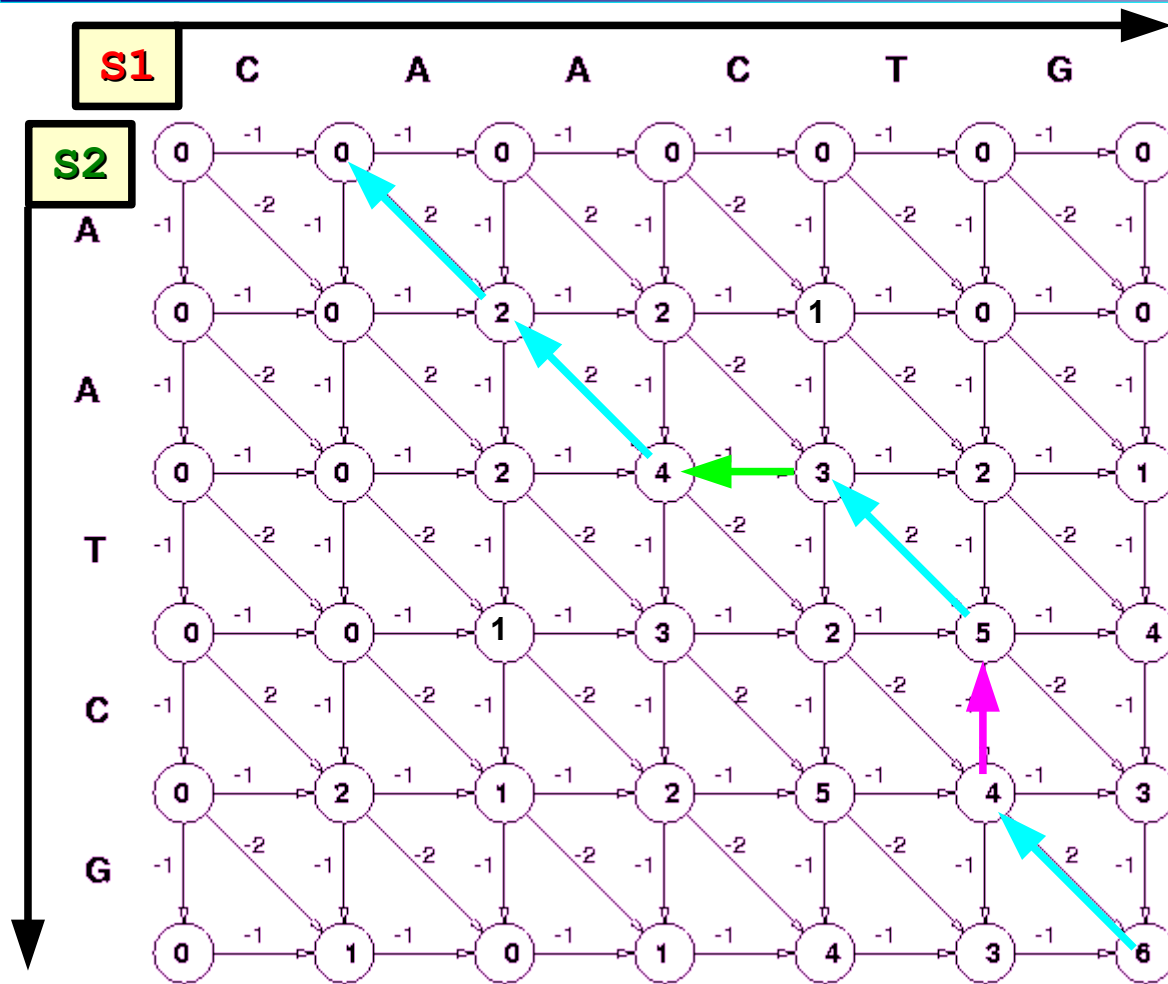
Ocena

score = 2 - 1 + 2 - 1 + 2 + 2 = 6

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Dopasowanie sekwencji: optymalne dopasowanie lokalne



System punktacji

	A	G	C	T
A	2	-2	-2	-2
G	-2	2	-2	-2
C	-2	-2	2	-2
T	-2	-2	-2	2

Kara z przerwę = -1

Dopasowanie

A1: AACT-G

A2: AA-TCG

Ocena

$$\text{score} = 2 - 1 + 2 - 1 + 2 + 2 = 6$$

Każda ścieżka prowadząca od górnego lewego rogu do dolnego prawego dolnego reprezentuje jeden z możliwych sposobów dopasowania sekwencji **S1** i **S2**. O jakości dopasowania świadczy jego ocena będąca sumą „kar” i „nagród” (o wartościach zależnych od systemu punktacji) związanych z możliwymi ruchami:

- **Ruch w pionie:** wstawienie przerwy do sekwencji **A1** i nukleotydu do **A2** („kara” -1)
- **Ruch w poziomie:** wstawienie przerwy do sekwencji **A2** i nukleotydu do **A1** („kara” -1)
- **Ruch po przekątnej:** wstawienie nukleotydów do **A1** i **A2** („kara” -2 jeżeli nukleotydy na danych pozycjach sekwencji **S1** i **S2** różnią się lub „nagroda” +2 jeżeli nukleotydy są takie same)

Sekwencja konsensusowa

Sekwencja konsensusowa jest to sekwencja powstała na podstawie częstości występowania nukleotydów na poszczególnych pozycjach dopasowania.

Reguły wyznaczania sekwencji konsensusowej dla przykładowego algorytmu (TIGR)

Sekwencje dopasowanych obszarów: A-CTAC

A-CAAC

A-GTAT

A-GTAT

A-GTGA

A-GTGG

A-GAGC

A-GAGG

Sekwencja konsensusowa: A-Gtmn

A, G, C, T, - : więcej niż 2/3 jednakowych znaków na danej pozycji

a, g, c, t : więcej niż 1/2 jednakowych znaków na danej pozycji

m, r, w, s, y, k : równa liczba wstąpień dwóch znaków na danej pozycji

n : w pozostałych przypadkach

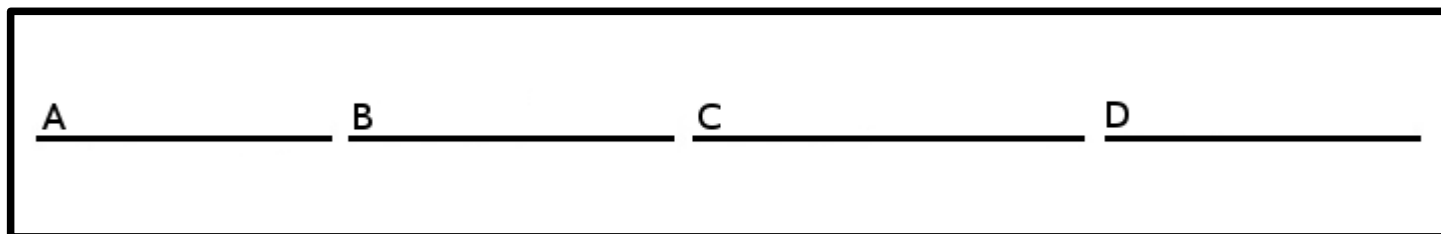
Składanie sekwencji genomu w strategii *shotgun*: algorytm zachłanny

Najprostszym podejściem do składania pełnej sekwencji genomu jest zastosowanie **zmodyfikowanego algorytmu zachłannego** służącego do rozwiązywania problemu najkrótszego wspólnego nadłańcucha:

1. Wyznaczenie dopasowania lokalnego pomiędzy każdą parą sekwencji.
2. Iteracyjne powtarzanie dwóch kroków:
 - wybór dwóch niepołączonych sekwencji o najdłuższym dopasowaniu lokalnym (przy zadanych progach min. dopuszczalnych długości i jakości dopasowań);
 - połączenie wybranych sekwencji w kontig;

Proces kończy się gdy pozostaje jeden kontig lub zbiór kontigów nie mających obszarów dopasowania o wymaganej minimalnej długości.

3. Określenie sekwencji konsensusowej dla obszarów dopasowań.



	A	B	C	D
A	-			
B	200	-		
C	5	50	-	
D	16	14	150	-

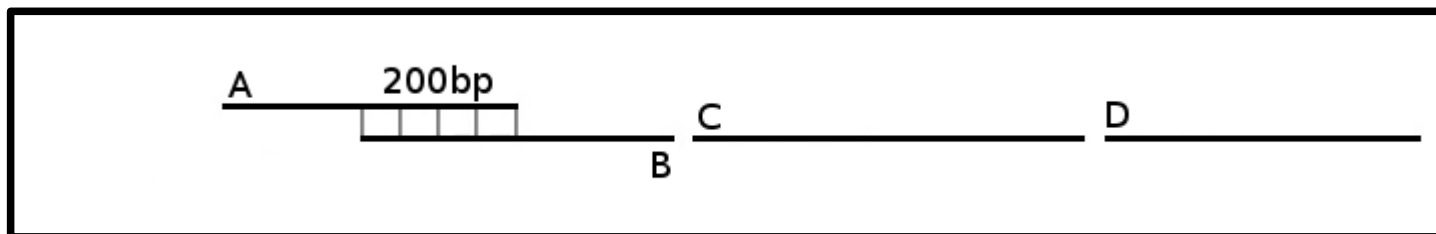
Składanie sekwencji genomu w strategii *shotgun*: algorytm zachłanny

Najprostszym podejściem do składania pełnej sekwencji genomu jest zastosowanie **zmodyfikowanego algorytmu zachłannego** służącego do rozwiązywania problemu najkrótszego wspólnego nadłańcucha:

1. Wyznaczenie dopasowania lokalnego pomiędzy każdą parą sekwencji.
2. Iteracyjne powtarzanie dwóch kroków:
 - wybór dwóch niepołączonych sekwencji o najdłuższym dopasowaniu lokalnym (przy zadanych progach min. dopuszczalnych długości i jakości dopasowań);
 - połączenie wybranych sekwencji w kontig;

Proces kończy się gdy pozostaje jeden kontig lub zbiór kontigów nie mających obszarów dopasowania o wymaganej minimalnej długości.

3. Określenie sekwencji konsensusowej dla obszarów dopasowań.



	A	B	C	D
A	-			
B	200*	-		
C	5	50	-	
D	16	14	150	-

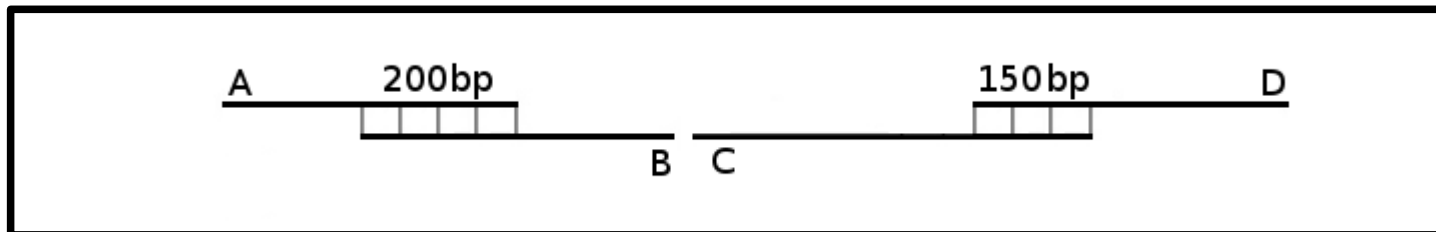
Składanie sekwencji genomu w strategii *shotgun*: algorytm zachłanny

Najprostszym podejściem do składania pełnej sekwencji genomu jest zastosowanie **zmodyfikowanego algorytmu zachłannego** służącego do rozwiązywania problemu najkrótszego wspólnego nadłańcucha:

1. Wyznaczenie dopasowania lokalnego pomiędzy każdą parą sekwencji.
2. Iteracyjne powtarzanie dwóch kroków:
 - wybór dwóch niepołączonych sekwencji o najdłuższym dopasowaniu lokalnym (przy zadanych progach min. dopuszczalnych długości i jakości dopasowań);
 - połączenie wybranych sekwencji w kontig;

Proces kończy się gdy pozostaje jeden kontig lub zbiór kontigów nie mających obszarów dopasowania o wymaganej minimalnej długości.

3. Określenie sekwencji konsensusowej dla obszarów dopasowań.



	A	B	C	D
A	-			
B	200*	-		
C	5	50	-	
D	16	14	150*	-

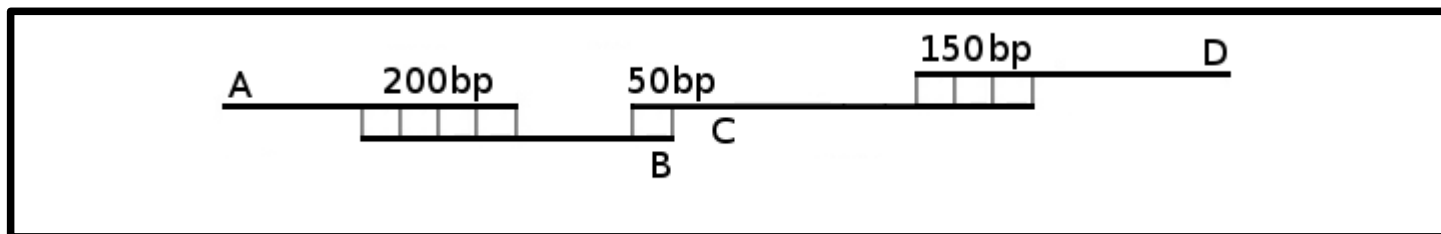
Składanie sekwencji genomu w strategii *shotgun*: algorytm zachłanny

Najprostszym podejściem do składania pełnej sekwencji genomu jest zastosowanie **zmodyfikowanego algorytmu zachłannego** służącego do rozwiązywania problemu najkrótszego wspólnego nadłańcucha:

1. Wyznaczenie dopasowania lokalnego pomiędzy każdą parą sekwencji.
2. Iteracyjne powtarzanie dwóch kroków:
 - wybór dwóch niepołączonych sekwencji o najdłuższym dopasowaniu lokalnym (przy zadanych progach min. dopuszczalnych długości i jakości dopasowań);
 - połączenie wybranych sekwencji w kontig;

Proces kończy się gdy pozostaje jeden kontig lub zbiór kontigów nie mających obszarów dopasowania o wymaganej minimalnej długości.

3. Określenie sekwencji konsensusowej dla obszarów dopasowań.



	A	B	C	D
A	-			
B	200*	-		
C	5	50*	-	
D	16	14	150*	-

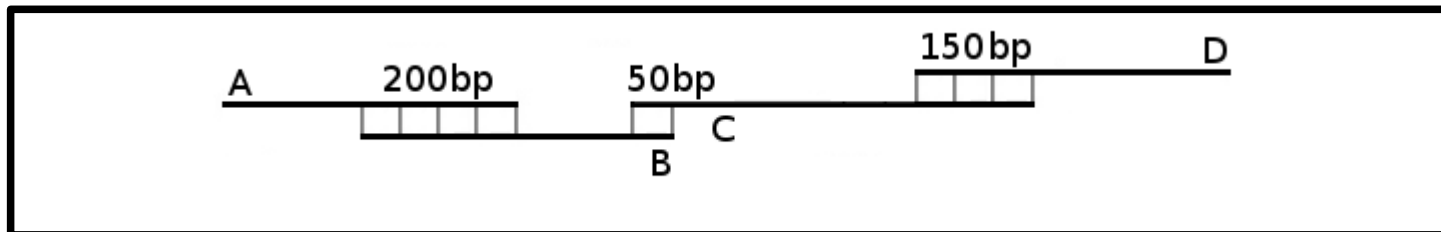
Składanie sekwencji genomu w strategii *shotgun*: algorytm zachłanny

Najprostszym podejściem do składania pełnej sekwencji genomu jest zastosowanie **zmodyfikowanego algorytmu zachłannego** służącego do rozwiązywania problemu najkrótszego wspólnego nadłańcucha:

1. Wyznaczenie dopasowania lokalnego pomiędzy każdą parą sekwencji.
2. Iteracyjne powtarzanie dwóch kroków:
 - wybór dwóch niepołączonych sekwencji o najdłuższym dopasowaniu lokalnym (przy zadanych progach min. dopuszczalnych długości i jakości dopasowań);
 - połączenie wybranych sekwencji w kontig;

Proces kończy się gdy pozostaje jeden kontig lub zbiór kontigów nie mających obszarów dopasowania o wymaganej minimalnej długości.

3. Określenie sekwencji konsensusowej dla obszarów dopasowań.



	A	B	C	D
A	-			
B	200*	-		
C	5	50*	-	
D	16	14	150*	-

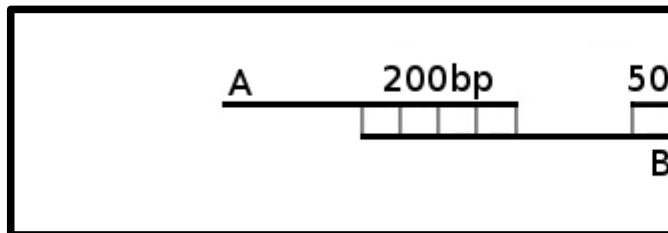
Składanie sekwencji genomu w strategii *shotgun*: algorytm zachłanny

Najprostszym podejściem do składania pełnej sekwencji genomu jest zastosowanie **zmodyfikowanego algorytmu zachłannego** służącego do rozwiązywania problemu najkrótszego wspólnego nadłańcucha:

1. Wyznaczenie dopasowania lokalnego pomiędzy każdą parą sekwencji.
2. Iteracyjne powtarzanie dwóch kroków:
 - wybór dwóch niepołączonych sekwencji o najdłuższym dopasowaniu lokalnym (przy zadanych progach min. dopuszczalnych długości i jakości dopasowań);
 - połączenie wybranych sekwencji w kontig;

Proces kończy się gdy pozostaje jeden kontig lub zbiór kontigów nie mających obszarów dopasowania o wymaganej minimalnej długości.

3. Określenie sekwencji kons

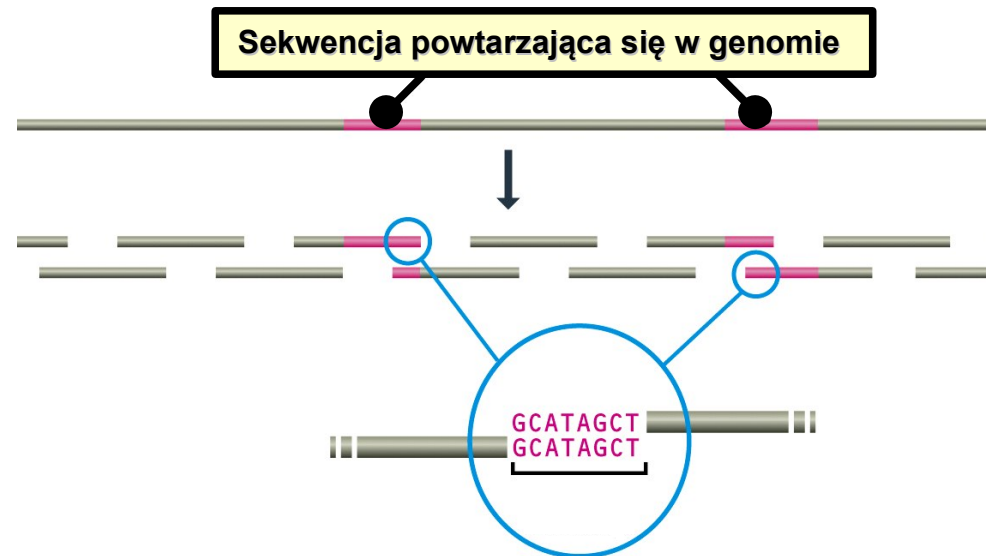
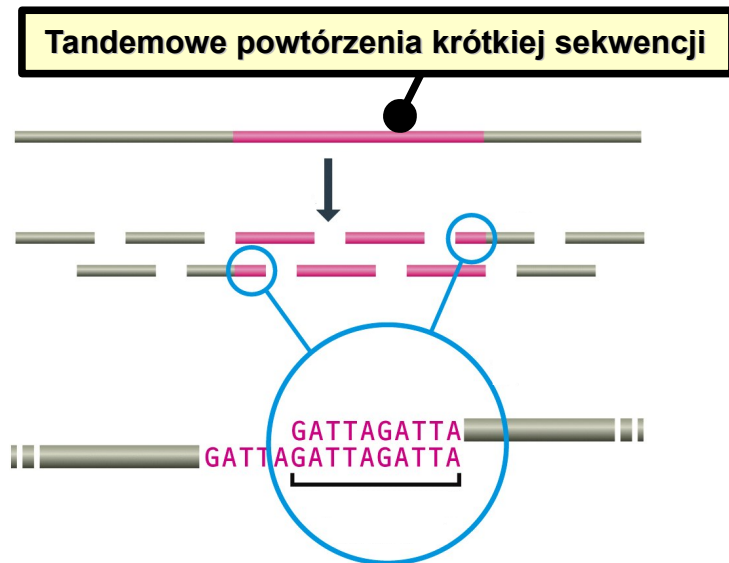


Ograniczeniem algorytmu zachłannego jest fakt, że w każdym kroku może on dokonać wyboru rozwiązania częściowego, optymalnego jedynie w sensie lokalnym (w danym momencie działania algorytmu), nie weryfikując przy tym skutków tego wyboru dla optymalności rozwiązania całego problemu.

Skuteczność takich algorytmów zależy do rodzaju problemu: w znacznej części zagadnień mogą one nie być w stanie dojść do rozwiązania optymalnego w sensie globalnym.

Składanie sekwencji genomu w strategii *shotgun*: algorytm zachłanny

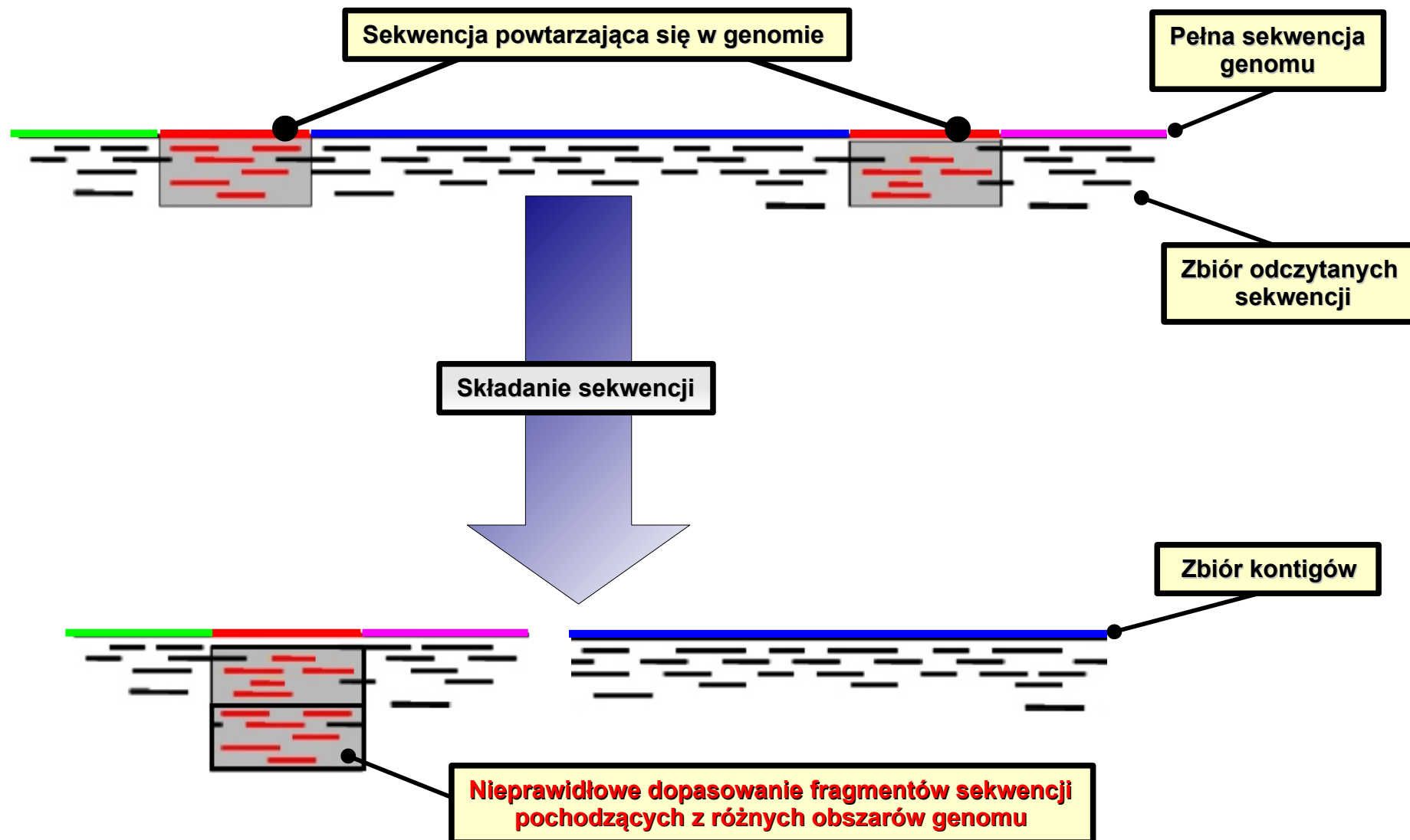
W przypadku składania genomu główną **przeszkodą na drodze do osiągnięcia rozwiązania optymalnego w sensie globalnym jest występowanie identycznych lub bardzo zbliżonych sekwencji w różnych obszarach genomu**, co niestety jest cechą charakterystyczną genomów wyższych organizmów eukariotycznych.



[T. A. Brown: *Genomy*. PWN, 2009]

Składanie sekwencji genomu w strategii *shotgun*: algorytm zachłanny

Obecność powtórzeń podważa podstawowe założenie strategii *shotgun*, mówiące że dwie takie same sekwencje pochodzą z jednego miejsca genomu. Może to prowadzić do poważnych błędów w procesie składania, skutkujących połączeniem fragmentów sekwencji, które w rzeczywistości są od siebie odległe.

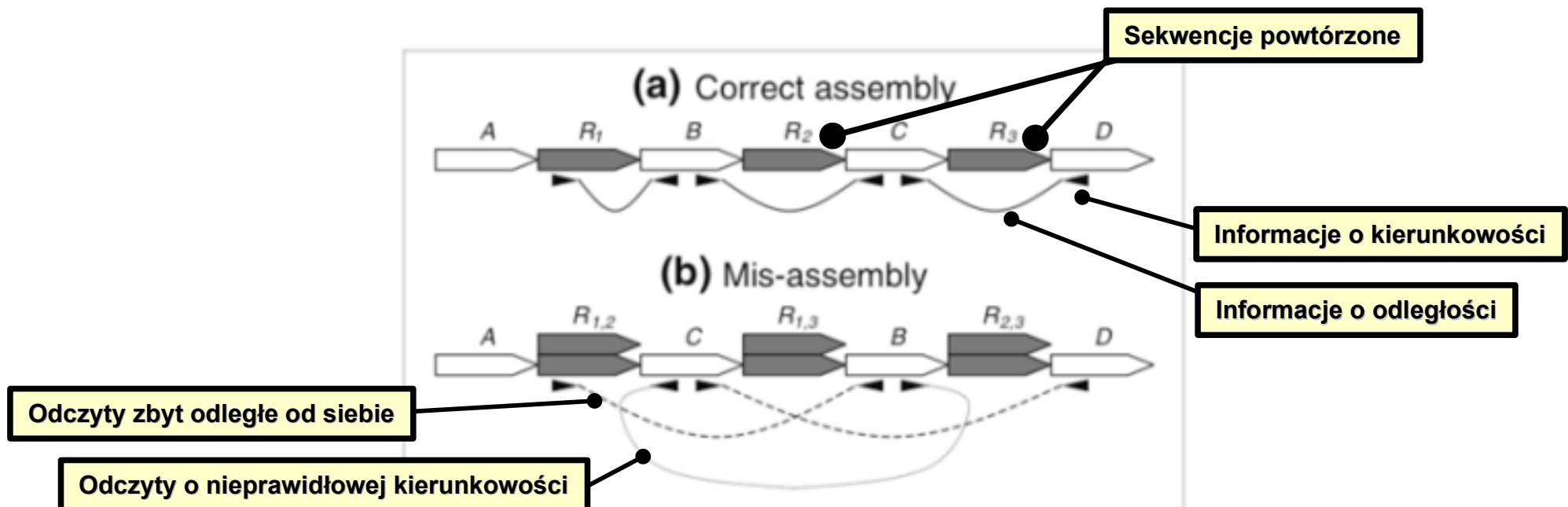


Składanie sekwencji genomu w strategii *shotgun*: algorytm zachłanny

Algorytm składania genomu możliwy do zastosowania w praktyce musi uwzględniać możliwość powtórzeń sekwencji w różnych miejscach genomu.

W tym kontekście bardzo istotne jest **wykorzystanie informacji o kierunkowości par odczytanych końców** (jest ona zapamiętywana na etapie sekwencjonowania) oraz **przybliżonej odległości między nimi** (wynika ona ze znajomości długości fragmentu, z którego pochodzą odczyty).

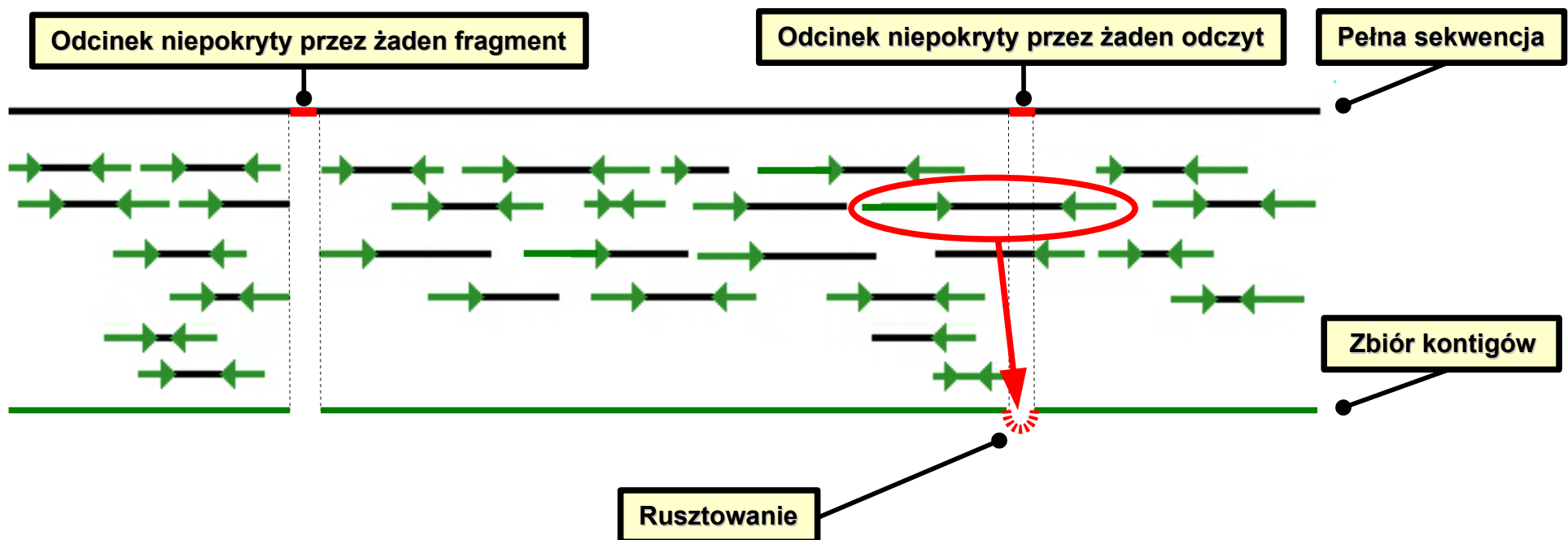
Użycie tych informacji pozwala wykryć błędy w procesie składania poprzez sprawdzanie czy w wyjściowej sekwencji pary odczytanych końców nie znajdują się zbyt blisko/daleko od siebie i czy ich wzajemna orientacja jest prawidłowa.



Składanie sekwencji genomu w strategii *shotgun*: łączenie kontigów

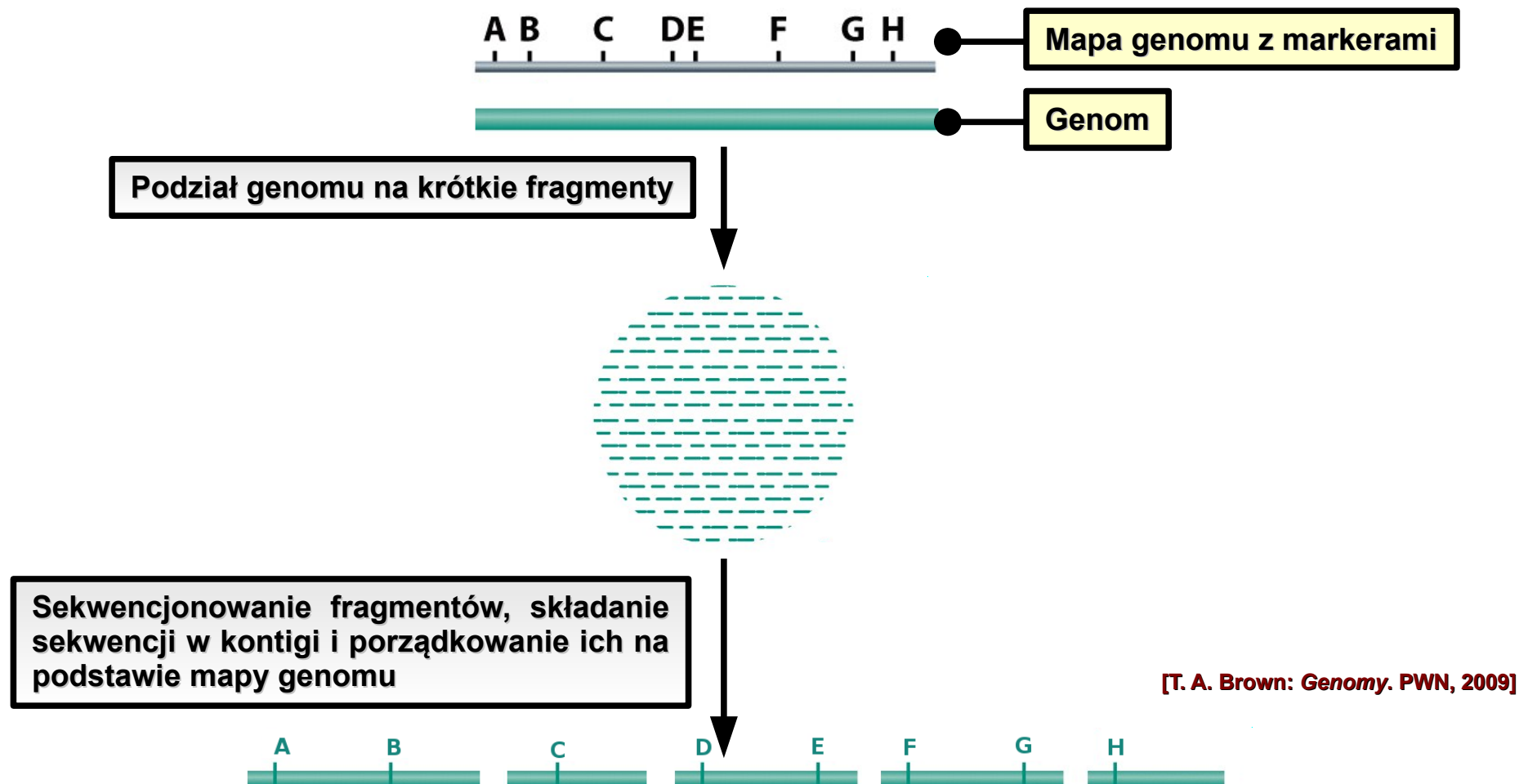
Po zakończeniu składania konieczne jest jeszcze ustalenie kolejności kontigów i uzupełnienie luk pomiędzy nimi. To drugie zadanie wymaga dodatkowych pomiarów, ukierunkowanych na wybrane obszary genomu. Natomiast ustalenie kolejności można osiągnąć przez analizę danych z sekwencjonowania i/lub wykorzystanie pewnej wiedzy wstępnej o budowie badanego genomu.

W szczególnych przypadkach możliwe jest automatyczne określenie kolejności i orientacji kontigów na podstawie kierunkowości rozdzielonych pomiędzy nimi par odczytów, co określane jest jako **tworzenie rusztowania (*scaffolding*)**.



Składanie sekwencji genomu w strategii *shotgun*: łączenie kontigów

Najczęściej jednak skuteczne łączenie kontigów genomów o dużej złożoności nie jest możliwe bez wykorzystania wiedzy *a priori* o ich budowie. Dlatego też etap sekwencjonowania zwykle jest poprzedzany etapem **mapowania**, czyli określania położenia pewnych charakterystycznych sekwencji, które służą następnie jako punkty odniesienia (markery) podczas ustalania kolejności kontigów.



Uczenie maszynowe w bioinformatyce

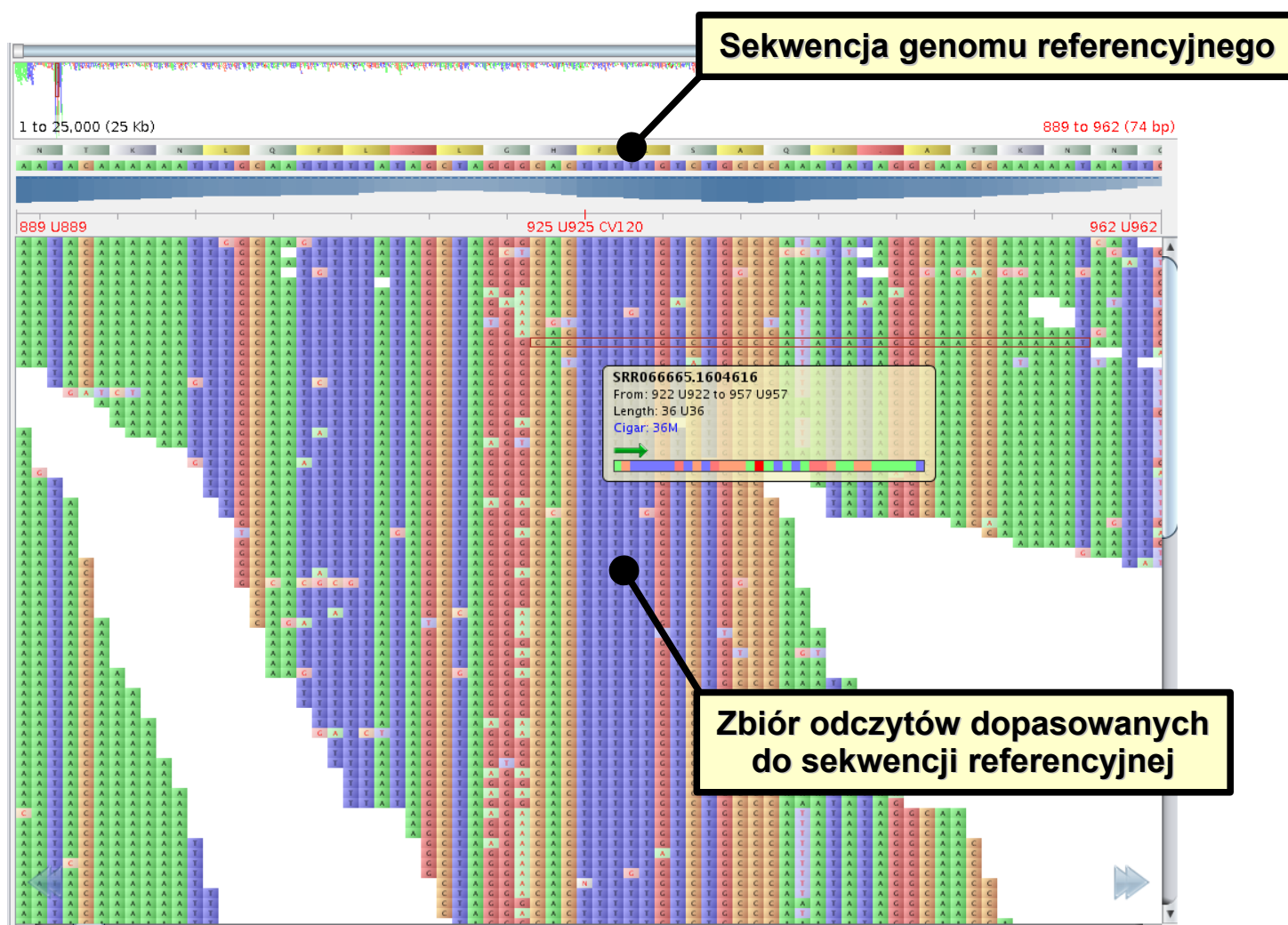
Wykład 4: sekwencjonowanie genomu osobnika

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Sekwencjonowanie genomu: od organizmu do osobnika

Nie zawsze jest konieczność wykonywania składania całkowitej sekwencji *de novo*. W przypadku gdy wykonujemy resekwencjonowanie znanego już genomu (np. dla konkretnego osobnika) możliwe jest wykonanie dopasowywania (mapowania) odczytów do genomu referencyjnego. Polega ono na określeniu pozycji nowych odczytów w odniesieniu do znanej sekwencji referencyjnej.



Sekwencjonowanie genomu: od organizmu do osobnika

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US\$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	<i>De novo</i>	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2 ⁺	250 [§]	7.4	Aligned*	95	3.32 (BLAT)	234	1,000,000 [¶]
Yoruban male (NA18507)	Illumina/Solexa	93% MP: 200 bp	3,410 ⁺	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000 [¶]
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/Solexa	66% Frag: 150–250 bp	1,921 ⁺	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000 [¶]
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/Solexa	21% Frag: 130 bp & 440 bp	393 ⁺	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000 [¶]
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36,88,106						
Korean male (SJK)	Illumina/Solexa	MP: 100 bp, 200 bp & 300 bp	1,647 ⁺	35,74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000 ^{¶,§}
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211 ⁺	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000 ^{¶,***}
		91% MP: 600–3,500 bp	2,075 ⁺	25,50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725 ⁺	32 [§]	28	Aligned*	90	2.81 (IndexDP)	4	48,000 [¶]
AML female	Illumina/Solexa	Frag: 150–200 bp ^{††}	2,730 ^{†,††}	32	32.7	Aligned*	91	3.81 ^{††} (MAQ)	98	1,600,000 ^{¶¶}
		Frag: 150–200 bp ^{§§}	1,081 ^{†,§§}	35	13.9		83	2.92 ^{§§} (MAQ)	34	
AML male	Illumina/Solexa	MP: 200–250 bp ^{††}	1,620 ^{†,††}	35	23.3	Aligned*	98.5	3.46 ^{††} (MAQ)	16.5	500,000 ^{¶¶}
		MP: 200–250 bp ^{§§}	1,351 ^{†,§§}	50	21.3		97.4	3.45 ^{§§} (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238 ⁺	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000 ^{¶,¶¶}
		84% MP: 600–3,500 bp	1,211 ⁺	25,50						

[Michael L. Metzker: Sequencing technologies – the next generation. Nature Reviews Genetics, 2010, 11, pp. 31-46.]

Uczenie maszynowe w bioinformatyce

Wykład 4: bazy danych genomów

Tymon Rubel

**Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW**

Sekwencjonowanie genomu: bazy danych genomów

EMBL-EBI

Enter Text Here

Databases Tools Research Training Industry About Us Help

- Complete genomes
- Archaea
- Archaeal virus
- Bacteria
- Eukaryota
- Organelle
- Phage
- Plasmid
- Viroid
- Virus
- Links
 - WGS info
 - EnsemblGenomes
 - Genome Reviews
 - Integr8 (proteomes)
 - Fasta33 Server
 - Ensembl

EBI > Databases > Nucleotide > The European Nucleotide Archive > Complete Genomes

Genomes Pages - At the EBI

Access to Completed Genomes

The first completed genomes from [viruses](#), [phages](#) and [organelles](#) were deposited into the EMBL Database in the early 1980's. Since then, molecular biology's shift to obtain the complete sequences of as many genomes as possible combined with major developments in sequencing technology resulted in hundreds of complete genome sequences being added to the database, including [Archaea](#), [Bacteria](#) and [Eukaryota](#). These web pages give access to a large number of complete genomes, [help](#) is available to describe the layout.

Whole Genome Shotgun Sequences (WGS)

Methods using whole genome shotgun data are used to gain a large amount of genome coverage for an organism. WGS data for a growing number of organisms are being submitted to DDBJ/EMBL/GenBank.

[More information about WGS projects...](#)

Last 40 Genome Entries

Date	Accession	Description
14-OCT-2011	FO082060.1	Methylobacterium alcaliphilum str. 20Z chromosome
14-OCT-2011	FO082061.1	Methylobacterium alcaliphilum str. 20Z plasmid MEALZ_p
14-OCT-2011	JN242813.1	Heterocephalus glaber mitochondrion

<http://www.ebi.ac.uk/genomes/>

<http://www.ncbi.nlm.nih.gov/sites/genome>



Format FASTA (informacje o sekwencji)

```
>ENA|CM000252|CM000252.1 Homo sapiens chromosome 1, whole genome shotgun sequence.
TCACCTGGGTGTGTGGGTGCCGTTCCAGGCTGTCAGAGCTCGCGTGGGGGTGTGGGTGCT
GCTCCAGGCTTTTCGGAGCTCACCTGGGGGTGCAGGGTGCTGTTCCAGGCTGTCAGATGCT
CACCTGGGGGTGTGGTTGCTGCTCCAGGCTGTCAGATGCTCACTTGGGGGTGCAGCGTGC
TGTTCCAGGCTGTCAGATGCTAACCTGGGGTTGTGAGAGCTGTTGCAGGCTGTCAGATGC
TCACTGGGGGTGTGCGTGCTGCTCCAGCCTGTCAGATGCTCACCTGGGGGTGTGGGTGCT
GTTCCAGGATATCAGATGCTCACCTGGGGGTGTGGGTGCTGCTCCAGGCTGTCGGATGCT
CACCTGGGGGTGTCGGTGCTGTTCCAGGCTGTTGGATGCTCACCTGGGGGTGTGGTTGCT
GCTCCAGGCTGTCAGGTGCTCACTTGGGGGTGCAGCGTGCTGTTCCGGGCTGTCAGATGC
TCACCTGGGGTTGTGAGAGCTGTTCCAGGCTGTCAGATGCTCACCTGGGGGTGTGGGTGC
```

Sekwencjonowanie genomu: bazy danych genomów

EMBL-EBI

Enter Text Here

Databases Tools Research Training Industry About Us Help

Complete genomes Archaea

EBI > Databases > Nucleotide > The European Nucleotide Archive > Complete

Genomes Pages - At the EBI

<http://www.ebi.ac.uk/genomes/>

<http://www.ncbi.nlm.nih.gov/sites/genome>



Since then,
its in
Bacteria and

Format GenBank (informacje o sekwencji i jej adnotacjach)

LOCUS NC_012920 16569 bp DNA circular PRI 31-OCT-2014
DEFINITION Homo sapiens mitochondrion, complete genome.
ACCESSION NC_012920 AC_000021
VERSION NC_012920.1 GI:251831106
KEYWORDS RefSeq.
SOURCE mitochondrion Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Mammalia; Eutheria; Euarchontoglires; Primates;
Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 16569)
AUTHORS Andrews,R.M., Kubacka,I., Chinnery,P.F., Lightowlers,C.N.,
TITLE Reanalysis and revision of the Cambridge reference
human mitochondrial DNA

FEATURES

	Location/Qualifiers
source	1..16569 /organism="Homo sapiens" /organelle="mitochondrion" /mol_type="genomic DNA" /db_xref="taxon:9606"
gene	577..647 /gene="TRNF"
	...

ORIGIN

1 gatcacaggt ctatcaccct attaaccact cacgggagct ctccatgcat ttggtatattt
61 cgtctggggg gtatgcacgc gatagcattg cgagacgctg gagccggagc accctatgtc
... ..

//

