Algorytmy w bioinformatyce sekwencji

Istotność wyników badania podobieństw i wyszukiwania

Robert Nowak

2024

Powtórzenie

Podobieństwo:

- 2 sekwencji (globalne, lokalne, ...)
- wielu sekwencji (dokładne, profile)

Wyszukiwanie:

- sekwencji podobnych do danej
- badanie podobieństw: heurystyczne (FASTA, BLAST)

Istotność wyników

Istotność odnalezionych sekwencji

- dwie dowolne sekwencje zawsze można dopasować
- wyszukiwanie w bazie sekwencji podobnych do danej zawsze zwróci wyniki

- Czy dopasowanie wskazuje na podobieństwo, czy jest wynikiem przypadku?

Rozkład ocen dopasowań dla par losowych sekwencji

Uproszczenia:

- sekwencje bez przerw o tej samej długości n
- wszystkie k symbole występują z tym samym prawdopodobieństwem $p = \frac{1}{k}$
- prawdopodobieństwo identycznych symboli

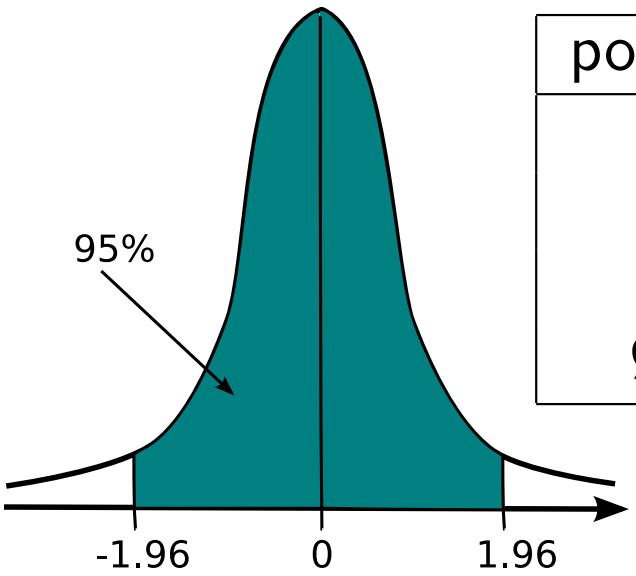
$$\sum_{0}^{k} p^2 = kp^2 = p$$

- miarą podobieństwa m jest liczba identycznych symboli

$$\begin{aligned} P(m) &= \binom{n}{m} p^m (1-p)^{n-m} \\ m_{sr} &= np \\ \sigma &= \sqrt{np(1-p)} \end{aligned}$$

Rozkład ocen dopasowań dla par losowych sekwencji (2)

Gdy $np \geq 3$ oraz $n(1-p) \geq 3$ to rozkład dwumianowy[1] można przybliżać rozkładem normalnym $P(m) \approx N(m_{sr}, \sigma)$



poz. ufności	zakres
90%	$(m_{sr} - 1.65\sigma, m_{sr} + 1.65\sigma)$
95%	$(m_{sr} - 1.96\sigma, m_{sr} + 1.96\sigma)$
99%	$(m_{sr} - 2.58\sigma, m_{sr} + 2.58\sigma)$
99.74%	$(m_{sr} - 3\sigma, m_{sr} + 3\sigma)$

[1] rozkład Bernoulliego

Rozkład ocen dopasowań dla par losowych sekwencji (3)

Przykład:

GAATTCGAATTC
GATGAAGATTAA

$m_1 = 5, n = 12, p = 0.25, m_{sr} = 3, \sigma = 1.5$

$$z = \frac{m_1 - m_{sr}}{\sigma} = 1.34 \text{(podobieństwo jest przypadkowe)}$$

Przykład 2: $m_2 = 350, n = 1200, p = 0.25, m_{sr} = 300, \sigma = 15$

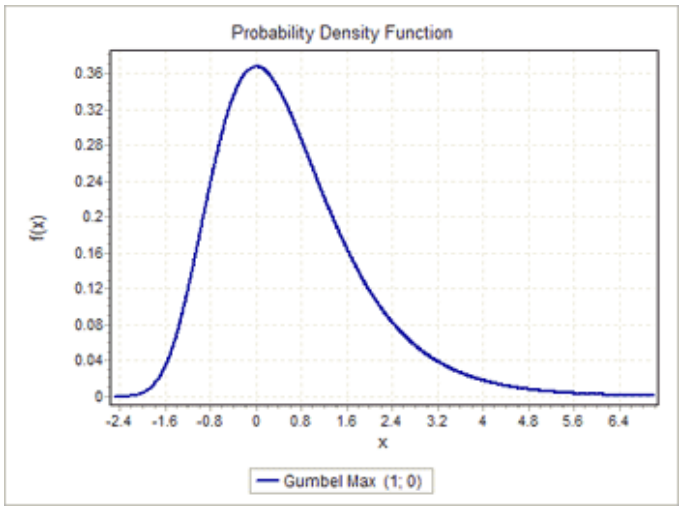
$$z = \frac{m_2 - m_{sr}}{\sigma} = 3.33 \text{(podobieństwo nie jest przypadkowe)}$$

Prawdopodobieństwo, że sekwencje o takim (lub większym) dopasowaniu są przypadkowe wynosi $\approx 0.05\%$

Rozkład ocen dla maksymalnie zgodnej sekwencji z bazy

Uproszczenia:

- sekwencje bez przerw o tej samej długości n
- wszystkie symbole występują z tym samym prawd.
- miarą podobieństwa m jest liczba identycznych symboli
- ocena to m_{max} dla wszystkich N sekwencji z bazy



$$P(m_{max}) = \lambda e^{-\lambda(m_{max}-u)} e^{-e^{-\lambda m_{max}-u}}$$

rozkład wartości ekstremalnej (EVD, extreme value distribution) lub rozkład Gumbela

- rozkład uwzględniając przerwy jest podobny (symulacje)
- dla danego algorytmu wyszukiwania estymuje się parametry tego rozkładu

E-value - spodziewana liczba wyników

$$E = kmne^{-\lambda S}$$

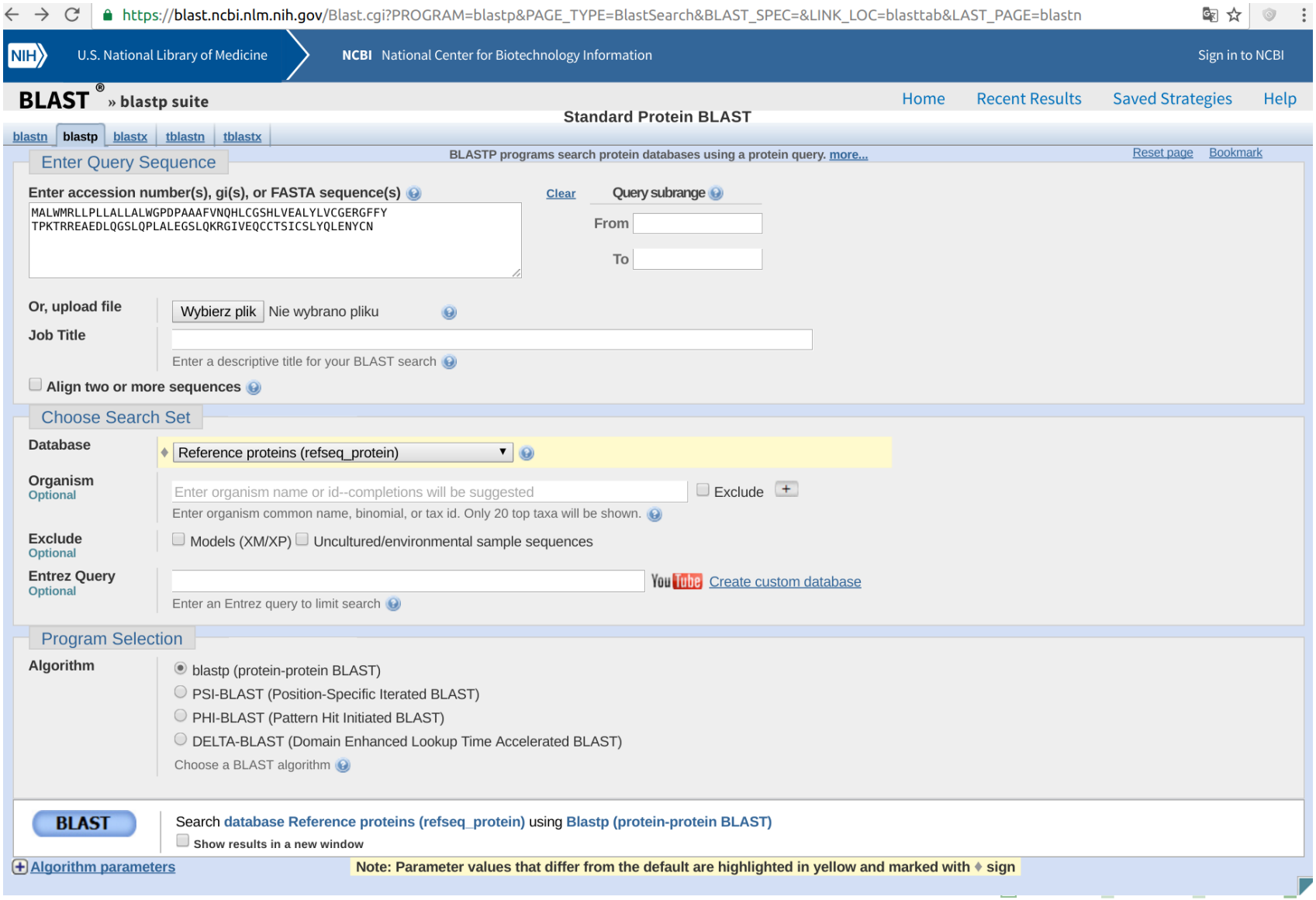
gdzie:

- m długość poszukiwanej sekwencji
- n liczbie rekordów w bazie danych
- S ocena podobieństwa
- λ, k współczynniki zależne od macierzy podobieństwa

Przykład: $E = 5$, baza danych z losowymi sekwencjami zwróci 5 rekordów

Jeżeli $E \ll 1$ oznacza to, że wynik jest istotny.

Przykład - uruchomienie algorytmu



Przykład - wyniki

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
insulin preproprotein [Homo sapiens]	192	192	100%	8e-65	89%	NP_000198.1
insulin-2 preproprotein [Mus musculus]	140	140	100%	4e-44	73%	NP_032413.1
insulin-1 preproprotein [Mus musculus]	137	137	100%	5e-43	73%	NP_032412.3
insulin, isoform 2 precursor [Homo sapiens]	112	112	63%	5e-32	100%	NP_001035835.1
insulin-like growth factor II isoform 1 preproprotein [Homo sapiens]	53.9	53.9	92%	2e-09	38%	NP_000603.1
insulin-like growth factor II isoform 2 [Homo sapiens]	53.9	53.9	92%	3e-09	38%	NP_001121070.1
insulin-like growth factor II isoform 2 preproprotein [Mus musculus]	50.8	50.8	98%	3e-08	32%	NP_001116208.1

insulin-1 preproprotein [Mus musculus]
Sequence ID: NP_032412.3 Length: 108 Number of Matches: 1

Range 1: 1 to 108 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
137 bits(346)	5e-43	Compositional matrix adjust.	79/108(73%)	83/108(76%)	10/108(9%)

```
Query  1    MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60
            MAL +  LPLLALLALW P P  AFV QHLCG HLVEALYLVCGERGFFYTPK+RRE ED
Sbjct  1    MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVED  60
Query  61   LQ----------GSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  98
             Q          G LQ LALE + QKRGIV+QCCTSICSLYQLENYCN
Sbjct  61   PQVEQLELGGSPGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN  108
```

Related Information
Gene - associated gene details
UniGene - clustered expressed sequence tags
Map Viewer - aligned genomic context

Download GenPept Graphics ▼ Next ▲ Previous ▲ Descriptions

insulin, isoform 2 precursor [Homo sapiens]
Sequence ID: NP_001035835.1 Length: 200 Number of Matches: 1

Range 1: 1 to 62 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
112 bits(281)	5e-32	Compositional matrix adjust.	62/62(100%)	62/62(100%)	0/62(0%)

```
Query  1   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60
           MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
Sbjct  1   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60
Query  61  LQ  62
           LQ
Sbjct  61  LQ  62
```

Related Information
Gene - associated gene details
Map Viewer - aligned genomic context

Parametry testów binarnych

Parametry testów binarnych

macierz pomyłek		stan	
		plus	minus
wynik	dodatni	prawdziwie dodatni(TP)	fałszywie dodatni(FP)
	ujemny	fałszywie ujemny(FN)	prawdziwie ujemny(TN)

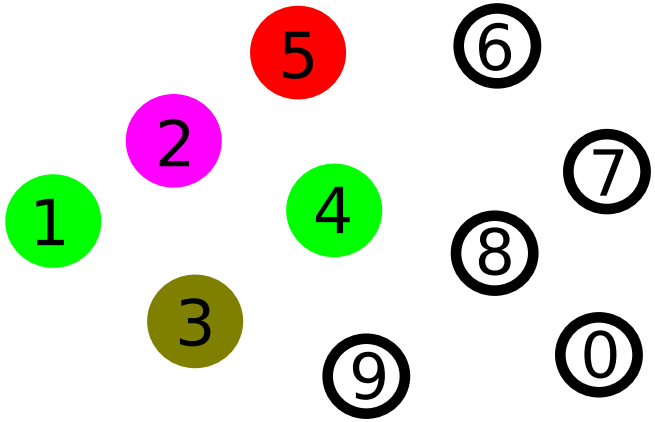
$$\text{czułość (sensitivity)} = \frac{TP}{TP + FN}, \text{swoistość (specifity)} = \frac{TN}{TN + FP}$$

$$\text{precyzja (precision)} = \frac{TP}{TP + FP}$$

$$\text{dokładność} = \frac{TP + TN}{TP + TN + FP + FN}, \text{błąd} = 1 - \text{dokładność}$$

$$\text{F1-score} = 2 * \frac{PPV * TPR}{PPV + TPR}$$

Parametry testów binarnych (2)



TP = 4	FP = 2
FN = 1	TN = 3

czułość = 0.8
swoistość = 0.6
precyzja = 0.67
dokładność = 0.7
F1-score = 0.73

Test na kolor:

nr	stan	wynik testu
0	NIE	NIE
1	TAK	NIE
2	TAK	TAK
3	TAK	TAK
4	TAK	TAK
5	TAK	TAK
6	NIE	TAK
7	NIE	TAK
8	NIE	NIE
9	NIE	NIE

Krzywe ROC (Receiver Operating Characteristics)

Graficzna ocena skuteczności testu, oś X to FPR = 1 - swoistość, oś Y to czułość (TPR).

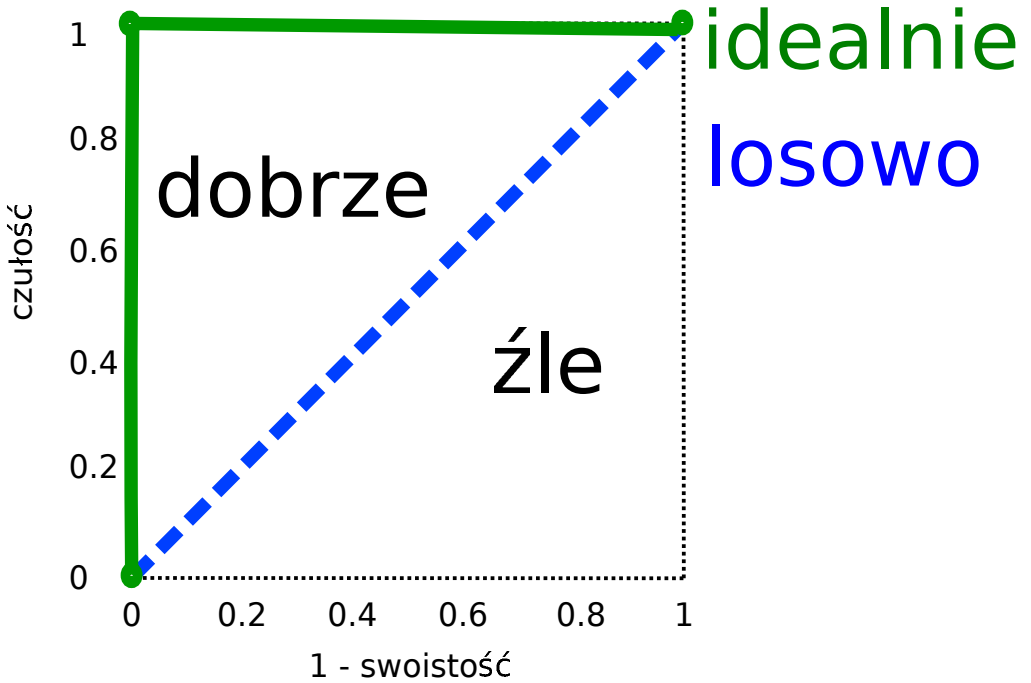
Punkty charakterystyczne:

(0,0) wszystkie przykłady są ujemne (TP=0)

(1,1) wszystkie przykłady są dodatnie (TN=0)

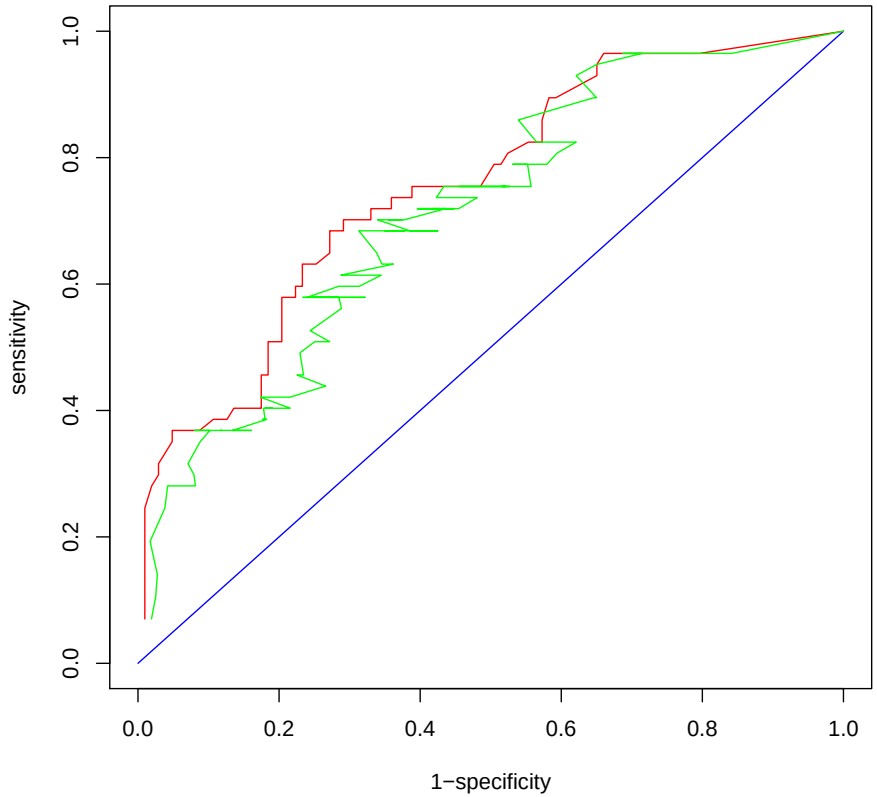
(0,1) test idealny

$y = x$ strategia losowa



Krzywe ROC - porównanie modeli

AUC ROC (area under ROC curve) – jakość klasyfikatora



model A jest lepszy niż model B, jeżeli w każdym punkcie krzywa ROC dla A jest powyżej ROC dla B

Krzywa PR (Precision-Recall)

graficzna ocena skuteczności testu binarnego; oś X to precyzja (precision) oś Y to czułość (recall, TPR).

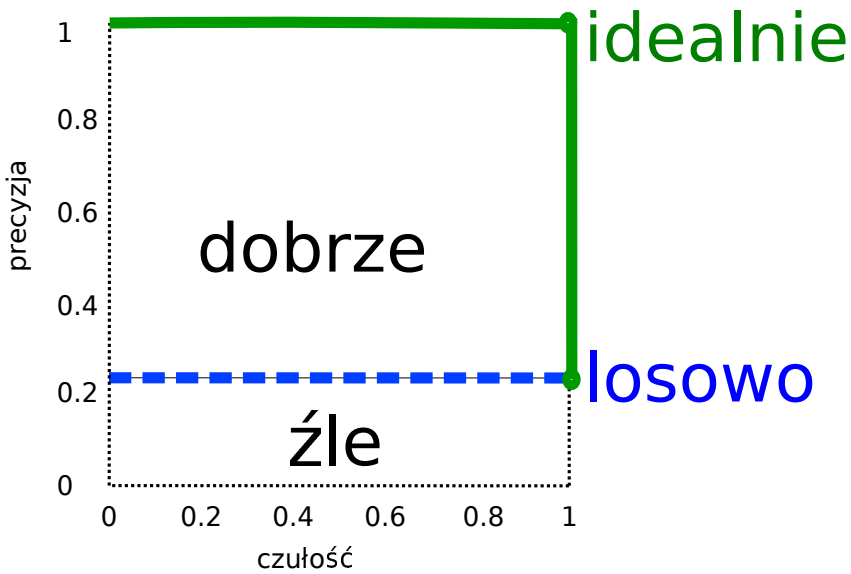
Punkty charakterystyczne:

- nie można obliczyć, gdy wszystkie przykłady są ujemne (TP=0)

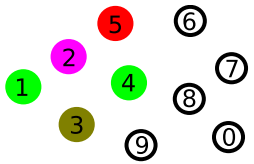
$(1,1)$ test idealny

$(1, \frac{p}{p+n})$ wszystkie przykłady są dodatnie

$y = \frac{p}{p+n}$ strategia losowa (odestek przykładów pozytywnych w zbiorze)



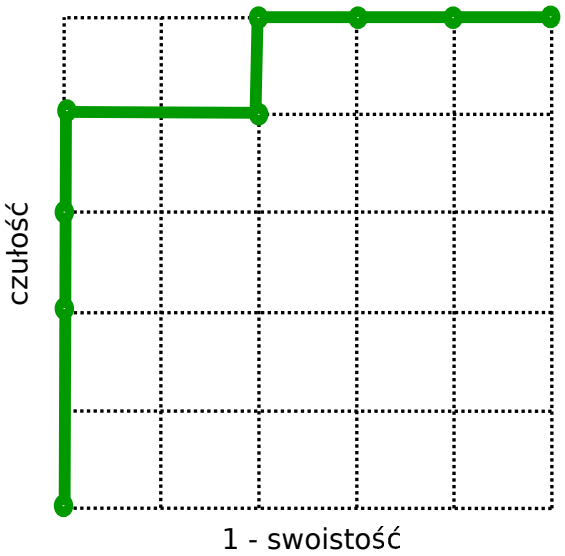
Krzywa ROC i PR - przykład tworzenia



nr	stan	wynik
0	NIE	0.1
1	TAK	0.4
2	TAK	0.9
3	TAK	0.9
4	TAK	0.8
5	TAK	0.7
6	NIE	0.6
7	NIE	0.6
8	NIE	0.3
9	NIE	0.2

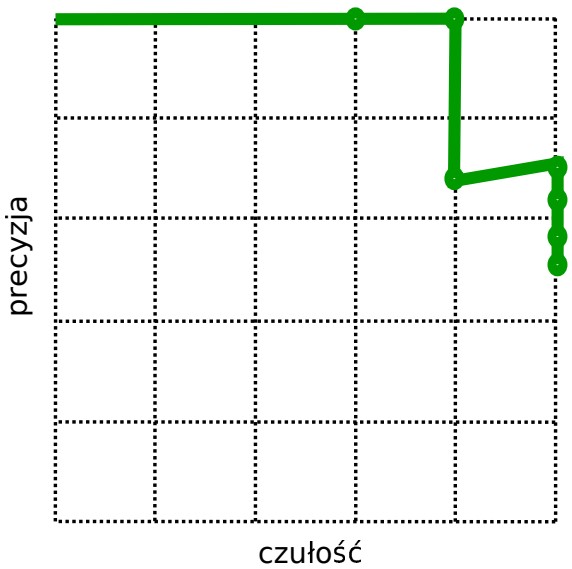
próg	macierz pom.		czuł	sw.	prec.
0.05	TP = 5	FP = 5	1.0	0.0	0.500
	FN = 0	TN = 0			
0.15	TP = 5	FP = 4	1.0	0.2	0.556
	FN = 0	TN = 1			
0.25	TP = 5	FP = 3	1.0	0.4	0.625
	FN = 0	TN = 2			
0.35	TP = 5	FP = 2	1.0	0.6	0.714
	FN = 0	TN = 3			
0.45	TP = 4	FP = 2	0.8	0.6	0.667
	FN = 1	TN = 3			
0.55	TP = 4	FP = 2	0.8	0.6	1.000
	FN = 1	TN = 3			
0.65	TP = 4	FP = 0	0.8	1.0	1.000
	FN = 1	TN = 5			
0.75	TP = 3	FP = 0	0.6	1.0	1.000
	FN = 2	TN = 5			
0.85	TP = 2	FP = 0	0.4	1.0	1.000
	FN = 3	TN = 5			
0.95	TP = 0	FP = 0	0.0	1.0	-
	FN = 5	TN = 5			

Krzywa ROC i PR - przykład tworzenia



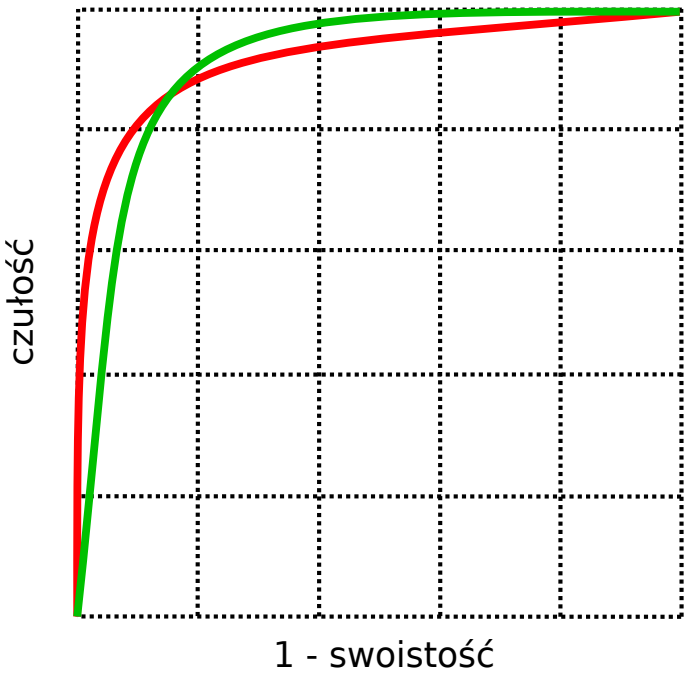
próg	macierz pom.		czuł	sw.	prec.
0.05	TP = 5	FP = 5	1.0	0.0	0.500
	FN = 0	TN = 0			
0.15	TP = 5	FP = 4	1.0	0.2	0.556
	FN = 0	TN = 1			
0.25	TP = 5	FP = 3	1.0	0.4	0.625
	FN = 0	TN = 2			
0.35	TP = 5	FP = 2	1.0	0.6	0.714
	FN = 0	TN = 3			
0.45	TP = 4	FP = 2	0.8	0.6	0.667
	FN = 1	TN = 3			
0.55	TP = 4	FP = 2	0.8	0.6	1.000
	FN = 1	TN = 3			
0.65	TP = 4	FP = 0	0.8	1.0	1.000
	FN = 1	TN = 5			
0.75	TP = 3	FP = 0	0.6	1.0	1.000
	FN = 2	TN = 5			
0.85	TP = 2	FP = 0	0.4	1.0	1.000
	FN = 3	TN = 5			
0.95	TP = 0	FP = 0	0.0	1.0	-
	FN = 5	TN = 5			

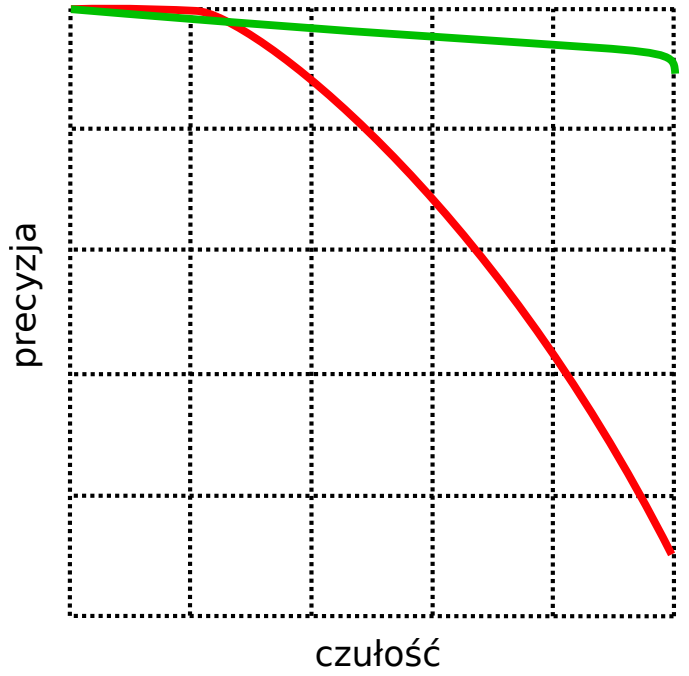
Krzywa ROC i PR - przykład tworzenia



próg	macierz pom.		czuł	sw.	prec.
0.05	TP = 5	FP = 5	1.0	0.0	0.500
	FN = 0	TN = 0			
0.15	TP = 5	FP = 4	1.0	0.2	0.556
	FN = 0	TN = 1			
0.25	TP = 5	FP = 3	1.0	0.4	0.625
	FN = 0	TN = 2			
0.35	TP = 5	FP = 2	1.0	0.6	0.714
	FN = 0	TN = 3			
0.45	TP = 4	FP = 2	0.8	0.6	0.667
	FN = 1	TN = 3			
0.55	TP = 4	FP = 2	0.8	0.6	1.000
	FN = 1	TN = 3			
0.65	TP = 4	FP = 0	0.8	1.0	1.000
	FN = 1	TN = 5			
0.75	TP = 3	FP = 0	0.6	1.0	1.000
	FN = 2	TN = 5			
0.85	TP = 2	FP = 0	0.4	1.0	1.000
	FN = 3	TN = 5			
0.95	TP = 0	FP = 0	0.0	1.0	-
	FN = 5	TN = 5			

Krzywe ROC i PR gdy nie ma równowagi klas

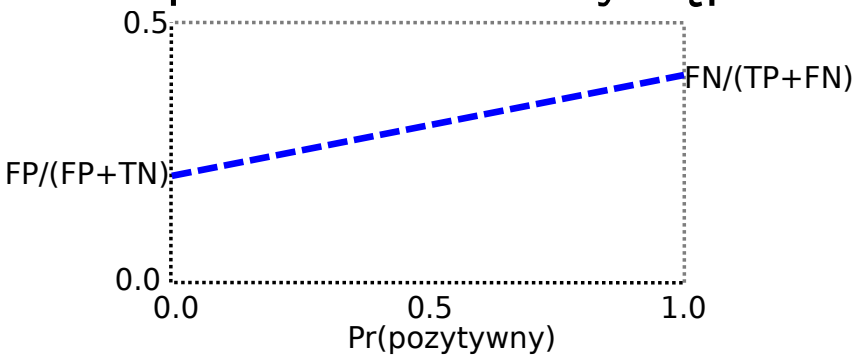


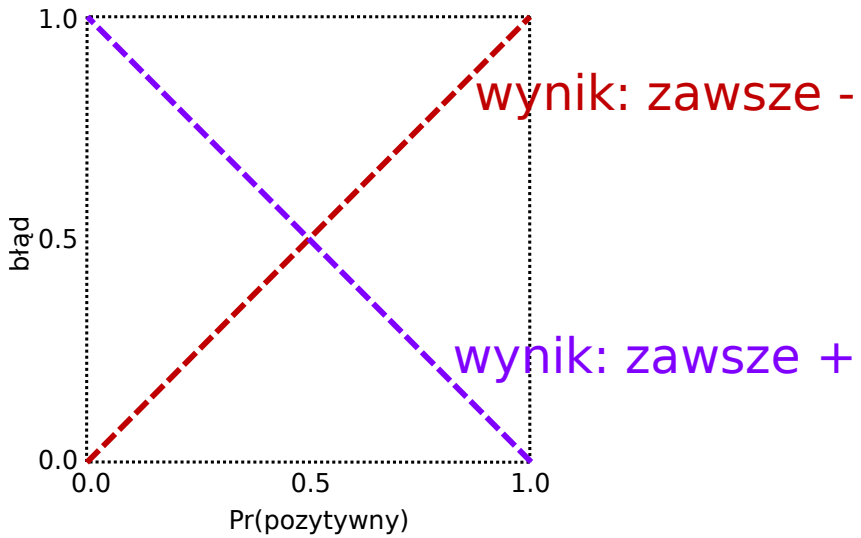


Krzywa kosztu (cost curve)

zakładamy koszt(FP) = koszt(FN)

graficzna ocena skuteczności testu binarnego: oś X to prawdopodobieństwo wystąpienia klasy plus, oś Y to błąd.



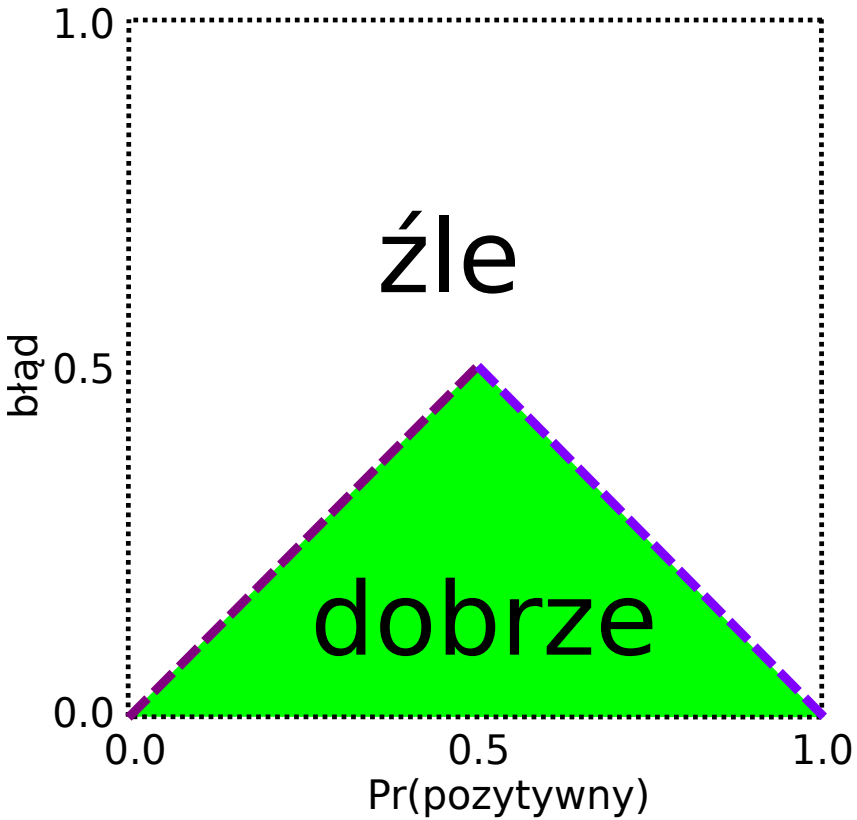


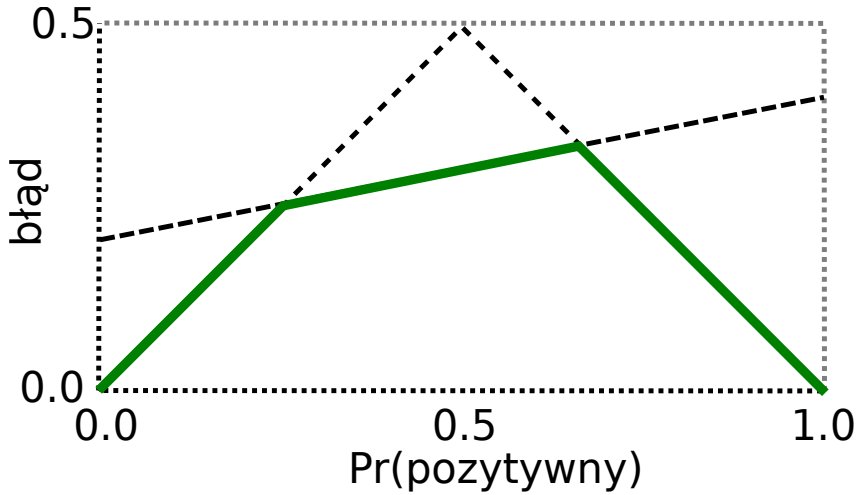
macierz pomyłek to prosta na krzywej kosztu, łączy punkty $(0, \frac{FN}{\Sigma})$ i $(1, \frac{FP}{\Sigma})$

Krzywa kosztu (2)

Możemy wybierać najlepszy z wyników:

- zawsze zwracamy etykietę 'negatywny'
- zwracamy wynik testu
- zawsze zwracamy etykietę 'pozytywny'





Krzywa ROC, PR i kosztu - przykład 2

osoba	stan	wynik testu
A	zdrowa	0.1
B	zdrowa	0.1
C	zdrowa	0.2
D	zdrowa	0.3
E	zdrowa	0.6
F	chora	0.6
G	chora	0.7
H	chora	0.8
I	chora	0.9
J	chora	0.9

Próg 0.35:

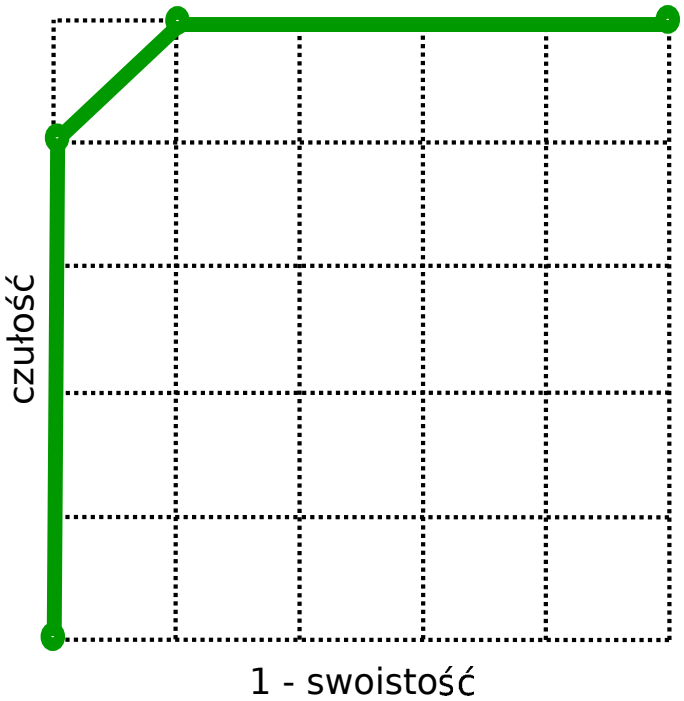
$TP = 5$	$FP = 1$
$FN = 0$	$TN = 4$

Próg 0.5:

$TP = 5$	$FP = 1$
$FN = 0$	$TN = 4$

Próg 0.65:

$TP = 4$	$FP = 0$
$FN = 1$	$TN = 5$



Krzywa ROC, PR i kosztu - przykład 2

osoba	stan	wynik testu
A	zdrowa	0.1
B	zdrowa	0.1
C	zdrowa	0.2
D	zdrowa	0.3
E	zdrowa	0.6
F	chora	0.6
G	chora	0.7
H	chora	0.8
I	chora	0.9
J	chora	0.9

Próg 0.35:

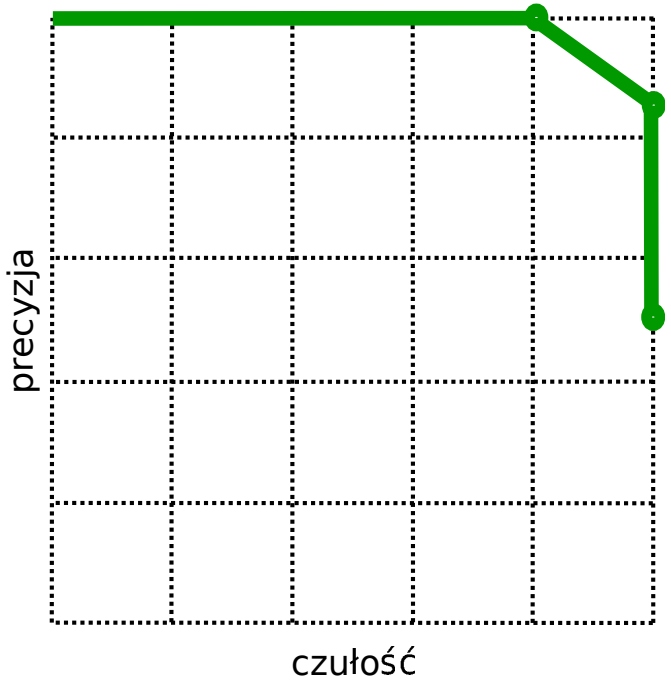
$TP = 5$	$FP = 1$
$FN = 0$	$TN = 4$

Próg 0.5:

$TP = 5$	$FP = 1$
$FN = 0$	$TN = 4$

Próg 0.65:

$TP = 4$	$FP = 0$
$FN = 1$	$TN = 5$



Krzywa ROC, PR i kosztu - przykład 2

osoba	stan	wynik testu
A	zdrowa	0.1
B	zdrowa	0.1
C	zdrowa	0.2
D	zdrowa	0.3
E	zdrowa	0.6
F	chora	0.6
G	chora	0.7
H	chora	0.8
I	chora	0.9
J	chora	0.9

Próg 0.35:

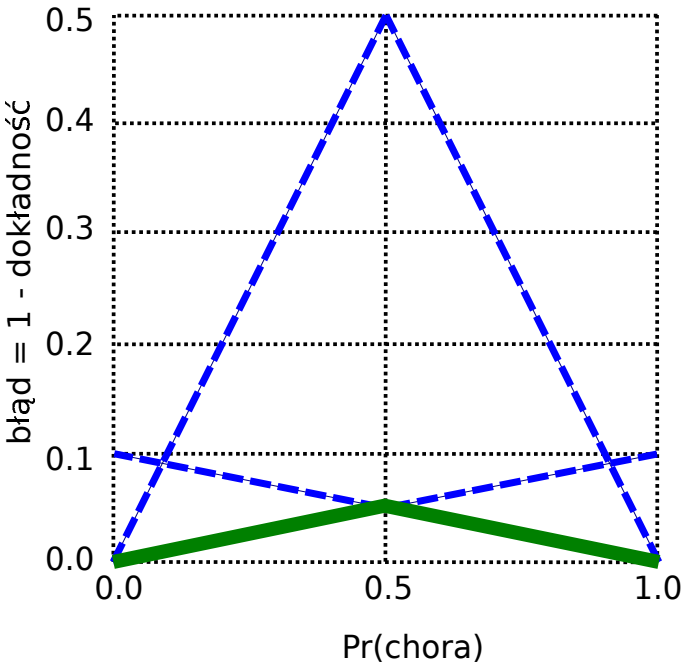
$TP = 5$	$FP = 1$
$FN = 0$	$TN = 4$

Próg 0.5:

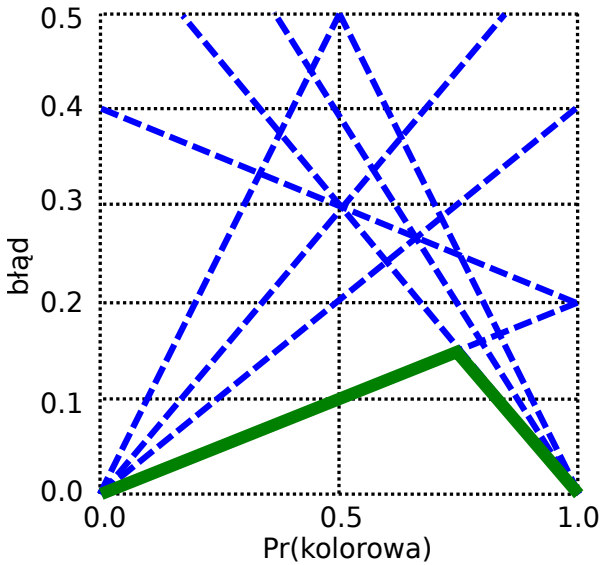
$TP = 5$	$FP = 1$
$FN = 0$	$TN = 4$

Próg 0.65:

$TP = 4$	$FP = 0$
$FN = 1$	$TN = 5$



Krzywa kosztu - przykład 2



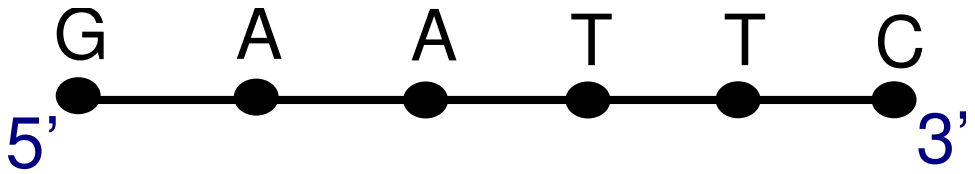
próg	macierz pom.		czuł	sw.	prec.
0.05	TP = 5 FN = 0	FP = 5 TN = 0	1.0	0.0	0.500
0.15	TP = 5 FN = 0	FP = 4 TN = 1	1.0	0.2	0.556
0.25	TP = 5 FN = 0	FP = 3 TN = 2	1.0	0.4	0.625
0.35	TP = 5 FN = 0	FP = 2 TN = 3	1.0	0.6	0.714
0.45	TP = 4 FN = 1	FP = 2 TN = 3	0.8	0.6	0.667
0.55	TP = 4 FN = 1	FP = 2 TN = 3	0.8	0.6	1.000
0.65	TP = 4 FN = 1	FP = 0 TN = 5	0.8	1.0	1.000
0.75	TP = 3 FN = 2	FP = 0 TN = 5	0.6	1.0	1.000
0.85	TP = 2 FN = 3	FP = 0 TN = 5	0.4	1.0	1.000
0.95	TP = 0 FN = 5	FP = 0 TN = 5	0.0	1.0	-

Sekwenatory

Sekwencjonowanie

Sekwencjonowanie DNA - proces ustalania kolejności nukleotydów tworzących cząsteczkę DNA.

- sekwencja pierwszorzędowa - nić DNA jest reprezentowana przez napis
- mapa fizyczna



Obecnie proces składa się z:

1. odczytu sekwencji fragmentów
2. składania (asemblacja)
3. wykańczania

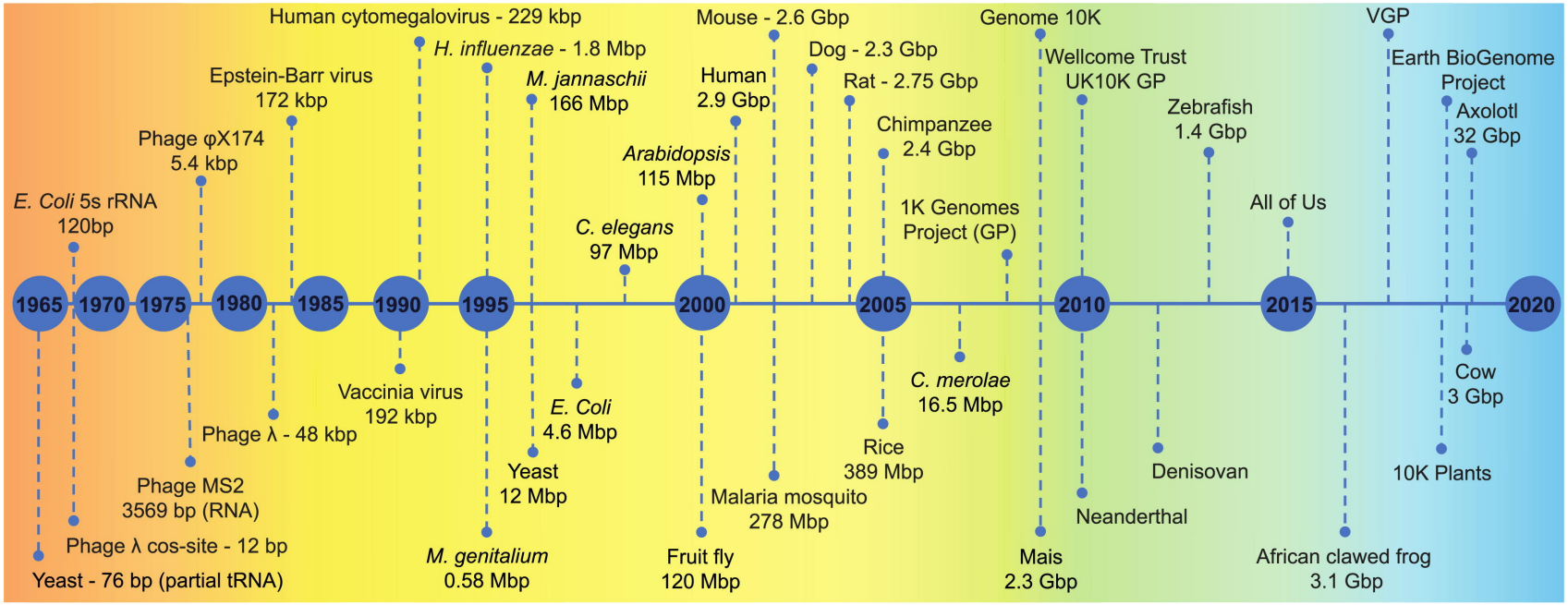
Poznawanie sekwencji genomów - historia

- fag Φ174, 5400 bp - 1977r
- wirus, 170kbp - 1984r
- bakteria, 1,8Mbp - 1995r
- drożdże, 12Mbp - 1996r
- ludzki, 3.3Gbp - 2004r (start w 1990r)
- obecnie 388790 genomów[2]: archea (3856), bakterie (341318), eukariotyczne (33844), wirusy (9772)

Planuje się, że koszt odczytu genomu człowieka spadnie do 100$ (w 2023 jest to 650$) Przewiduje się, że do 2025 roku będzie odczytanych pomiędzy 100 Mln a 2 Mld sekwencji człowieka. Każdy rekord to 100-150GB.

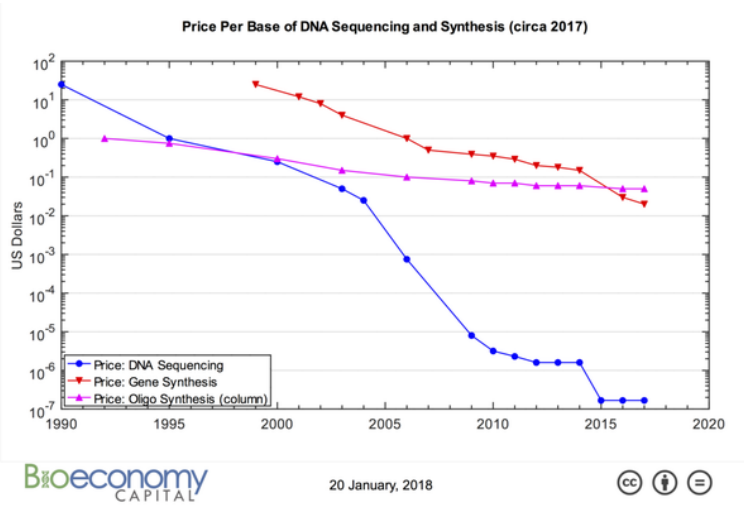
[2] GOLD (Genomes Online Database), `http://genomesonline.org/`

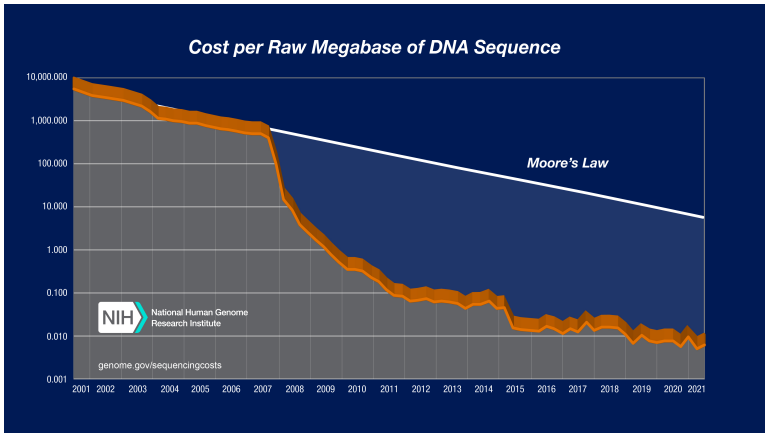
Najważniejsze projekty sekwencjonowania



Prawo Moore'a dla sekwencji genomów

- liczba poznawanych genomów podwaja się co 15 miesięcy
- wielkość sekwencjonowanych genomów podwaja się co 18 miesięcy
- koszty sekwencjonowania (na bp) zmniejszają się dwukrotnie co 18 miesięcy

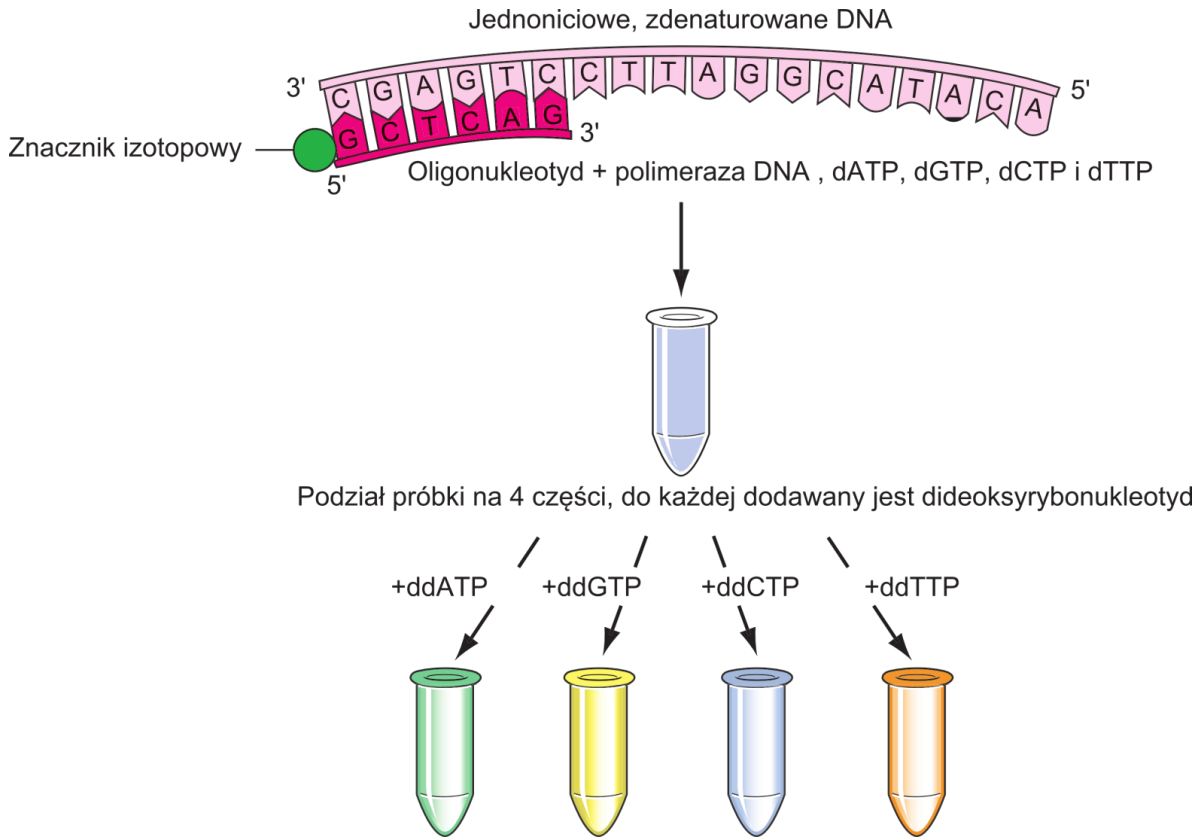




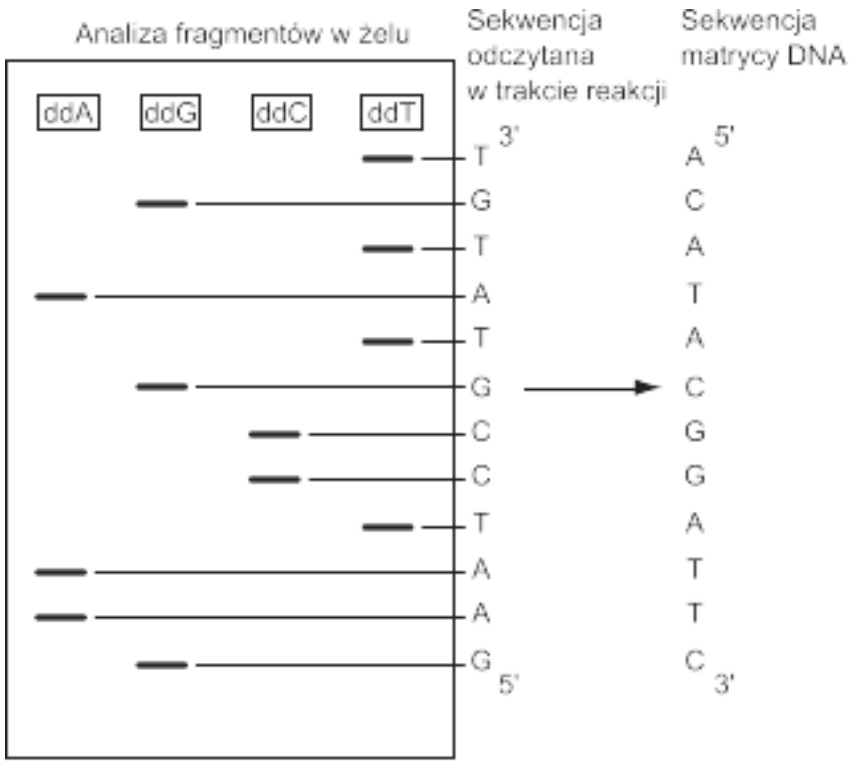
Sekwenatory pierwszej generacji

Metoda Sangera (1) - metoda terminacji łańcucha

Matryca DNA jest denaturowana i dodawany jest znakowany izotopem starter, polimeraza DNA i nukleotydy (dNTP)



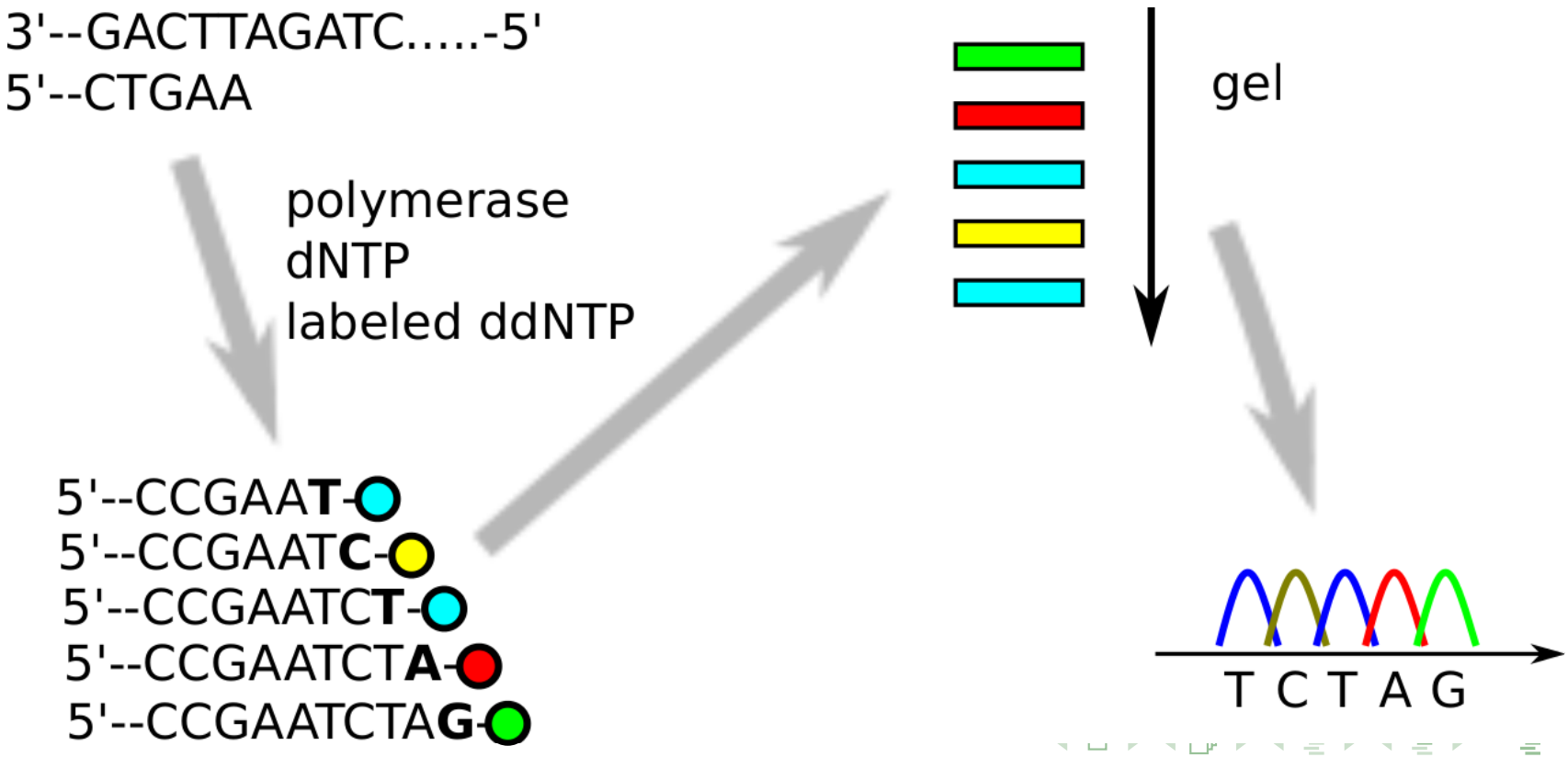
Metoda Sangera (2)



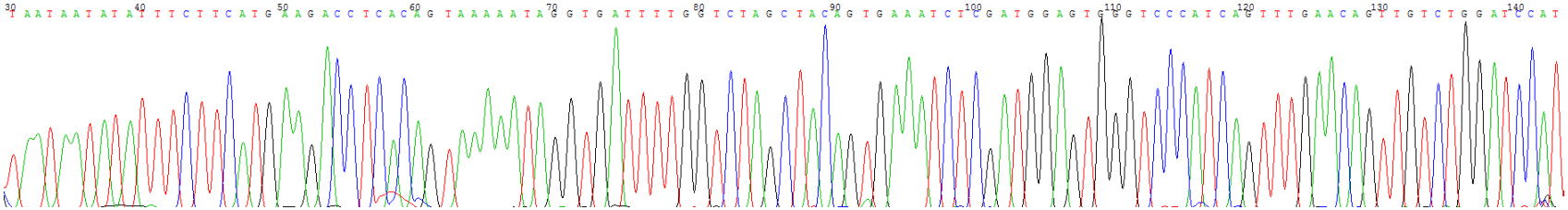
- do rozdziału fragmentów używana jest elektroforeza w żelu poliakrylamidowym
- długość fragmentu odpowiada specyficznym dd-nukleotydom
- sekwencja od 5' do 3' syntezowanej nici jest czytana od dołu żelu

Automatyczne sekwencjonowanie metodą Sangera

- dd-nukleotydy ze znacznikiem fluorescencyjnym
- pojedyncza reakcja i ta sama linia żelu (kapilara)
- sygnał rejestrowany jako chromatogram fluorescencji



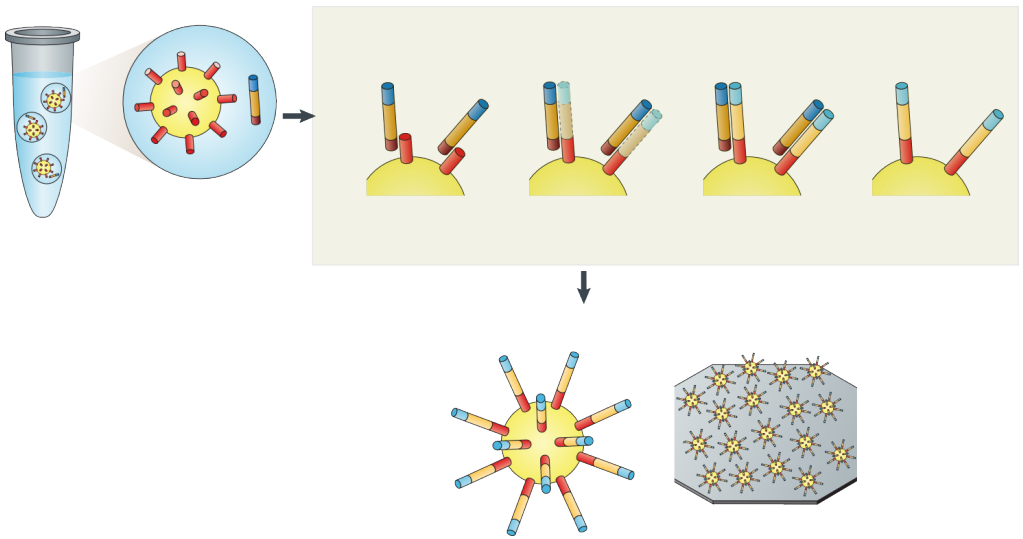
Automatyczna metoda Sangera (2)



- pozwala odczytać 2000 bp
- dokładność 99.999%
- zdominowała sekwencjonowanie na 20 lat
- niska przepustowość i wysokie koszty

Sekwenatory drugiej generacji

Klonalna amplifikacja fragmentów, emulsyjny PCR

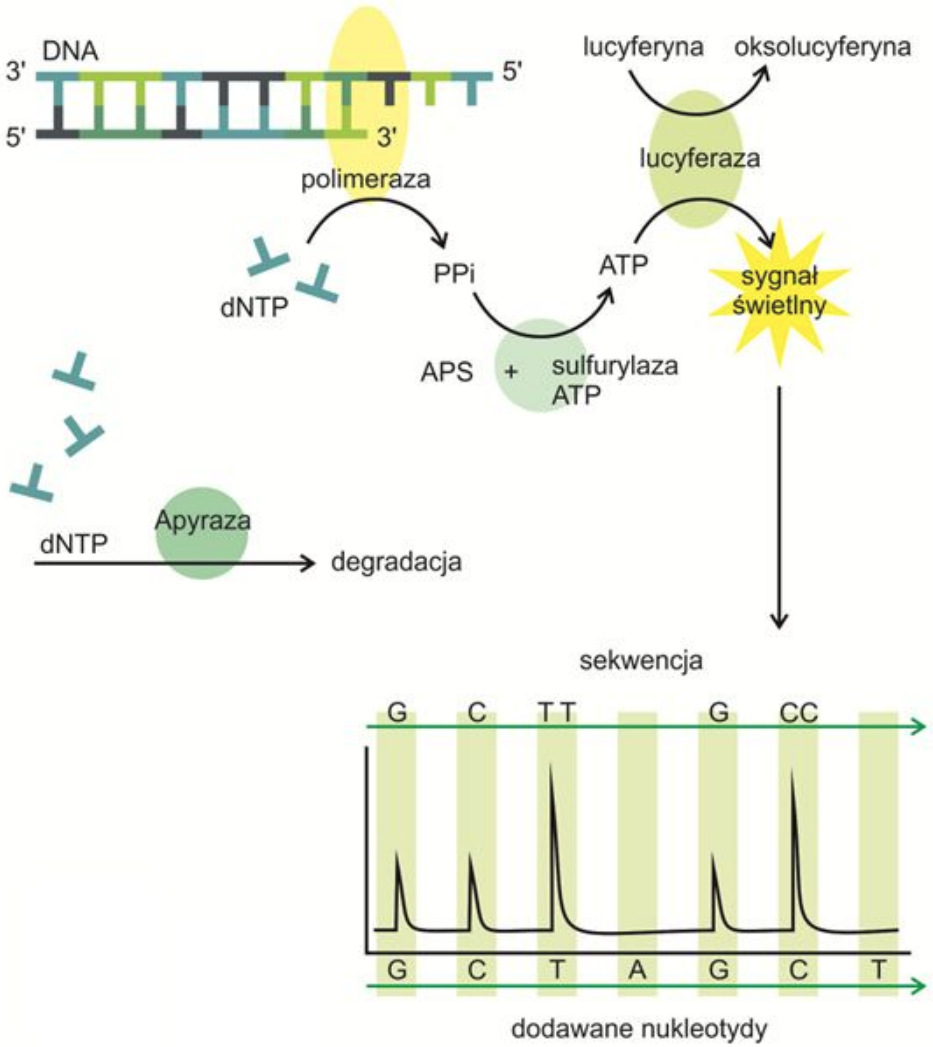


- reaktory: krople w oleju
- kropla zawiera: kulkę, startery, dNTP, matrycę, polimeraza
- jedna cząsteczka DNA na kulkę

- produkty PCR związane do kulek są wtłaczane do studzienek reakcyjnych
- stosowana w urządzeniach: Roche 454, SOLID, Ion Torrent

Piro-sekwencjonowanie, platforma 454

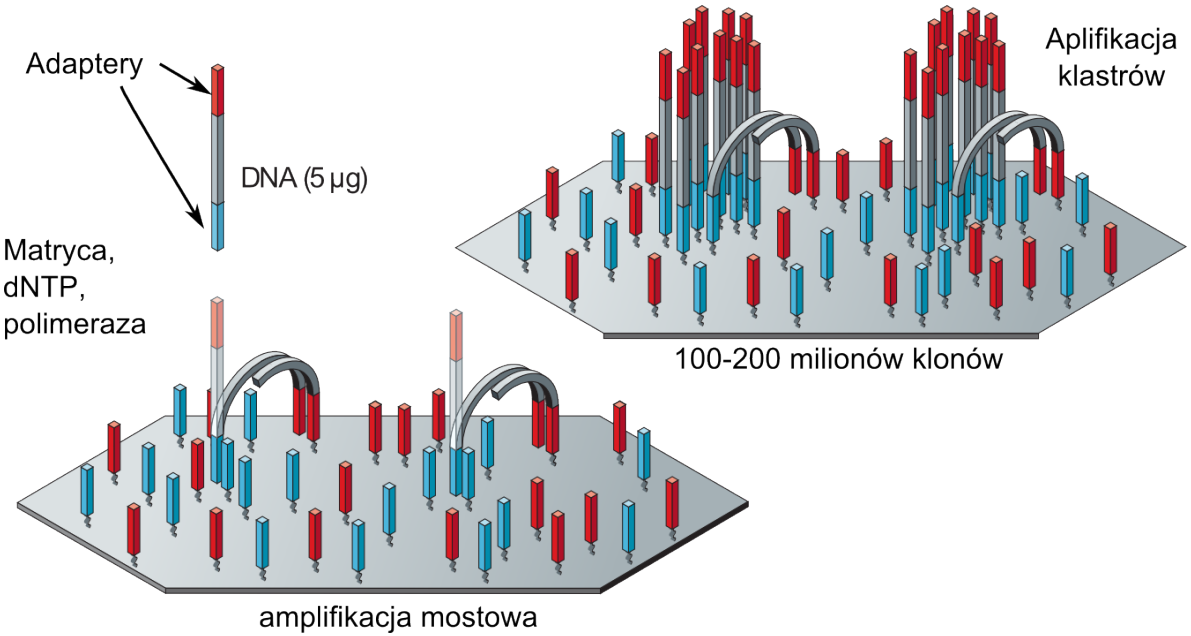
- przeprowadzana jest polimeryzacja z dATP, później dCTP, dGTP, dTTP
- sulfurylaza wychwytuje pirofosforany (PPi) i generuje ATP
- lucyferaza generuje sygnał świetlny
- siła sygnału zależna od ilości zużytego ATP



Platforma Roche 454

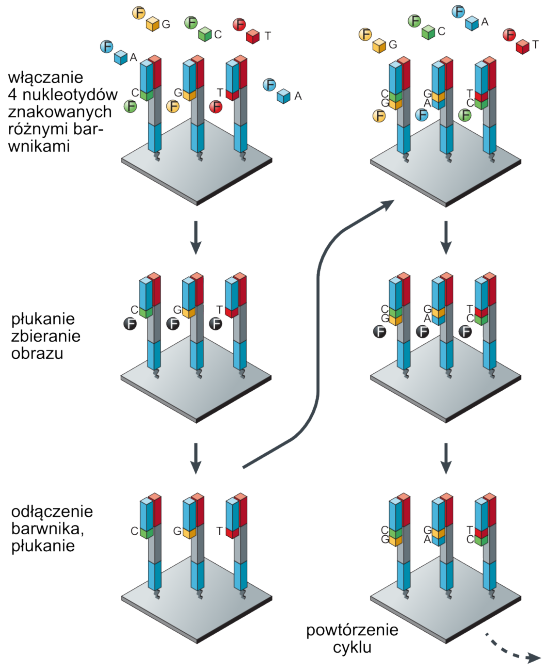
- pierwszy sekwenator drugiej generacji
- początek sprzedaży w 2005 r, koniec produkcji w 2016 r
- sekwencjonowanie $\approx 10x$ tańsze niż metodą Sangera
- $\approx 300x$ bardziej wydajne

Klonalna amplifikacja fragmentów, mostowy PCR, Illumina



- biblioteka losowych fragmentów z adapterami wiązana dwoma końcami do powierzchni płytki
- każdy klaster powiela jeden fragment

Sekwencjonowanie metodą Illumina



- do syntezy używane usuwalne ddNTP wyznakowane fluorescencyjnie
- iteracyjnie skanuje się a następnie usuwa ddNTP

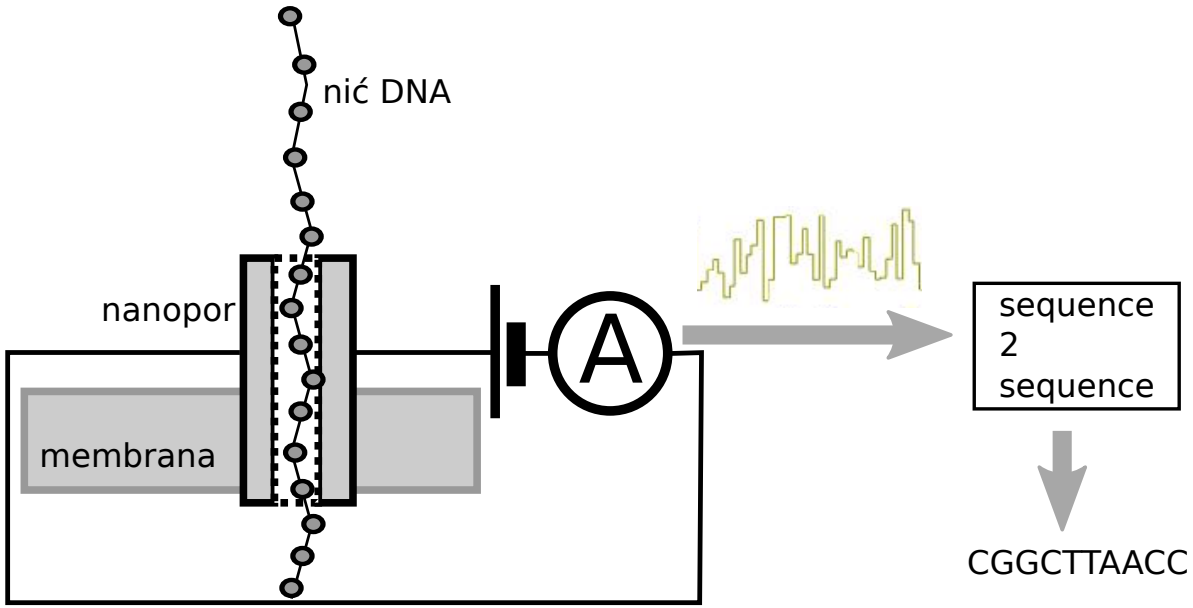
Illumina

- sekwencjonowanie $\approx 100x$ tańsze niż Roche 454
- $\approx 10x$ bardziej wydajne
- obecnie (2023) dominuje na rynku

Illumine HiSeq

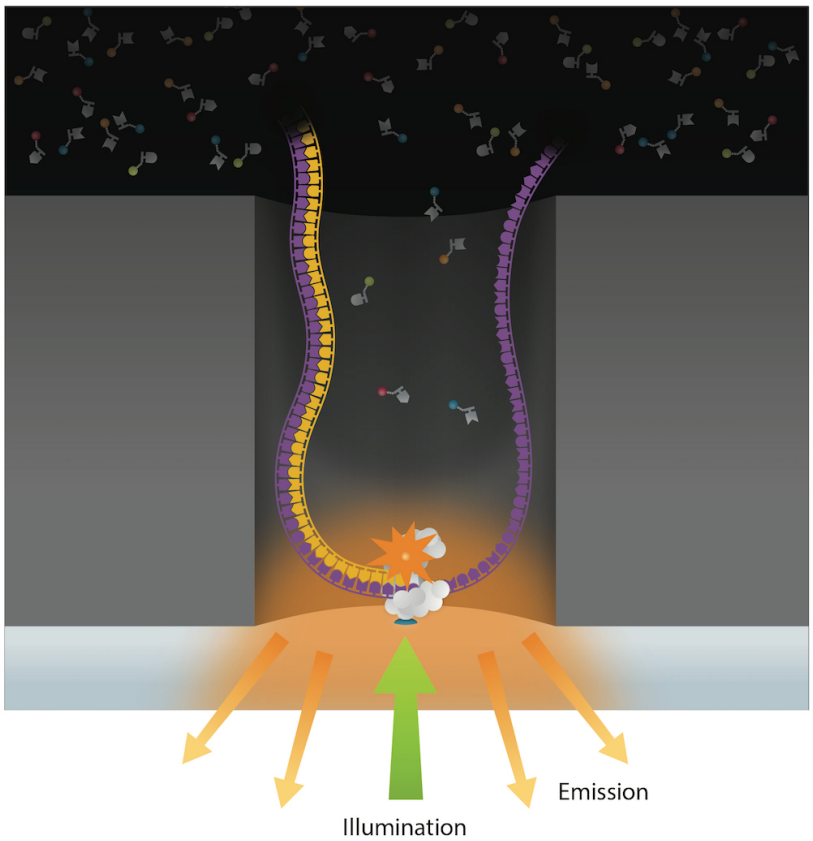
Sekwenatory trzeciej generacji

Pomiar pola elektrycznego w nanoporach



- metoda wykorzystywana przez Oxford Nanopore
- długie odczyty
- 15% błędów

Synteza pojedynczych molekuł w czasie rzeczywistym



- reakcja w nano-komorach
- wyznakowane dNTP gdy są przyłączane, emitują światło
- używany PacBio

Sekwenatory PacBio i Oxford Nanopore

Dziękuję