

Metody bioinformatyki (MBI)

Wykład 14 - Badanie istotności wyników. Przeszukiwanie z odcinaniem - mapy restrykcyjne, wyszukiwanie motywów

Robert Nowak

2024L

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

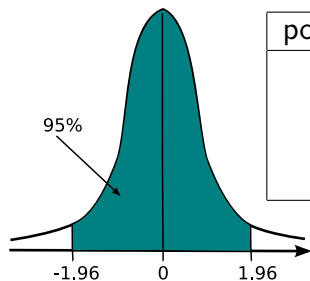
$$P(m) = \binom{n}{m} p^m (1-p)^{n-m}$$

$$m_{sr} = np$$

$$\sigma = \sqrt{np(1-p)}$$

Rozkład ocen dopasowań dla par losowych sekwencji (2)

Gdy $np \geq 3$ oraz $n(1 - p) \geq 3$ to rozkład dwumianowy¹ można przybliżyć 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%	$(m_{sr} - 3\sigma, m_{sr} + 3\sigma)$

¹rozkład 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: $m_2 = 350, n = 1200, p = 0.25, m_{sr} = 300, \sigma = 15$

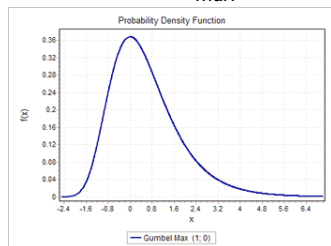
$$z = \frac{m_2 - m_{sr}}{\sigma} = 3.33 \text{ (podobieństwo nie jest przypadkowe)}$$

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

← → ↻ https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=&LINK_LOC=blasttab&LAST_PAGE=blastn

NIH U.S. National Library of Medicine **NCBI** National Center for Biotechnology Information [Sign in to NCBI](#)

BLAST » **blastp suite** [Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

blastn **blastp** **blastx** **tblastn** **tblastx** **Standard Protein BLAST**

Enter Query Sequence BLASTP programs search protein databases using a protein query. [more...](#) [Reset page](#) [Bookmark](#)

Enter accession number(s), gi(s), or FASTA sequence(s)

HALMRRLLPLALLALMGPDPAAAFVNOHLGSHLVEALYLVCGERGFFY
TPKTRREAEDLQSLQPLALEGSLQKRGIVEQCCTICSLYQLENYCN

[Clear](#) [Query subrange](#)

From

To

Or, upload file Nie wybrano pliku

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Organism Optional Exclude

Enter organism name or id--completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search

Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST Search database **Reference proteins (refseq_protein)** using **Blastp (protein-protein BLAST)**

show results in a new window

Przykład - wyniki

Alignments Download GenPept Graphics Distance tree of results Multiple alignment							
	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	insulin preproprotein [Homo sapiens]	192	192	100%	8e-65	89%	NP_000198.1
<input type="checkbox"/>	insulin-2 preproprotein [Mus musculus]	140	140	100%	4e-44	73%	NP_032413.1
<input type="checkbox"/>	insulin-1 preproprotein [Mus musculus]	137	137	100%	5e-43	73%	NP_032412.3
<input type="checkbox"/>	insulin, isoform 2 precursor [Homo sapiens]	112	112	63%	5e-32	100%	NP_001035835.1
<input type="checkbox"/>	insulin-like growth factor II isoform 1 preproprotein [Homo sapiens]	53.9	53.9	92%	2e-09	38%	NP_000603.1
<input type="checkbox"/>	insulin-like growth factor II isoform 2 [Homo sapiens]	53.9	53.9	92%	3e-09	38%	NP_001121070.1
<input type="checkbox"/>	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  MALWMRLLP L L L L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T R R E A E D 60
          MAL + L P L L A L L A L W P P A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K + R R E E D
Sbjct 1  MALLVHFLP L L A L L A L W E P K P T Q A F V K Q H L C G P H L V E A L Y L V C G E R G F F Y T P K S R R E V E D 60
Query 61  L Q ----- G S L O P L A L E G S L Q K R G I V E Q C C T S I C S L Y Q L E N Y C N 98
          0
Sbjct 61  P Q V E Q L E L G G S P G D L Q T L A L E V A R Q K R G I V D Q C C T S I C S L Y Q L E N Y C N 108
  
```

[Download](#) [GenPept](#) [Graphics](#)

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  MALWMRLLP L L L A L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T R R E A E D 60
          MALWMRLLP L L L A L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T R R E A E D 60
Sbjct 1  MALWMRLLP L L L A L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T R R E A E D 60
Query 61  L Q 62
          L Q
Sbjct 61  L Q 62
  
```

Related Information

[Gene](#) - associated gene details

[UniGene](#) - clustered expressed sequence tags

[Map Viewer](#) - aligned genomic context

[Next](#) [Previous](#) [Descriptions](#)

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)

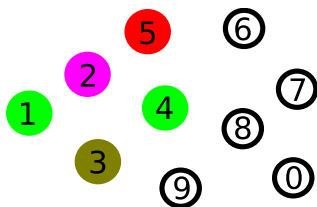
$$\text{czułość (sensitivity)} = \frac{TP}{TP + FN}, \text{ swoistość (specifity)} = \frac{TN}{TN + FP}$$

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

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

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

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 $FPR = 1 - \text{swoistość}$, oś Y to czułość (TPR).

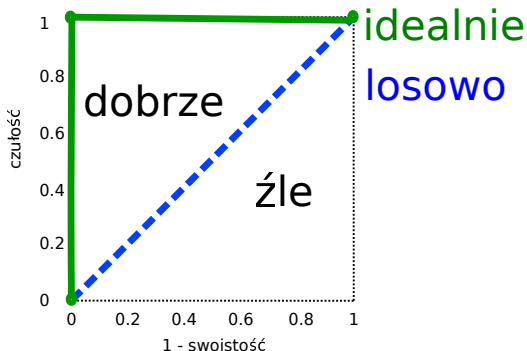
Punkty charakterystyczne:

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

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

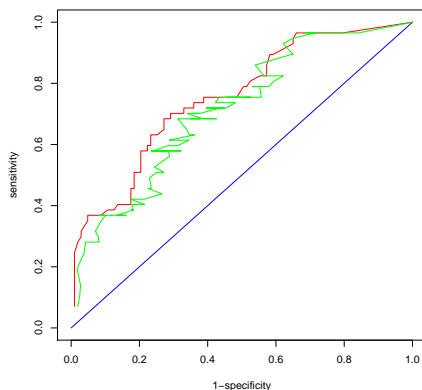
(0,1) test idealny

$y = x$ 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).

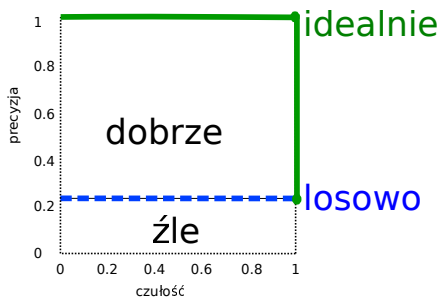
Punkty charakterystyczne:

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

$(1,1)$ test idealny

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

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



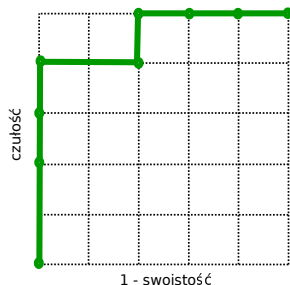
Krzywa ROC i PR - przykład tworzenia



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

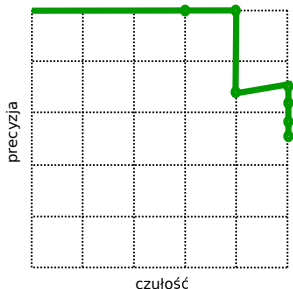
próg	macierz pom.	czuł	sw.	prec.	
0.05	TP = 5 FN = 0	FP = 5 TN = 0	1.0	0.0	0.500
0.15	TP = 5 FN = 0	FP = 4 TN = 1	1.0	0.2	0.556
0.25	TP = 5 FN = 0	FP = 3 TN = 2	1.0	0.4	0.625
0.35	TP = 5 FN = 0	FP = 2 TN = 3	1.0	0.6	0.714
0.45	TP = 4 FN = 1	FP = 2 TN = 3	0.8	0.6	0.667
0.55	TP = 4 FN = 1	FP = 2 TN = 3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

Krzywa ROC i PR - przykład tworzenia



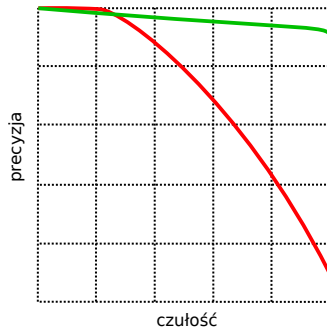
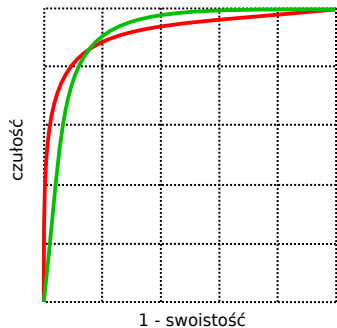
próg	macierz pom.	czuł	sw.	prec.	
0.05	TP = 5 FN = 0	FP = 5 TN = 0	1.0	0.0	0.500
0.15	TP = 5 FN = 0	FP = 4 TN = 1	1.0	0.2	0.556
0.25	TP = 5 FN = 0	FP = 3 TN = 2	1.0	0.4	0.625
0.35	TP = 5 FN = 0	FP = 2 TN = 3	1.0	0.6	0.714
0.45	TP = 4 FN = 1	FP = 2 TN = 3	0.8	0.6	0.667
0.55	TP = 4 FN = 1	FP = 2 TN = 3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

Krzywa ROC i PR - przykład tworzenia



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	-

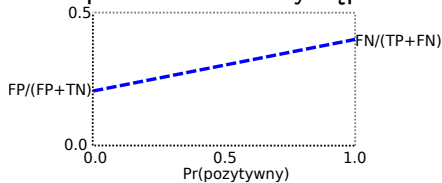
Krzywe ROC i PR gdy nie ma równowagi klas



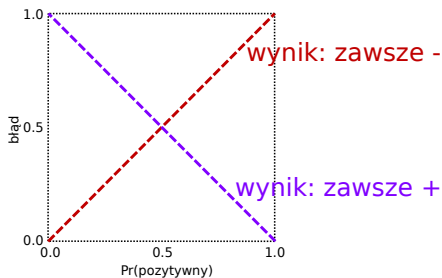
Krzywa kosztu (cost curve)

zakładamy koszt(FP) = koszt(FN)

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



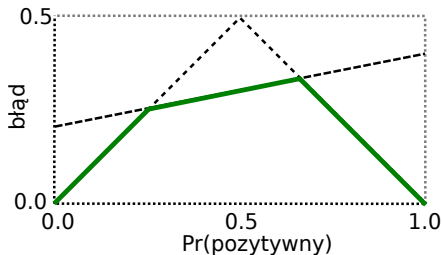
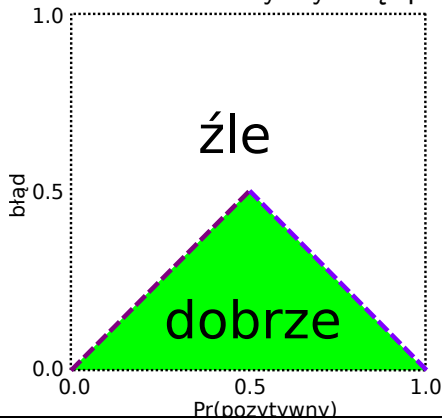
macierz pomyłek to prosta na krzywej kosztu, łączy punkty $(0, \frac{FN}{\Sigma})$ i $(1, \frac{FP}{\Sigma})$



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:

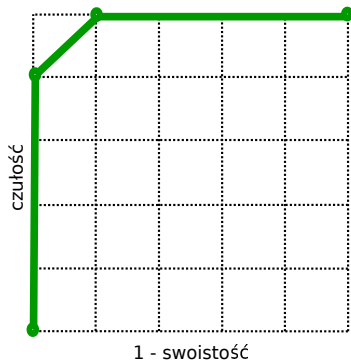
TP = 5	FP = 1
FN = 0	TN = 4

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:

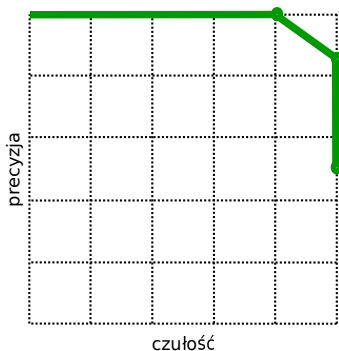
TP = 5	FP = 1
FN = 0	TN = 4

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:

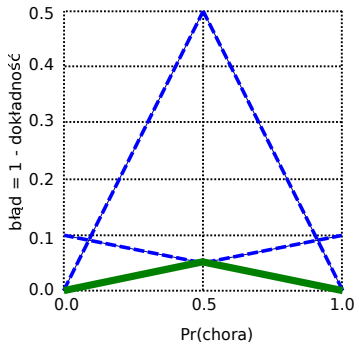
TP = 5	FP = 1
FN = 0	TN = 4

Próg 0.5:

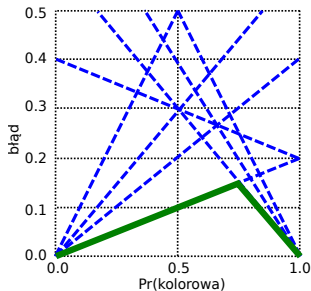
TP = 5	FP = 1
FN = 0	TN = 4

Próg 0.65:

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



Krzywa kosztu - przykład 2



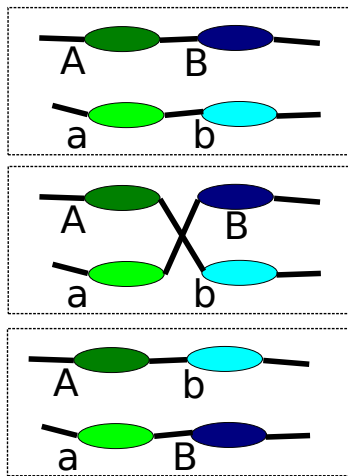
próg	macierz pom.	czuł	sw.	prec.	
0.05	TP = 5 FN = 0	FP = 5 TN = 0	1.0	0.0	0.500
0.15	TP = 5 FN = 0	FP = 4 TN = 1	1.0	0.2	0.556
0.25	TP = 5 FN = 0	FP = 3 TN = 2	1.0	0.4	0.625
0.35	TP = 5 FN = 0	FP = 2 TN = 3	1.0	0.6	0.714
0.45	TP = 4 FN = 1	FP = 2 TN = 3	0.8	0.6	0.667
0.55	TP = 4 FN = 1	FP = 2 TN = 3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

Mapy restrykcyjne

Mapa genetyczna i fizyczna

- ▶ Mapa genetyczna - mapa lokalizacji genów lub markerów na chromosomie powstała poprzez badanie osobników w krzyżówce testowej. Jednostka – centymorgany (cM).
- ▶ Mapa fizyczna - mapa powstała poprzez odczyt sekwencji. Jednostka - nt (nukleotydy) lub bp (pary zasad).

cM - prawdopodobieństwo rozdzielenia w jednym pokoleniu podczas rekombinacji (cross-over) wynosi 1%



Tworzenie mapy genetycznej

zlicza się cechy u osobników potomnych:

- ▶ markery są położone blisko na chromosomie → są dziedziczone wspólnie (małe prawd. ich rozdzielania)
- ▶ markery są oddalone → są dziedziczone niezależnie (duże prawdopodobieństwo ich rozdzielania)

Przykład, krzyżowanie AB/ab i ab/ab^a

^akrzyżówka z podwójną homozygotą recesywną upraszcza obliczenia

- ▶ jeżeli markery są odległe (niesprzężone), to

$$P(AB/ab) = P(Ab/ab) = P(aB/ab) = P(ab/ab) = 0.25$$

- ▶ markery sprzężone zupełnie (w 100%)

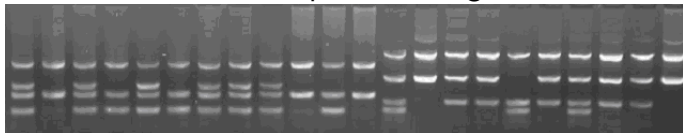
$$P(AB/ab) = P(ab/ab) = 0.5, P(aB/ab) = P(Ab/ab) = 0.0$$

- ▶ inne częstości oznaczają sprzężenie częściowe

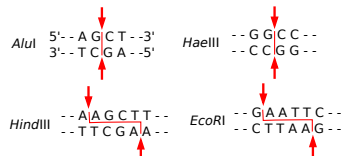
Mapy uproszczone - mapy restrykcyjne

Mapy restrykcyjne można uzyskać bez konieczności odczytywania sekwencji

- ▶ cząsteczki DNA trudno obserwować bezpośrednio
- ▶ elektroforeza - metoda pomiaru długości łańcucha



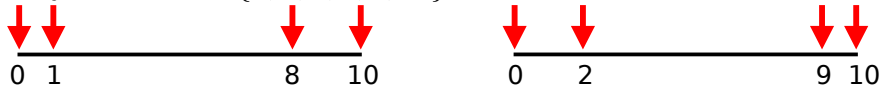
- ▶ enzym restrykcyjny tnie DNA w miejscu, które zawiera wzorec specyficzny dla danego enzymu



Problem częściowego strawienia (partial digest problem)

 $X = \{x_1, x_2, \dots, x_n\}$ zbiór n pozycji $\Delta X = \{x_j - x_i : i < j\}$ multi-zbiór odległości $|\Delta X| = \binom{n}{2}$ Przykład 1: $X = \{0, 5, 7\}$, $\Delta X = \{2, 5, 7\}$ Przykład 2: $X = \{0, 5, 7, 9\}$, $\Delta X = \{2, 2, 4, 5, 7, 9\}$ PDP: dla danego ΔX znaleźć (wszystkie) X

- ▶ typowo 1 rozwiązanie
- ▶ $H(n)$ - liczba rozwiązań dla zbioru n elementowego,
 $H(n) \leq \frac{1}{2}n^{1.233}$

Przykład: $\Delta X = \{1, 2, 7, 8, 9, 10\}$ 

$$X_1 = \{0, 1, 8, 10\}, X_2 = \{0, 2, 9, 10\}$$

PDP strawienia - rekurencja z nawrotami

```

bool partialDigest(set L) { //L - zbiór różnic
    width = deleteMax(L);
    X = { 0, width };
    return place(L,X);
}

bool place(set L, set X) { //L - zbiór różnic, X - zbiór miejsc
    if ( empty L ) return true;
    x = deleteMax(L); //nowy fragment umieszczony od lewej
    if ( delta(x, X ) in L ) { //oblicza różnice odległości
        new_X = X + x;
        new_L = L - delta(x, X);
        if( place(new_L, new_X) ) return true;
    }
    x = width - x; //nowy fragment umieszczony od prawej
    if ( delta(x, X in L ) {
        new_X = X + x; new_L = L - delta(x, X);
        if( place(new_L, new_X) ) return true;
    }
    return false;
}

```


PDP - rekurencja z nawrotami (2)

Przykład $L = \{1, 2, 7, 8, 9, 10\}$

width = 10 place()	X={0,10}, L={1,2,7,8,9} x = 9, delta(x,X) = {1,9} place()	X = {0,9,10}, L = {2,7,8} x = 8, delta(x,X) = {1,2,8} x = 2, delta(x,X) = {2,7,8} place()	X = {0,2,9,10}, L = { } return true
return true		return true	

- ▶ złożoność pesymistyczna:

$$O(2^n n \log n)$$

- ▶ złożoność średnia

$$O(n^2 \log n)$$

PDP - problemy podobne

- ▶ uproszczony problem częściowego strawienia (SPDP), przeprowadzane są dwa doświadczenia: częściowe i pełne trawienie, problem NP-trudny (złożoność wykładnicza)
- ▶ znakowany problem częściowego strawienia (labeled PDP), końce cząsteczki znakowane (np. radioaktywnie), następnie częściowe trawienie, złożoność wielomianowa
- ▶ cięcie dwoma enzymami (Double Digest Problem, DDP), przeprowadzane są trzy doświadczenia dla dwóch różnych enzymów, trawienie pełne jednym enzymem, trawienie pełne drugim, trawienie pełne jednym i drugim, problem NP-trudny

Znaczenie mapowania restrykcyjnego maleje, obecnie używamy sekwencjonowania.

Wyszukiwanie motywów

Wyszukiwanie motywów

Wejście: t sekwencji o długości n

Wyjście: znaleźć pozycje $\mathbf{s} = [s_0, s_1, \dots, s_n]$, które są początkiem motywu o długości l

Przykład:

$$t = 5, n = 18, l = 8$$

$$\mathbf{s} = (5, 0, 2, 9, 6)$$

```
CGGGGCTATCCAGCTGGG
CTATCCAGATCATCATCA
TTCTATCCAGAAAGTCAC
AATTTTTCTCTATCCAGT
AAGGCCCTATCCAGTGAT
```

Typowo: $t = 50..100, n = 500..1000, l = 8..30$

motywy nie są identyczne, ale bardzo podobne

Wyszukiwanie motywów (2)

Ocena dla \mathbf{s} to $\sum_{j=0}^{l-1} \max P_j(\mathbf{s})$, gdzie $P(\mathbf{s})$ to profil dla \mathbf{s} , gdzie $P_j(\mathbf{s})$ jest j kolumną profilu Złożoność: $O(lt)$

Przykład:

CGGGGCTATCCAGCTGGG
CTATCCAGATCATCATCA
TTCTATCCAGAAAGTCAC
AATTTTTCTCTATCCAGT
AAGGCCCTATCCAGTGAT

	G	G	G	G	C	T	A	T
	A	T	C	C	A	G	A	T
	T	A	T	C	C	A	G	A
	T	A	T	C	C	A	G	T
	A	A	G	G	C	C	C	T
A	2	3	0	0	1	2	2	1
C	0	0	1	3	4	1	1	0
G	1	1	2	2	0	1	2	0
T	2	1	2	0	0	1	0	4
	A	A	G	C	C	A	G	T

$$\mathbf{s} = [1, 2, 3, 10, 0]$$

$$\text{ocena} = 2 + 3 + 2 + 3 + 4 + 2 + 2 + 4 = 22$$

Wyszukiwanie motywów (3)

Algorytm pełnego przeglądu przestrzeni:

Dla wszystkich $(n - l + 1)^t$ pozycji początkowych s znaleźć to, o najwyższej ocenie

Złożoność (dla $n \gg l$): $O(ltn^t)$

Obserwacja: Maksymalna ocena = $l * t$ (wszystkie symbole identyczne)

można więc badać tylko pierwszych i pozycji s_0, s_1, \dots, s_i i jeżeli profil jest zły, to nie robić badań dla wszystkich $(n - l + 1)^{(t-i)}$ pozycji s_{i+1}, \dots, s_{t-1}

```

function MOTIFSEARCH( $X, t, n, l$ )
   $s \leftarrow [0, 0, \dots, 0]$ ,  $best \leftarrow 0$ ,  $i \leftarrow 1$ 
  while  $s \in \{[0, 0, \dots, 0], \dots, [n - l + 1, \dots, n - l + 1]\}$  do
    if  $i < t$  then
      optimistic = score( $X, s, i$ ) + (t-i)*l;
      if optimistic < best then
        pomiń poddrzewo
      else
         $i \leftarrow i + 1$ 
      end if
    else
      if score( $X, s$ ) > best then
         $best \leftarrow score(X, s)$ ,  $motif \leftarrow s$ 
      end if
       $s \leftarrow next(s)$ ,  $i \leftarrow 1$ 
    end if
  end while
  return motif
end function

```

Wyszukiwanie motywów (4) - mediana napisów

Mediana napisów, znaleźć napis v , $|v| = l$, który daje minimalną odległość dla t napisów $[X_0, X_1, \dots, X_t]$, czyli

$$d_H(v, X) = \sum_{i=0}^{t-1} d_H(v, X_i)$$

- ▶ napisów jest 4^l (dla alfabetu DNA)
- ▶ obliczenie $d_H(v, X)$ jest rzędu $O(nt)$ (dla każdego z t napisów znajdujemy pozycję, która jest najbardziej podobna do v (np. algorytm KMP))
- ▶ złożoność znajdowania mediany napisów $O(nt4^l) \ll O(lnt^t)$

Mediana napisów jest motywem ^a

^adowód w N. Jones, P. Pevzner, An Introduction to Bioinformatics Algorithms, 2004


```

function MEDIANSEARCH( $X, t, n, l$ )
   $v \leftarrow AA...A, |v| = l, best \leftarrow \infty, i \leftarrow 1$ 
  while  $v \in \{AA...A, \dots, TT...T\}$  do
    if  $i < l$  then
       $prefix \leftarrow v[0..i], optimistic \leftarrow distance(X, prefix)$ 
      if  $optimistic > best$  then
        pomiń poddrzewo
      else
         $i \leftarrow i + 1$ 
      end if
    else
      if  $distance(X, v) < best$  then
         $best \leftarrow distance(X, v), median \leftarrow v$ 
      end if
       $v \leftarrow next(v), i \leftarrow 1$ 
    end if
  end while
  return median
end function

```

Dziękuję