Wprowadzenie do Informatyki Biomedycznej

Wykład 3:

Wyszukiwanie w bazach sekwencji
Przewidywanie genów

Marek Krętowski
pokój 206
e-mail: m.kretowski@pb.edu.pl
http://aragorn.pb.bialystok.pl/~mkret

Wersja 1.05

Wyszukiwanie podobnych sekwencji w bazach danych

- Ogromne projekty, w których całe genomy organizmów są sekwencjonowane, pozwoliły upowszechnić informację genomiczną
  - przykłady: bacteria *Escherichia coli*, drożdże *Saccharomyces cerevisiae*, itd. czy najbardziej spektakularny projekt dotyczący *Homo sapiens*
  - intensywna analiza biologiczna tych modelowych organizmów pozwala odkrywać funkcje genów i zakodowanych białek
- Tworzenie baz danych sekwencji i ich przeszukiwanie jest jednym z najistotniejszych zadań bioinformatyki:
  - wykrywanie podobieństwa nowych analizowanych sekwencji do zgromadzonych i rozpoznanych sekwencji umożliwia np. przewidywanie funkcji genów u zbliżonych gatunków
  - alternatywnie, sekwencja o znanej funkcji może być wykorzystana do przeszukiwania genomu konkretnego organizmu w celu wskazania genów mogących pełnić analogiczne funkcje

Wyszukiwanie w sekwencji nukleotydów czy aminokwasów?

- Sekwencje DNA składają się z 4 nukleotydów, podczas gdy sekwencje białek wykorzystują 20 aminokwasów
  - w związku z pięciokrotnie większą różnorodnością znaków w sekwencji, wykrycie podobieństwa pomiędzy sekwencjami białek niż DNA jest znacznie łatwiejsze
- Zasadą przy przeszukiwaniu baz jest poprzedzenie wyszukiwania przetłumaczeniem sekwencji DNA na sekwencję aminokwasów
  - udowodniono, że prowadzi to do daleko bardziej znaczących dopasowań
  - w większość narzędzi wyszukiwania wbudowano narzędzia dokonujące niezbędnych tłumaczeń
  - odstępstwo od powyższej zasady, ma jedynie uzasadnienie, gdy porównujemy sekwencje nukleotydów tego samego organizmu w celu wykrycia pozostałych wpisów tej samej sekwencji

Czułość i selektywność

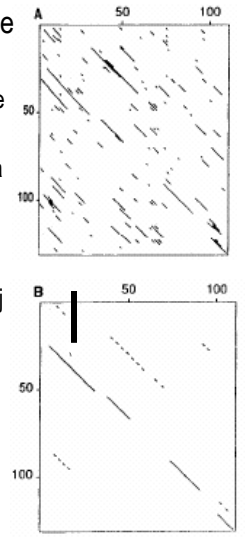
- Porównując metody przeszukiwania baz białek należy brać pod uwagę zarówno **czułość** (ang. sensitivity) jak i **selektywność** (ang. selectivity)
- Czułość - odnosi się do zdolności metody do wyszukiwania większości członków rodziny białek reprezentowanej przez sekwencje z zapytania
- Selektywność - odnosi się natomiast do zdolności metody do nie odszukiwania przedstawicieli innych rodzin białek - błędne rozpoznanie, klasyfikacja (ang. *false positive*)
- W idealnym przypadku zarówno czułość jak i selektywność powinny być jak najwyższe
- Wygodnym sposobem opisywania obu cech jest podawanie poziomu pokrycia rodziny białek przy danym poziomie błędnej klasyfikacji

Wyszukiwanie przy wykorzystaniu FASTA

- Zamiast porównywania pojedynczych znaków w sekwencji, FASTA wyszukuje dopasowane wzorce lub słowa
  - słowa lub krotki (ang. k-tuples) obejmują dopasowanie *k* kolejnych znaków
- Następnie program próbuje stworzyć lokalne ustawienie bazując na dopasowaniu słów
- FASTA3 - aktualna wersja programu jest rezultatem całej serii usprawnień (zarówno dopasowanie sekwencji jak i wyliczanie statystycznego ich znaczenia), które pozwoliły m.in. poprawić znacząco wykrywanie słabo spokrewnionych sekwencji
  - przyjmuje się, że dla fragmentów sekwencji FASTA jest równie skuteczny jak algorytm Smith-Waterman (p.dynam.)
  - w przypadku przeszukiwania sekwencji DNA, FASTA może być nawet szybszy niż BLASTN w wyszukiwaniu dopasowań, gdyż dopuszcza krótsze słowa

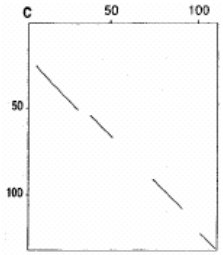
Działanie FASTA (1)

- (**A**)10 najlepiej dopasowanych obszarów w każdej parze sekwencji jest lokalizowanych poprzez szybki przegląd
  - najpierw wszystkie dopasowania o długości *k* są odnajdywane (sekwencje DNA: *k* jest zwykle 4-6, białka: 1-2)
  - następnie dopasowane fragmenty w pewnej odległości (białka 32 dla *k*=1, 16 dla *k*=2) są łączone w dłuższe dopasowane fragmenty bez przerw i obszary z najwyższym dopasowaniem są wyznaczane
  - obliczenia są zbliżone do metodę kropkowej, ale jest ich mniej
  - odcinki ukośne reprezentują odnalezione wspólne odcinki
- (**B**) najbardziej dopasowane obszary wyznaczone w poprzednim kroku są oceniane przy użyciu macierzy zastępstw (aminokwasów - PAM lub BLOSUM)
  - najwyżej ocenione, nazywane najlepszymi początkowymi obszarami (INIT1) są wykorzystywane dalej



Działanie FASTA (2)

- (**C**) dłuższe obszary (INITN) są tworzone poprzez łączenie początkowych obszarów z wynikami wyższymi niż założony próg
  - wynik odpowiadający INITN jest sumą wyników tworzących go regionów minus stałe kary za każdą wprowadzoną przerwę
  - najnowsze wersje FASTA zawierają krok optymalizacyjny: gdy wynik INITN osiąga pewien próg wówczas pełne lokalne dopasowanie jest wykonywane przy wykorzystaniu programowania dynamicznego (Smith-Waterman)
  - pozwala to na poprawienie wyniku (zwiększa czułość, ale zmniejsza selektywność)
- (**D**) Optymalne lokalne dopasowanie pomiędzy wejściową sekwencją a najlepszą znalezioną sekwencją jest wykonywane (alg. Smith-Waterman)



Wykorzystanie haszowania w FASTA

- Tablica zawierająca położenie każdego wystąpienia słowa o długości *k* jest tworzona dla każdej sekwencji; następnie względne położenie każdego słowa jest wyliczane poprzez odjęcie pozycji
- Słowa, które mają takie samo przesunięcie są w fazie i ujawniają obszary dopasowania
- Przy wykorzystaniu haszowania liczba porównań rośnie liniowo wraz z średnią długością sekwencji
- W przypadku białek oraz *k*=2 liczba rozpatrywanych krotek wynosi 20x20=400

```
position    1  2  3  4  5  6  7  8  9  10  11
sequence 1  n  c  s  p  t  a  ·  ·  ·  ·   ·

position    1  2  3  4  5  6  7  8  9  10  11
sequence 2              a  c  s  p  r  k
```

amino acid	position in protein A	position in protein B	offset pos A - pos B
a	6	6	0
c	2	7	-5
k	-	11	
n	1	-	
p	4	9	-5
r	-	10	
s	3	8	-5
t	5	-	

```
protein 1 n c s p t a
            | | |
protein 2 a c s p r k
```

BLAST - Basic Local Alignment Search Tool

- W założeniu miał być szybszy niż FASTA przy tej samej czułości
  - opracowany przez S. Altschula w 1990 r.
  - wykorzystywany bardzo intensywnie m.in. w NCBI, National Library of Medicine
- Podobnie jak FASTA wyszukuje najpierw wspólne słowa:
  - przy czym w przeciwieństwie do FASTA, który poszukuje wszystkich możliwych słów zadanej długości, BLAST ogranicza się tylko do najbardziej znaczących słów
  - w przypadku białek znaczenie jest wyznaczane przy wykorzystaniu np. BLOSUM62
- Długość słowa jest ustalona na 3 (białka) lub 11 (DNA)
  - długość ta jest minimum niezbędnym do uzyskania oceny dopasowania wystarczającej wysokiego aby uznane zostało za znaczące, ale nie zbyt długie aby pominąć krótkie, ale znaczące wzorce
- Wersja BLAST2 pozwala filtrować odcinki o niskiej złożoności zarówno w sekwencjach DNA jak i białkach
  - odcinki takie mogą prowadzić do zawyżonych ocen dopasowania

Szkic algorytmu BLAST

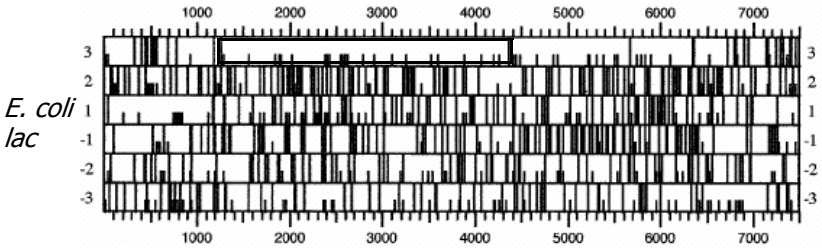
- Eliminacja/maskowanie regionów o niskiej złożoności (LCR- Low-Complexity Region)
  - obszary w sekwencji z wielokrotnie powtórzoną niewielką liczbą elem.
  - ok. 15 % zawartości baz
  - powodują tworzenie nieprawdziwych dopasowań i/lub sztucznego zawyżania ocen
- Stworzenie listy słów (ustalonej długości) na podstawie sekwencji poszukiwanej
- Wybranie tylko wysoko ocenianych podobnych słów
  - ocena przy użyciu macierzy zastępstw
- Dla każdej wysoko ocenionej pary HSP (High Scoring Segment Pair) następuje próba rozszerzenia przyrównania w obu kierunkach, z jednoczesnym wyliczaniem oceny
  - do momentu gdy ocena nie spadnie poniżej ustalonego progu
- Podawane są wszystkie przyrównania o odpowiednio wysokiej ocenie i weryfikowane jest ich statystyczne znaczenie
- Można próbować połączyć dopasowanie dwóch (lub więcej) HSP
- Nowsza wersja BLAST dopuszcza też przerwy

Przewidywanie genów (ang. gene prediction)

- Obejmuje przewidywanie położenia genów kodujących białka oraz identyfikację sekwencji (takich jak promotory), które regulują aktywność tych genów
- ORF (Open Reading Frame) - odcinek DNA zawierający zbiór przyległych kodonów, z których każdy określa jeden aminokwas
- W każdej sekwencji jest 6 sposobów odczytu informacji (ang. *reading frames*) - ram:
  - rozpoczynające się od pozycji 1, 2 lub 3 w kierunku od 5' do 3' oraz w przeciwnym kierunku w komplementarnej sekwencji
- W genomie prokariotów, DNA kodujące białka jest przepisywane na mRNA i zwykle mRNA jest tłumaczone na białko bez istotnych zmian
  - w takiej sytuacji najdłuższy ORF ciągnący się od pierwszego dostępnego kodonu MET do następnego kodonu oznaczającego koniec (TER) może być z dużym prawdopodobieństwem rozpoznany jako obszar kodujący białko

Predykcja genów prokariotów

- rama odczytu, która nie koduje białka posiada krótkie ORF w związku z występowaniem dużej liczby kodonów kończących
- występowanie wielu genów oraz możliwość pokrywania się genów (dwa różne białka są zakodowane na różnych ramach tego samego mRNA, na tym samym albo komplementarnym łańcuchu DNA)



AUG (początek) - połowa pionowej linii, TER (koniec) - |

gen *lacZ* - ORF od pozycji 1284 do 4355 w ramie 3

Predykcja genów eukariotów

- Generalnie przewidywanie genów jest znacznie trudniejszym zadaniem
- Po transkrypcji obszarów kodujących białka (rozpoczynających się od specyficznych sekwencji promotora) następuje usunięcie niekodujących sekwencji (intronów) z pre-mRNA przy wykorzystaniu mechanizmu sklejania, pozostawiając jedynie kodujące eksony
  - zakładając, że introny zostały usunięte oraz po pewnych dodatkowych modyfikacjach dojrzałego RNA może nastąpić translacja zwykle od pierwszego kodonu startu do pierwszego stopu
- Trzy elementy dotyczące fazy po transkrypcji mają ponadto wpływ na translację i w konsekwencji jakość predykcji genów
  - kod genetyczny konkretnego organizmu może różnić się od uniwersalnego
  - poszczególna tkanka może wykorzystywać różne mechanizmy sklejania, tworząc podobne, ale różne mRNA kodujące powiązane ale różniące się białka
  - mRNA może być edytowane, zmieniając sekwencję

Uniwersalny kod genetyczny

UUU-Phe	F	UCU-Ser	S	UAU-Tyr	Y	UGU-Cys	C
UUC-Phe	F	UCC-Ser	S	UAC-Tyr	Y	UGC-Cys	C
UUA-Leu	L	UCA-Ser	S	UAA-	TER	UGA-	TER
UUG-Leu	L	UCG-Ser	S	UAG-	TER	UGG--Trp	W
CUU-Leu	L	CCU-Pro	P	CAU-His	H	CGU-Arg	R
CUC-Leu	L	CCC-Pro	P	CAC-His	H	CGC-Arg	R
CUA-Leu	L	CCA-Pro	P	CAA-Gln	Q	CGA-Arg	R
CUG-Leu	L	CCG-Pro	P	CAG-Gln	Q	CGG-Arg	R
AUU-Ile	I	ACU-Thr	T	AAU-Asn	N	AGU-Ser	S
AUC-Ile	I	ACC-Thr	T	AAC-Asn	N	AGC-Ser	S
AUA-Ile	I	ACA-Thr	T	AAA-Lys	K	AGA-Arg	R
AUG-MET	M	ACG-Thr	T	AAG-Lys	K	AGG-Arg	R
GUU-Val	V	GCU-Ala	A	GAU-Asp	D	GGU-Gly	G
GUC-Val	V	GCC-Ala	A	GAC-Asp	D	GGC-Gly	G
GUA-Val	V	GCA-Ala	A	GAA-Glu	E	GGA-Gly	G
GUG-Val	V	GCG-Ala	A	GAG-Glu	E	GGG-Gly	G

Sprawdzanie poprawności predykcji

- Sekwencje DNA, które kodują białka nie są losowymi łańcuchami dostępnych kodonów, ale raczej uporządkowanymi listami określonych kodonów, które odzwierciedlają ewolucyjne pochodzenie genu oraz ograniczenia związane z ekspresją genu
  - ten brak losowości może być wykorzystany podczas predykcji
  - każdy gatunek ma charakterystyczne wzorce wykorzystania równoznacznych kodonów, dodatkowo te wzorce mogą się różnić pomiędzy silnie i słabo wyrażonymi genami
- Zaproponowano 3 testy pozwalające utwierdzić się w przekonaniu, że rozpoznany ORF koduje rzeczywiście białko
  - pierwszy opiera się na obserwacji, że w ORF każda trzecia baza jest często taka sama (dużo częściej niż losowo); własność ta jest niezależna od gatunku
  - analiza porównawcza kodonów wykorzystywanych w innych genach tego organizmu
  - porównywanie wynikowej sekwencji aminokwasów z rozpoznanymi wcześniej

Predykcja genów prokariotów

- Predykcja genów jest łatwiejsza, ponieważ nie ma intronów oraz występuje kilka wysoce zakonserwowanych wzorców w obszarze promotora i wokół miejsca startu transkrypcji
- Na przykładzie obszary -10 i -35 (żółte) oznaczają obszary interakcji z polimerazą RNA

DNA PATTERNS IN THE *E. coli lexA* GENE

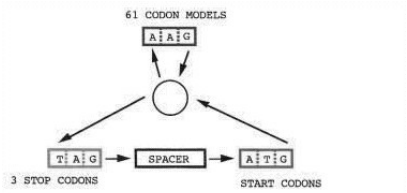
GENE SEQUENCE — PATTERN

1 GAATTCGATAAATCTCTGGTTTATTGTGCAGTTTATGGTT — CTGNNNNNNNNNNCAG
TT — TTGACA
41 CCAAAATCGCCTTTTGCTGTATATACTCACAGCATAACTG — CTGNNNNNNNNNNCAG
CCAA -35 -10 TATACT > — TATAAT, > mRNA start
81 TATATACACCCAGGGGGCGGAATGAAAGCGTTAACGGCCA — CTGNNNNNNNNNNCAG
+10 GGGGG Ribosomal binding site — GGAGG
121 GGCAACAAGAGGTGTTTGATCTCATCCGTGATCACATCAG
161 CCAGACAGGTATGCCGCCGACGCGTGCGGAAATCGCGCAG — ATG
201 CGTTTGGGGTTCCGTTCCCAAACGCGGCTGAAGAACATC
241 TGAAGGCGCTGGCACGCAAAGGCGTTATTGAAATTGTTTC
281 CGGCGCATCACGCGGATTCGTCTGTTGCAGGAAGAGGAA
321 GAAGGGTTGCCGCTGGTAGGTCGTGTGGCTGCCGGTGAAC
361 CACTTCTGGCGCAACAGCATATTGAAGGTCATTATCAGGT — OPEN READING FRAME
401 CGATCCTTCCTTATTCAAGCCGAATGCTGATTTCCTGCTG
441 CGCGTCAGCGGGATGTCGATGAAAGATATCGGCATTATGG
481 ATGGTGACTTGCTGGCAGTGCATAAAACTCAGGATGTACG
521 TAACGGTCAGGTCGTTGTCGCACGTATTGATGACGAAGTT
561 ACCGTTAAGCGCCTGAAAAAACAGGGCAATAAAGTCGAAC
601 TGTTGCCAGAAAATAGCGAGTTTAAACCAATTGTCGTTGA
641 CCTTCGTCAGCAGAGCTTCACCATTGAAGGGCTGGCGGTT
681 GGGGTTATTCGCAACGGCGACTGGCTGTAACATATCTCTG — TAA
721 AGACCGCGATGCCGCCTGGCGTCGCGGTTTGTTTTTCATC
761 TCTCTTCATCAGGCTTGTCTGCATGGCATTCCTCACTTCA
801 TCTGATAAAGCACTCTGGCATCTCGCCTTACCCATGATTT
841 TCTCCAATATCACCGTTCCGTTGCTGGGACTGGTCGATAC
881 GGCGGTAATTGGTCATCTTGATAGCCCGGTTTATTTGGGC
921 GGCGTGGCGGTTGGCGCAACGGCGGACCAGCT

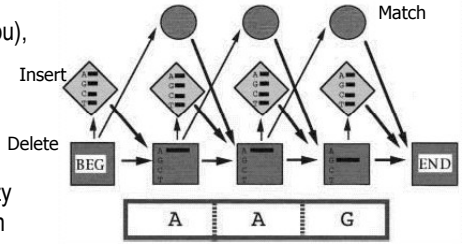
Shown are matches to approximate consensus binding sites for LexA repressor (CTGNNNNNNNNNNCAG), the -10 amd -35 promoter regions relative to the start of the mRNA (TTGACA and TATAAT), the ribosomal binding site on the mRNA (GGAGG), and the open reading frame (ATG...TAA). Only the second two of the predicted LexA binding sites actually bind the repressor.

Wykorzystanie ukrytych modeli Markowa (ang. hidden Markov model)

- Wysoce zakonserwowane cechy genów narzucają możliwość ich identyfikacji
- Na rysunku przedstawiony jest model genomu bakterii, w którym geny są gęsto upakowane z relatywnie krótkimi obszarami pomiędzy i brakiem intronów
- Model czyta sekwencję o nieznanej kompozycji genów i znajduje geny (serię kodonów pomiędzy kodonami startu i stopu), które najbardziej są zbliżone do znanych sekwencji genów wykorzystywanych do uczenia modelu
- Ponieważ wykorzystanie kodonów i obejmowanie sekwencji różni się pomiędzy genomami, model wytrenowany na danym organizmie nie musi pracować dla innego



Wprowadzenie modelu kodonu

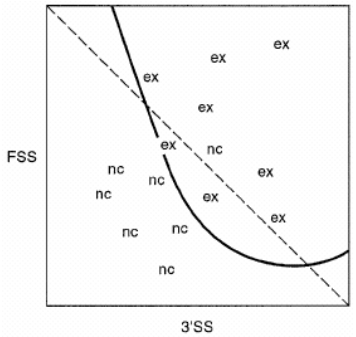


Model pojedynczego kodonu

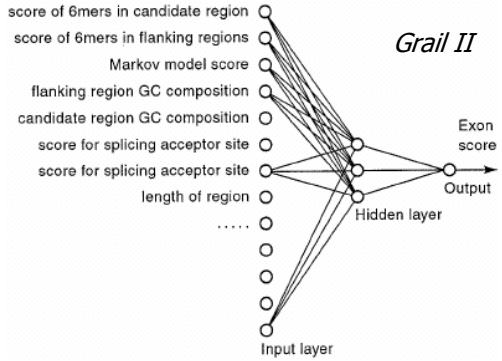
Predykcja genów eukariotów

- Najprostsza metoda odkrywania genów kodujących białka polega na translacji sekwencji na 6 sposobów i próbie przeszukania bazy sekwencji przy użyciu BLASTX lub FASTX
- Inną możliwością jest wykorzystanie metod klasyfikacji do rozpoznawania eksonów (odróżnienie eksonów od sekwencji niekodujących – metody *ab initio*)
  - w pierwszym kroku musi zostać stworzony (nauczony) klasyfikator na podstawie zbioru uczącego zawierającego charakterystyki znanych eksonów (wykorzystywane są głównie cechy związane z sygnałami i składem genu)
  - do sygnałów zaliczamy miejsca start i stop genu oraz przypuszczalne miejsca splicingowe, charakterystyczne sekwencje konsensusowe
  - skład genu odnosi się do statystyk kodowania (np. nielosowy rozkład nukleotydów i aminokwasów czy częstość występowania heksamerów)

Wykorzystanie analizy dyskryminacyjnej i sztucznych sieci neuronowych



Funkcja dyskryminacyjna:
- - - - - liniowa
——— kwadratowa



GRAIL (Gene Recognition and Assembly Internet Link): sieć została wytrenowana przy użyciu wielu cech (m. in. miejsca splicingowe, kodony start i stop, sekwencje promotorowe, ...); skanuje sekwencję wejściową oknem o zadanej długości i ocenia p-wo, że fragment jest kodujący => efektem jest zestaw potencjalnych eksonów