Bioinformatyka

wykład 5: poszukiwanie motywów

prof. dr hab. inż. **Marta Kasprzak**
Instytut Informatyki, Politechnika Poznańska

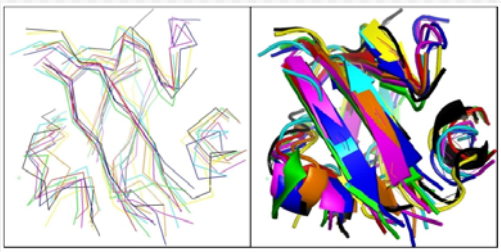
Poszukiwanie motywów

- Poszukiwanie charakterystycznych ciągów w sekwencjach nukleotydowych lub aminokwasowych jest częstym problemem biologicznym. Dotyczy np. regionów promotorowych w sekwencjach nukleotydowych lub fragmentów odpowiedzialnych za strukturę białka w sekwencjach aminokwasowych
- Poszukiwanie może dotyczyć motywów o stałej strukturze, identycznej w badanych sekwencjach (np. najdłuższe powtórzenia), ale najczęściej wystąpienia tego samego motywu w różnych organizmach różnią się w szczegółach (np. wskutek ewolucji), za to posiadają charakterystyczny wspólny „szkielet" odpowiedzialny za ich funkcję

Poszukiwanie motywów

- Przedmiotem poszukiwań może być pojedyncze wystąpienie motywu w sekwencji lub też cała seria różnych motywów. W tym drugim przypadku kolejne wystąpienia motywów mogą następować (bądź nie) w zadanym porządku
- Dodatkowo można nakładać ograniczenia na odległości pomiędzy wystąpieniami motywów. Przykładowo, region promotorowy ma pojawić się w przedziale zdefiniowanym w pewnej odległości przed kodonem startu, kluczowe fragmenty sekwencji białkowej mają zawierać się w oknach o zadanym położeniu wewnątrz sekwencji
- Można poszukiwać motywy znane lub nieznane

Motywy w białkach



Motywy w białkach a struktura przestrzenna.

Uwaga: dopasowanie sekwencyjne utworzone zostało tutaj na podstawie dopasowania strukturalnego.

descriptor name	segment 1	segment 2	segment 3	segment 4	segment 5	segment 6
d1p1da2_A_206_LEU	FHVKLPK	LGITI	DPLVISD	SVAHRTGTLEL	DKLLAIDN	QILQQCEDLVKLKIRK
d1q3oa_A_679_VAL	KTVLLQK	FGFVL	..QYLES	GVAWR.AGLRM	DFLIEVNG	NMIRQ..NTLMVKVVM
d1y7na1_A_84_MET	TTVLIRR	LGFSV	..GIICS	GIAER.GGVRV	HRIIEING	HILSN..GEIHMKTMP
d1x6da1_A_98_ILE	HVTILHK	AGLGF	..ITVHR	GLASQ.GTIQK	NEVLSING	RQARE..RQAVIVTRK
d1v62a_A_96_LEU	..VEIVK	LGISL	..ITIDR	SVVDR.GALHP	DHILSIDG	KLLASISEKVRLEILP
d1w9ea1_A_188_MET	REVILCK	LRLKS	..IFVQL	SPASL.VGLRF	DQVLQING	KVLKQ..EKITMTIRD
d2cssa1_A_110_ILE	GRVILNK	LKVVG	..AFITK	SLADVVGHLRA	DEVLEWNG	NIILE..PQVEIIVSR
d1uf1a_A_98_LEU	KKVNLVL	LTIRG	..IYITG	SEAEG.SGLKV	DQILEVNG	RLLKS..RHLILTVKD

[M. Antczak i in., *BMC Bioinformatics* 17 (2016)]

Poszukiwanie motywów

- Wystąpienia tego samego motywu w różnych sekwencjach mogą różnić się sekwencją znaków, długością (istotnie), pozycją (znacznie), może się także zdarzyć brak wystąpienia motywu w którejś z sekwencji ze zbioru
- W grupie problemów związanych z poszukiwaniem motywów można wyróżnić poszukiwanie najdłuższego motywu w zbiorze sekwencji, poszukiwanie największej liczby motywów, poszukiwanie najbardziej zakonserwowanego zbioru motywów
- Problem poszukiwania motywów staje się trudny obliczeniowo przy wielu sekwencjach i zmiennych motywach

Wizualizacja sekwencji konsensusowej

[T.D.Schneider i R.M. Stephens, *Nucleic Acids Res.* 18 (1990)]

```
TACCGATTCATGCATCACGTACGTA
AAGCCATGCTTGTATTTA_TACTCA
A_GCGACGTA_CCCTCACTTGCGCA
TTGGATTTATTGTAATAGGAACGTT
_AGCGACTAC_GTATTAAATCCGCA
TTCCGATGAATGTTTTTGGACCGTA
AAGCGTTGCC_GCATTAATTACTCA
ATGCGATTCATGTATTACGAGCGTA
TTGCGTT_TTGGTATCTGGACCGTA
```

Graficzny sposób prezentowania sekwencji konsensusowej (ang. *consensus logo*, *sequence logo*) niekiedy ułatwia wizualne wyłuskanie pojedynczych motywów, o ile dla instancji jest uzasadnione obliczenie globalnego dopasowania sekwencji.

Poszukiwanie motywów

- Poszukiwanie jednego, nieznanego motywu w wielu sekwencjach można sprowadzić do dopasowania lokalnego wielu sekwencji i rozwiązywać stosownymi podejściami
- Poszukiwanie wielu nieznanych motywów można niekiedy zrealizować, podobnie jak w poprzednim przykładzie, drogą dopasowania globalnego wielu sekwencji i wyłuskania z nich wyróżniających się fragmentów
- Niektóre metody dopasowania globalnego wielu sekwencji bardziej niż inne nadają się do tego celu — te, które nie wymagają utrzymującej się zgodności sekwencji na całych ich długościach

Poszukiwanie motywów

[J. Kececioglu, *Lect. Notes Comput. Sci.* 684 (1993)]

- Progresywna strategia dopasowania wielu sekwencji (także inne) prowadzi do rozwiązania stosunkowo zgodnego z n–1 cząstkowymi dopasowaniami par sekwencji, choć pozostałe $\binom{n}{2} - n + 1$ mogą się z nim w dużej mierze nie zgadzać
- Informacja reprezentowana w zaproponowanym grafie ma umożliwić bardziej kompleksowe spojrzenie na optymalizację dopasowania
- Metoda nie wymaga obliczania dopasowania wszystkich par sekwencji, jednak im więcej, tym bardziej wiarygodna informacja

Poszukiwanie motywów

[J. Kececioglu, *Lect. Notes Comput. Sci.* 684 (1993)] – cd.

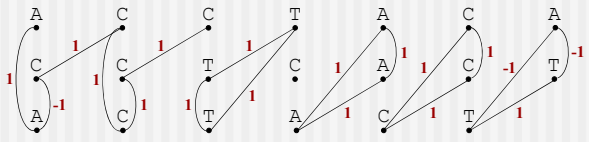
- W grafie dopasowania $G = (V, E, \prec)$ każdy znak ze zbioru wejściowego ma osobny wierzchołek. Relacja $\prec$ w zbiorze wierzchołków zachodzi, jeśli pierwszy znak bezpośrednio poprzedza w którejś sekwencji drugi znak
- Wierzchołki są łączone krawędziami, jeśli w dopasowaniu rozważanej pary sekwencji odpowiadające im znaki zostały sparowane. Spójna składowa grafu odpowiada z grubsza sąsiednim kolumnom poszukiwanego dopasowania
- W modelu z wagami, bardziej odpowiednim w problemie poszukiwania motywów, waga krawędzi oddaje stopień zgodności znaków

Poszukiwanie motywów

[J. Kececioglu, *Lect. Notes Comput. Sci.* 684 (1993)] – cd.

```
sekwencje: ACCTACA, CCTCACT, ACTACT
dopasowania par:    ACCT-ACA     ACCTACA     CCTCACT
                    -CCTCACT     AC-TACT     ACT-ACT

wagi: zgodność pary znaków +1, niezgodność -1
```



Zwarte, zbliżone do klik składowe o wysokich wagach potencjalnie reprezentują fragmenty motywów. Składowe porządkuje się na podstawie relacji poprzedzania. Metoda obejmuje w kolejnym etapie kompresję składowych do wierzchołków połączonych łukami oddającymi relację poprzedzania.

Poszukiwanie motywów

[B. Raphael i in., *Genome Res.* 14 (2004)]

- Graf konstruowany podobnie jak w metodzie Kececioglu, z tym że dla każdej sekwencji wejściowej dodano osobne źródło i ujście w celu łatwiejszej ich identyfikacji
- Kompresja w grafie też jest przeprowadzana, ale w odniesieniu do takich samych sparowanych znaków
- Łuki po skompresowaniu reprezentują bezpośrednie sąsiedztwo w sekwencjach i otrzymują wagę równą liczbie wystąpień danych połączeń w sekwencjach
- Metoda uzupełniona jest o heurystykę usuwającą z grafu krótkie cykle i wybrzuszenia

Poszukiwanie motywów

[B. Raphael i in., *Genome Res.* 14 (2004)] – cd.

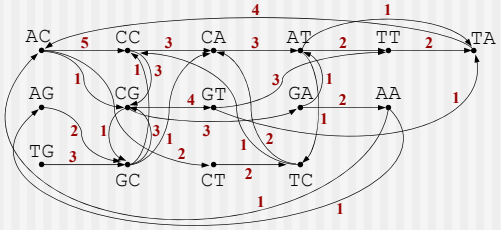
```
sekwencje: ACCTACA, CCTCACT, ACTACT
dopasowania par:    ACCT-ACA     ACCTACA     CCTCACT
                    -CCTCACT     A-CTACT     ACT-ACT
```

Graf przed zastosowaniem heurystyki redukującej wybrzuszenia. Krótkie ścieżki o wysokich wagach potencjalnie reprezentują motywy w sekwencjach.

Poszukiwanie motywów

[Y. Zhang i M.S. Waterman, *J. Comput. Biol.* 10 (2003)]

```
ACCATTACGTT
AGCCATACTCA
ACCGATCCGCA
TGCGAACCGTT
ACCATTACTCA
TGCGAAGCGTT
TGCGTACCGTA
```




```
ACCATTACGTT
AGCCATACTCA
ACCGATCCGCA
TGCGAACCGTT
ACCATTACTCA
TGCGAAGCGTT
TGCGTACCGTA
```


Poszukiwanie motywów

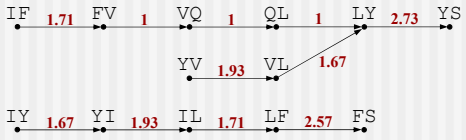
[R. Patwardhan i in., *Lect. Notes Comput. Sci.* 4316 (2006)]

- Graf z metody Zhanga i Watermana (2003) użyty do wyszukiwania nieznanych motywów w wielu sekwencjach
- Metoda została dostosowana do 20-literowego alfabetu aminokwasów, gdzie rzadziej spotyka się identyczne podciągi. Podobne funkcjonalnie białka mogą być zakodowane podobnymi, nieidentycznymi aminokwasami, stąd odwołanie do macierzy substytucji
- Dwa *k*-mery uznane są za podobne, jeśli wszystkie ich aminokwasy z tych samych pozycji mają podobieństwo w macierzy substytucji powyżej pewnego założonego progu (w artykule powyżej 0)

Poszukiwanie motywów

[R. Patwardhan i in., *Lect. Notes Comput. Sci.* 4316 (2006)] – cd.

```
sekwencje: IFVQLYS, IYILFS, YVLYS
3-mery: IFV, FVQ, VQL, QLY, LYS, IYI, YIL, ILF, LFS, YVL, VLY
```



Dla macierzy BLOSUM62 i progu podobieństwa 3, za podobne uznane są (oprócz identycznych) pary V-I, Y-F.

Wagi nie oznaczają już liczby wystąpień *k*-merów w sekwencjach, są przeliczone na podstawie wag z macierzy.

Rozwiązaniem w grafie są krótkie ścieżki wyróżniające się wyższą wagą.
Różne wystąpienia tego samego motywu mogą być reprezentowane różnymi ścieżkami.

Poszukiwanie motywów

[R. Patwardhan i in., *Lect. Notes Comput. Sci.* 4316 (2006)] – cd.

Algorytm obliczania wag w przykładowym grafie:

1. Waga początkowa w_A krawędzi A jak u Zhanga i Watermana
2. Współczynnik podobieństwa (asymetryczny) dla pary podobnych *k*-merów A, B wyliczany jest ze wzoru
$$r_{A,B} = \sum_i s(A_i,B_i) \,/\, \sum_i s(A_i,A_i)$$
gdzie $s(A_i,B_i)$ jest wartością z macierzy substytucji
3. Nowa waga w_A' obliczana jest przez dodanie do w_A wartości $r_{A,B} \cdot w_B$ dla wszystkich *k*-merów B podobnych do A

Poszukiwanie motywów

[C. Boucher i in., *Lect. Notes Comput. Sci.* 4645 (2007)]

- Metoda poszukiwania w zbiorze sekwencji motywów „słabych", zdegenerowanych, w których na stosunkowo wielu pozycjach dopuszcza się wystąpienie niezgodności
- Wykorzystywany jest tu model grafu nieskierowanego ważonego i w podejściu heurystycznym poszukiwane są gęste podgrafy o dużej wadze
- Zastosowanie wag daje nowej metodzie przewagę nad wcześniejszymi algorytmami, w których poszukiwano struktury zbliżone do klik w grafach bez wag. Autorzy wykryli widoczną różnicę w wagach klik odpowiadających i nieodpowiadających rzeczywistym motywom

Poszukiwanie motywów

[C. Boucher i in., *Lect. Notes Comput. Sci.* 4645 (2007)] − cd.

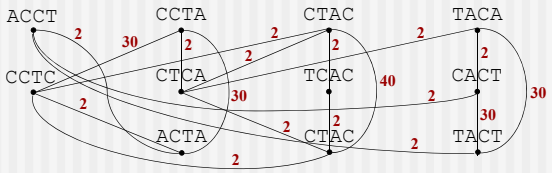
- W pierwszym etapie algorytm *MarkovCluster* poszukuje gęste i wysoko punktowane podgrafy w zdefiniowanym grafie, w drugim etapie dokładny algorytm znajduje w nich motywy
- Wierzchołki grafu odpowiadają wszystkim podciągom o długości l, krawędzie łączą wierzchołki z różnych sekwencji, jeśli odległość Hamminga pomiędzy podciągami jest nie większa niż $2d$, gdzie d jest liczbą dopuszczalnych zdegenerowanych pozycji w motywie. Waga krawędzi jest równa $l-k$ dla odległości Hamminga $d < k \leq 2d$, lub $10(l-k)$ dla $k \leq d$

Poszukiwanie motywów

[C. Boucher i in., *Lect. Notes Comput. Sci.* 4645 (2007)] – cd.

`sekwencje: ACCTACA, CCTCACT, ACTACT`
`długość podciągu` l=4
`limit zdegenerowanych pozycji` d=1

waga krawędzi jest równa
$l-k$, jeżeli $d < k \leq 2d$,
lub $10(l-k)$, jeżeli $k \leq d$



W grafie tym można wyróżnić wiele klik, ale wagi pozwalają łatwo wybrać preferowane czteroliterowe podciągi `CCTA`, `CTAC`, `TACT` i motyw `CCTACT`.

Poszukiwanie motywów

[M. Dogruel i in., *BMC Bioinformatics* 9 (2008)]

- Probabilistyczna metoda NestedMICA poszukiwania krótkich motywów (o długości ≥ 3) w sekwencjach aminokwasowych
- Autorzy oparli swoją metodę na obserwacji, że efektywne poszukiwanie krótkiego motywu zależy ściśle od konkretnej instancji problemu. Dlatego wstępnym etapem tej metody jest analiza „tła”, czyli sekwencji wejściowych
- Skuteczność metody jest większa dla krótkich i rzadkich motywów niż wcześniejszych algorytmów. Dla porównania przetestowano metodę MEME, która motyw trzech aminokwasów wykryła dopiero przy jego obecności w 80% sekwencji, podczas gdy nowa metoda już w 10% sekwencji

Poszukiwanie motywów

[M. Dogruel i in., *BMC Bioinformatics* 9 (2008)] − cd.

MCC (*Matthew's correlation coefficient*) jest wskaźnikiem, jak bardzo motyw odróżnia się od tła.

Distance to miara podobieństwa motywu wskazanego przez algorytm i oryginalnego.