Grafy i sieci – wybrane zagadnienia

wykład 2:
modele służące rekonstrukcji sekwencji

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

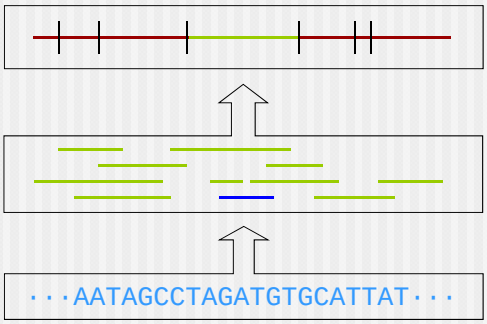
Plan wykładu

1. Modele grafowe problemu sekwencjonowania DNA
2. Grafy quasi-sprzężone
3. Sekwencjonowanie z błędami — przepływ w sieci, ścieżka komiwojażera
4. Modele grafowe problemu asemblacji DNA
5. Przepływ w sieci jako element rozwiązania problemu asemblacji

Strona przedmiotu:
http://www.cs.put.poznan.pl/mkasprzak/doc/doc.html

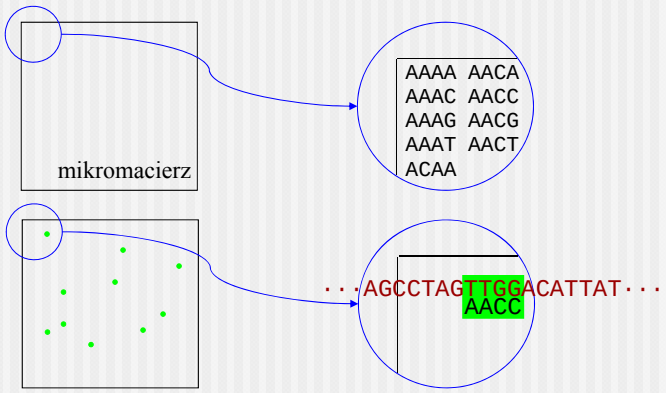
Poznawanie sekwencji genomowej

- Poznawanie sekwencji genomów na trzech poziomach



Sekwencjonowanie przez hybrydyzację

- Eksperyment z biblioteką oligonukleotydów o stałej długości



Sekwencjonowanie DNA

badana sekwencja: CAGTCAGAGTA
długość oligonukleotydów: 4
sekwencjonowanie bez błędów:
{AGAG, AGTA, AGTC, CAGA,
CAGT, GAGT, GTCA, TCAG}

rozwiązania:

CAGTCAGAGTA	CAGAGTCAGTA
CAGT	CAGA
AGTC	AGAG
GTCA	GAGT
TCAG	AGTC
CAGA	GTCA
AGAG	TCAG
GAGT	CAGT
AGTA	AGTA

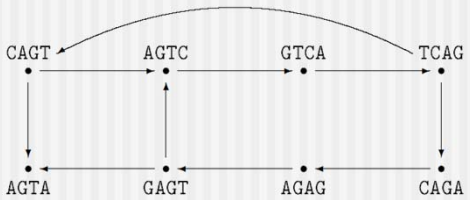
Sekwencjonowanie DNA

[Y.P. Lysov i in., *Doklady Akad. Nauk SSSR* 303 (1988) 1508–1511]

- Pierwszy algorytm odwołujący się do znanego problemu z teorii grafów
- Założenie o braku błędów w instancji
- Poszukiwanie ścieżki Hamiltona w grafie skierowanym, w którym oligonukleotydy odpowiadają wierzchołkom

S = {AGAG, AGTA, AGTC, CAGA, CAGT, GAGT, GTCA, TCAG}

Rozwiązania:
CAGTCAGAGTA
CAGAGTCAGTA



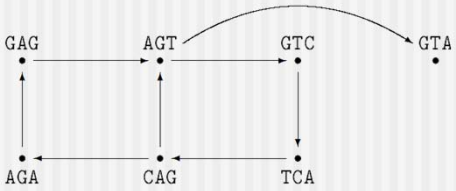
Sekwencjonowanie DNA

[P.A. Pevzner, *J. Biomol. Struct. Dyn.* 7 (1989) 63–73]

- Pierwszy algorytm rozwiązujący problem sekwencjonowania bez błędów w instancji w czasie wielomianowym
- Poszukiwanie ścieżki Eulera w grafie skierowanym, w którym oligonukleotydy odpowiadają łukom

S = {AGAG, AGTA, AGTC, CAGA, CAGT, GAGT, GTCA, TCAG}

Rozwiązania:
CAGTCAGAGTA
CAGAGTCAGTA



Grafy DNA

[J. Błażewicz i in., *Discrete Appl. Math.* 98 (1999) 1–19]

- Równoważność problemów modelowanych grafami z metod Lysova i in. oraz Pevznera stała się inspiracją do dalszych badań nad powiązaniami pomiędzy podobnymi klasami grafów
- Grafy z metody Lysova i in. (tzw. *grafy DNA*) należą do klasy *grafów liniowych* (krawędziowych) skierowanych, dla których problem poszukiwania ścieżki Hamiltona rozwiązywany jest w czasie wielomianowym. Ich *grafami oryginalnymi* są grafy z metody Pevznera
- Wierzchołki grafu liniowego $G=(V,A)$ reprezentują łuki grafu oryginalnego $H=(U,V)$, łuki w G łączą wierzchołek x z y, jeśli w H koniec łuku x pokrywa się z początkiem łuku y

Grafy DNA

[J. Błażewicz i in., *Discrete Appl. Math.* 98 (1999) 1–19] – cd.

- W grafie liniowym (który jest skierowanym 1-grafem) zachodzi następująca własność dla każdej pary $x, y \in V$

$$N^+(x) \cap N^+(y) \neq \emptyset \implies N^+(x) = N^+(y) \quad \wedge \quad N^-(x) \cap N^-(y) = \emptyset$$

- *Graf sprzężony* (który jest skierowanym 1-grafem) to graf, w którym zachodzi dla każdej pary $x, y \in V$ własność

$$N^+(x) \cap N^+(y) \neq \emptyset \implies N^+(x) = N^+(y)$$

Grafy quasi-sprzężone

[J. Błażewicz i in., *Discrete Appl. Math.* 156 (2008) 2573–2580]

- Poszukiwanie szerszej klasy grafów, dla których podobna transformacja byłaby możliwa, zainspirowane problemem izotermicznego sekwencjonowania przez hybrydyzację
- *Graf quasi-sprzężony* to skierowany graf, w którym zachodzi następująca własność dla każdej pary $x, y \in V$

$$N^+(x) \cap N^+(y) \neq \emptyset \implies N^+(x) = N^+(y) \quad \vee \quad N^+(x) \subset N^+(y) \quad \vee \quad N^+(y) \subset N^+(x)$$

- Problem poszukiwania ścieżki/cyklu Hamiltona rozwiązywany jest w takich grafach w czasie wielomianowym

Grafy quasi-sprzężone

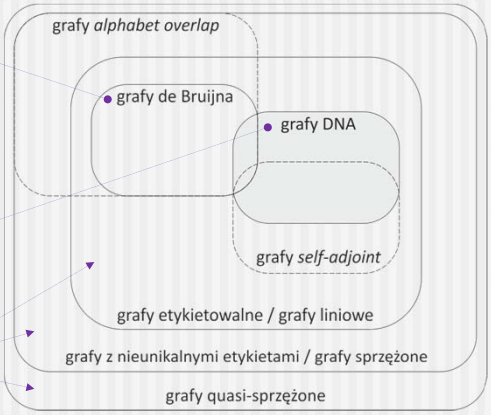
[M. Kasprzak, *Discrete Appl. Math.* 234 (2018) 86–92]

m.in. modele sieci komunikacyjnych

grafy Pevznera – zwykłe podgrafy grafów DNA

grafy Lysova i in. – indukowane podgrafy grafów de Bruijna (α=4)

grafy modelujące sekwencjonowanie izotermiczne



Sekwencjonowanie DNA z błędami

- Przykłady błędów eksperymentalnych

rodzaj błędu	sekwencja	zbiór oligonukleotydów S
negatywny	TTACATTCT	{ACAT, ATTC, TACA, TTAC, TTCT}
negatywny z powtórzeń	TTACATTAC	{ACAT, ATTA, CATT, TACA, TTAC}
pozytywny	TTACATT	{ACAT, CATT, TACA, TTAC, TTAT}

Sekwencjonowanie DNA z błędami

badana sekwencja: CAGTCAGAGTA

długość sekwencji $n = 11$

długość oligonukleotydów $l = 4$

sekwencjonowanie bez błędów:
{AGAG, AGTA, AGTC, CAGA, CAGT, GAGT, GTCA, TCAG}

błędy negatywne: AGAG, AGTC

błąd pozytywny: TCAC

sekwencjonowanie z błędami:
{AGTA, CAGA, CAGT, GAGT, GTCA, TCAC, TCAG}

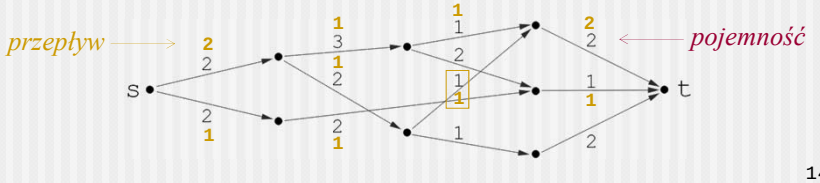
przykładowe rozwiązania:

CAGTCAGAGTA	CAGAGTCAGTA
CAGT	CAGA
GTCA	GAGT
TCAG	GTCA
CAGA	TCAG
GAGT	CAGT
AGTA	AGTA

Sieć przepływowa

- *Sieć przepływowa* jest grafem skierowanym bez pętli własnych wierzchołków z wyróżnionymi dwoma wierzchołkami: źródłem s i ujściem t, takimi że $in(s) = 0$ i $out(t) = 0$
- Każdy łuk takiej sieci ma określoną nieujemną *pojemność*
- *Przepływ* w sieci to funkcja φ przypisująca łukom taką wartość nieujemną, że nie przewyższa ona pojemności łuku i spełnione jest prawo zachowania przepływu:

$$\forall_{x \in V \setminus \{s,t\}} \quad \sum_{(y,x) \in A} \varphi(y,x) = \sum_{(x,y) \in A} \varphi(x,y)$$



Sieć przepływowa

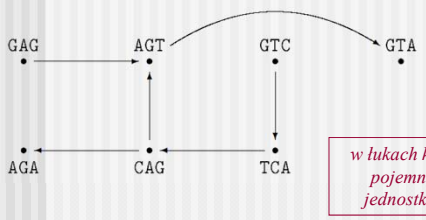
- Całkowita wartość przepływu w sieci wyrażona jest wzorem $\sum_{(s,x)\in A} \varphi(s,x)$. Przepływ o maksymalnej wartości może być znaleziony np. algorytmem Forda-Fulkersona
- W sieci mogą być zdefiniowane także inne wagi, np.:
  - koszty jednostkowego przepływu w łukach; odpowiedni problem może polegać na znalezieniu przepływu o zadanej wartości i minimalnym koszcie
  - dolne ograniczenia na przepływ w łukach; problem polega na znalezieniu przepływu spełniającego oba ograniczenia
- Powyższe problemy rozwiązywane są w wielomianowym czasie

Sekwencjonowanie DNA z błędami

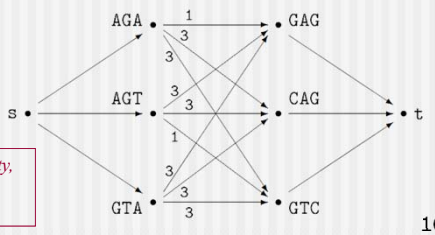
[P.A. Pevzner, *J. Biomol. Struct. Dyn.* 7 (1989) 63–73]

- Założenie o obecności błędów negatywnych w instancji
- Poszukiwanie brakujących oligonukleotydów jako przepływu o wartości m–1 i minimalnym koszcie w sieci opartej na grafie dwudzielnym $K_{m,m}$

$S = \{\mathtt{AGTA}, \mathtt{CAGA}, \mathtt{CAGT}, \mathtt{GAGT}, \mathtt{GTCA}, \mathtt{TCAG}\}$, $n = 11$, $l = 4$



w łukach koszty, pojemność jednostkowa



Sekwencjonowanie DNA z błędami

[P.A. Pevzner, *J. Biomol. Struct. Dyn.* 7 (1989) 63–73] − cd.

- Algorytm heurystyczny wielomianowy
- Niepowodzenie, gdy:
  - przepływ o minimalnym koszcie okaże się mieć koszt inny niż liczba brakujących oligonukleotydów
  - uzupełniony graf okaże się niespójny
  - wierzchołek z brakującymi incydentnymi łukami nie zostanie wstawiony do grafu dwudzielnego (nie zawsze źródło niepowodzenia)

Sekwencjonowanie DNA z błędami

- Sformułowanie problemu klasycznego sekwencjonowania DNA przez hybrydyzację z błędami — wersja optymalizacyjna

  Instancja: Zbiór oligonukleotydów o jednakowej długości l, długość sekwencji oryginalnej n.

  Odpowiedź: Sekwencja o długości $\leq n$ zawierająca maksymalną liczbę oligonukleotydów ze zbioru.
- Problem sekwencjonowania jest silnie NP-trudny, nawet po ograniczeniu błędów do tylko negatywnych lub tylko pozytywnych
- Nawet przy założeniu wiedzy o istnieniu unikalnego rozwiązania w instancji oba podproblemy pozostają trudne obliczeniowo

Sekwencjonowanie DNA z błędami

[J. Błażewicz i in., *J. Comput. Biol.* 6 (1999) 113–123]

- Pierwszy algorytm rozwiązujący problem z dowolnymi błędami negatywnymi i pozytywnymi w instancji
- Sformułowanie wariantu problemu selektywnego komiwojażera w grafie skierowanym pełnym

	AGTA	CAGA	CAGT	GAGT	GTCA	TCAC	TCAG
AGTA	—	4	4	4	4	4	4
CAGA	3	—	4	2	4	4	4
CAGT	1	4	—	4	2	3	3
GAGT	1	4	4	—	2	3	3
GTCA	3	2	2	4	—	1	1
TCAC	4	3	3	4	4	—	4
TCAG	2	1	1	3	3	4	—

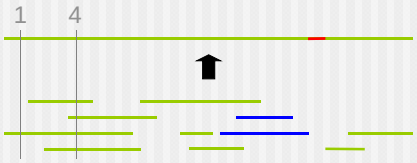
$n = 11, l = 4$

maks. zysk
koszt $\leq n-l$

rozwiązania:
CAGTCAGAGTA
CAGAGTCAGTA

Asemblacja DNA

- Asemblacja jest kolejnym po sekwencjonowaniu etapem składania sekwencji genomowej. Jest problemem trudnym obliczeniowo
- Na skomplikowanie problemu asemblacji wpływają liczba i różnorodność błędów występujących w instancji, a także jej znaczny rozmiar



Asemblacja DNA

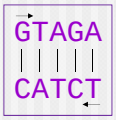
- Sekwencje mogą pochodzić z obu nici DNA i ich orientacja nie jest znana. Nukleotydy *komplementarne*: C–G, A–T
- Zawierają przekłamania przeniesione z etapu sekwencjonowania: insercje, delecje i zamiany nukleotydów

Fragmenty składowe:

TATGC, ATCAGCAC, GACTC, GTAGA, GCATCA

Jedno z możliwych rozwiązań:

```
TATGCAGCACTCTAC
---------------
TATGC G-ACTC
   GCATCA TCTAC
 AT-CAGCAC
```



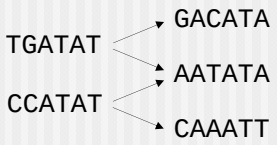
Asemblacja DNA

- Trzy modele grafowe problemu asemblacji DNA
  - model trójetapowy *overlap-layout-consensus* — sekwencje w wierzchołkach
  - model grafu dekompozycji — sekwencje w seriach łuków o nakładających się etykietach
  - model *string graphs* — sekwencje w seriach łuków o rozłącznych etykietach
- Niezależnie od modelu, ze względu na nieznaną orientację, sekwencje wejściowe są na wstępie dublowane w sensie odwrotnie komplementarnych odpowiedników

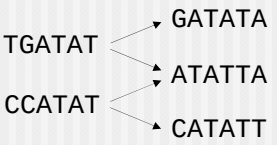
{TATGC, ATCAGCAC, GACTC, GTAGA, GCATCA,
 GCATA, GTGCTGAT, GAGTC, TCTAC, TGATGC}

Model *overlap-layout-consensus*

- Przypomina model grafowy z metody Lysova i in., lecz łuki mają wagi
- Wagi opisują stopień dopasowania sekwencji w wierzchołkach, zwykle długość oraz wartość dopasowania
- Dopuszczenie niedokładnych i/lub nierównej długości nałożeń sekwencji sprawia, że graf taki nie jest już w ogólności grafem liniowym, nie jest nawet grafem quasi-sprzężonym



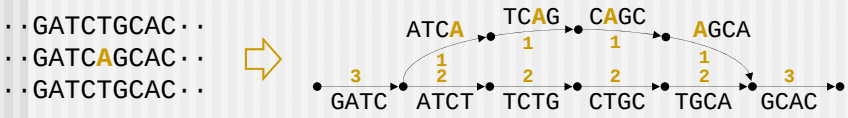
nałożenia niedokładne i równe – dopuszczony jeden błąd



nałożenia dokładne i nierówne

Model grafu dekompozycji

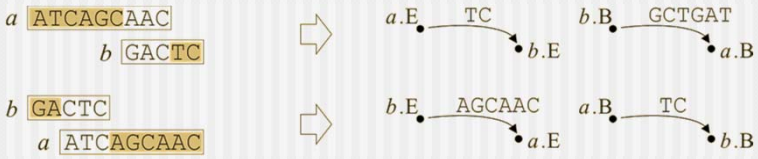
- Po wstępnej dekompozycji sekwencji na krótsze, nakładające się oligonukleotydy o równej długości konstruowany jest graf jak w metodzie Pevznera (częściowa utrata informacji)
- Łuki mogą mieć wagę reprezentującą liczbę sekwencji zawierających dany oligonukleotyd
- Błędy sekwencjonowania objawiają się charakterystycznymi strukturami w grafie o niskiej wadze, częściowo korygowanymi



Model *string graphs*

[E.W. Myers, *Bioinformatics* 21 suppl. 2 (2005) ii79–ii85]

- Dopasowanie sekwencji jak w modelu trójetapowym, wystające odcinki z dopasowania każdej pary sekwencji tworzą łuki. Wierzchołki odpowiadają punktom początkowym i końcowym odcinków i nie są związane z żadnymi ciągami znaków

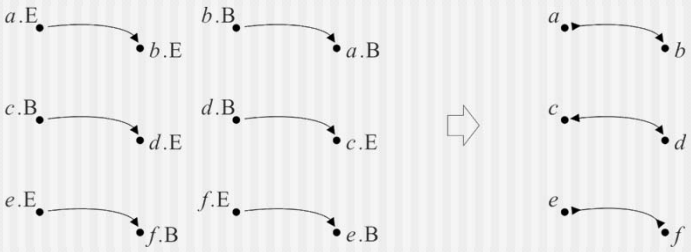


- Środkowe odcinki dopasowania reprezentowane są serią innych wystających odcinków

Model *string graphs*

[E.W. Myers, *Bioinformatics* 21 suppl. 2 (2005) ii79–ii85] – cd.

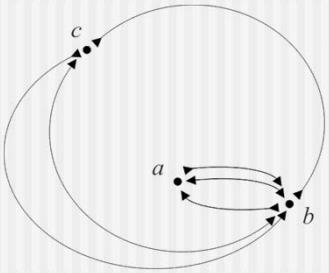
- Graf upraszczany jest przez usunięcie łuków „nadrzędnych” reprezentowanych przez inne łuki składowe. Ścieżki proste bez rozgałęzień są kompresowane do pojedynczych łuków
- Komplementarne odcinki genomu układają się w grafie w łuki przeciwrównoległe i zastępowane są dwukierunkową krawędzią



Model *string graphs*

[E.W. Myers, *Bioinformatics* 21 suppl. 2 (2005) ii79–ii85] – cd.

- W trakcie konstrukcji rozwiązania przechodzić przez wierzchołek można w obie strony, jednak z zachowaniem zasady, żeby zwrot przy wychodzeniu był przeciwny do zwrotu zaobserwowanego przy wchodzeniu do wierzchołka



Przykład zawiera kilka możliwych rozwiązań, wszystkie oparte na uszeregowaniu a–b–c o różnych orientacjach sekwencji.

W tym przypadku nie ma potrzeby wielokrotnego przechodzenia krawędzi.

Model *string graphs*

[E.W. Myers, *Bioinformatics* 21 suppl. 2 (2005) ii79–ii85] – cd.

- Oszacowanie, ile razy należy przejść daną krawędź, odbywa się na podstawie jej długości oraz pochodzenia. Odwzorowane jest wagami w grafie:
  - dokładnie 1, gdy krawędź ma odpowiednio długą etykietę w odniesieniu do rozmiaru całej instancji
  - ≥ 0, gdy krawędź nie powstała na drodze kompresji łuków ani nie jest potwierdzona relacją zawierania sekwencji
  - ≥ 1 wpp.
- Sekwencje wejściowe zawarte w innych nie biorą udziału w konstrukcji grafu, mają jednak wpływ na oszacowanie wwym. wag

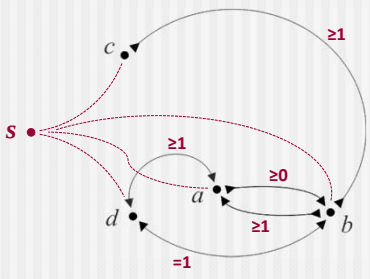
Model *string graphs*

[E.W. Myers, *Bioinformatics* 21 suppl. 2 (2005) ii79–ii85] – cd.

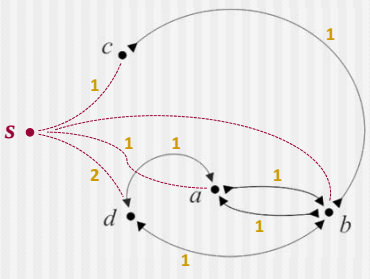
- Przez rozwiązanie problemu przepływu w sieci ustala się, które krawędzie powinny być włączone więcej niż raz
- Przepływ ma spełniać dolne i górne ograniczenia na jego wartość określone w poprzednim kroku i charakteryzować się minimalnym kosztem, gdzie każda krawędź ma koszt równy 1
- Dodawany jest wierzchołek *s* połączony przeciwrównoległymi łukami o nieograniczonej pojemności i koszcie równym 1 ze wszystkimi wierzchołkami grafu
- Każdy jednostkowy przepływ przez *s* interpretowany jest jako rozpoczęcie składania nowego kontigu

Model *string graphs*

[E.W. Myers, *Bioinformatics* 21 suppl. 2 (2005) ii79–ii85] – cd.



Jeden wierzchołek *s* można zastąpić parą *s*, *t* z łukami skierowanymi w jedną stronę



Przepływ o minimalnym koszcie zgodny ze zwrotami krawędzi