Grafy i sieci – wybrane zagadnienia

wykład 3:
modele służące porównywaniu sieci

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

Plan wykładu

1. Sieci jako modele interakcji pomiędzy obiektami
2. Graflety
3. Struktura biomolekuł jako rodzaj sieci przestrzennej
4. Deskryptory
5. Model kombinatoryczny problemu porównywania deskryptorów

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

Sieci

- Sieciami nazywane są grafy występujące w szczególnym kontekście — jako modele np. zagadnień transportowych, komunikacyjnych, biologii systemowej. Mogą być skierowane lub nieskierowane
- Wierzchołki sieci nazywane są węzłami
- Zagadnienia z zakresu bieżącego wykładu odnoszą się do grafów nieskierowanych

Sieci biologiczne

- Informacja o interakcjach, jakie zachodzą pomiędzy biomolekułami, daje możliwość badania systemów biologicznych jako złożoną sieć zależności
- Sieci odwzorowujące powiązania funkcjonalne pomiędzy genami lub odpowiadającymi im białkami mogą opierać się przykładowo na:
  - informacji o ekspresji genów
  - informacji o interakcjach cząsteczek białkowych
  - bazie danych ontologii genów
- Połączenie kilku źródeł danych daje bardziej kompleksowy obraz funkcjonowania systemu biologicznego

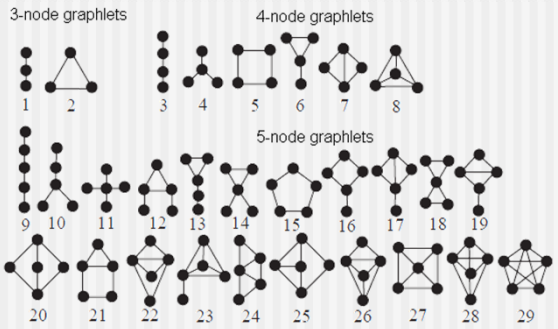
Sieci biologiczne

- Analiza sieci biologicznych może przykładowo polegać na:
  - porównywaniu całych sieci wygenerowanych dla różnych organizmów lub na podstawie różnych eksperymentów
  - porównywaniu fragmentów sieci
  - grupowaniu obiektów o podobnej charakterystyce
  - wyszukiwaniu aktywnych podsieci
- Wiele problemów kombinatorycznych związanych z analizą sieci biologicznych jest trudnych obliczeniowo (np. bazujących na izomorfizmie podgrafów, klastrowaniu). Dodatkowo dochodzi aspekt niepełnej i niepewnej informacji pozyskanej eksperymentalnie oraz zwykle duży rozmiar instancji

Graflety

[N. Przulj i in., *Bioinformatics* 20 (2004) 3508–3515]

- Graflety (ang. *graphlets*) to małe spójne struktury grafowe, które jako indukowane podgrafy sieci służą do analizy ich topologii



Graflety

[N. Przulj i in., *Bioinformatics* 20 (2004) 3508–3515] – cd.

- Wektor liczby wystąpień grafletów w sieci daje przybliżoną informację o jej strukturze
- Dwie sieci o zbliżonych wartościach w wektorach lub zbliżonych częstościach wystąpień grafletów mogą charakteryzować się pewnymi podobieństwami strukturalnymi
- W celu zniwelowania różnicy pomiędzy bardzo często i bardzo rzadko występującymi grafletami obliczany jest logarytm z częstości ich wystąpień

Przykładowa różnica w rzeczywistej sieci biologicznej:
liczba wystąpień grafletu nr 9 = 23854301
liczba wystąpień grafletu nr 29 = 869

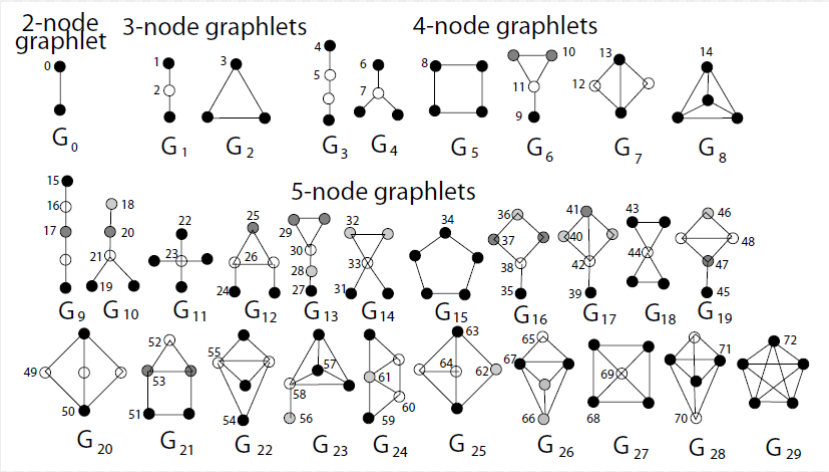
Graflety

[T. Milenkovic i N. Przulj, *Cancer Informatics* 6 (2008) 257–273]

- Wektory grafletów można obliczać nie tylko dla całej sieci, ale także dla pojedynczych węzłów — wtedy jest to liczba grafletów danego rodzaju, w skład których wchodzi dany węzeł
- Analiza grafletowa może dostarczyć więcej informacji, jeśli uwzględni się, które miejsce grafletu dany węzeł sieci pokrywa
- W grafletach wyróżniono tzw. *orbity*, czyli wierzchołki, które odwzorowują się na siebie w przekształceniu automorficznym. Są 73 różne orbity w grafletach o rozmiarach do pięciu wierzchołków i takiej długości wektory reprezentują węzły sieci

Graflety

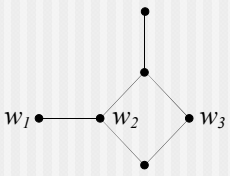
[T. Milenkovic i N. Przulj, *Cancer Inform.* 6 (2008) 257–273] – cd.



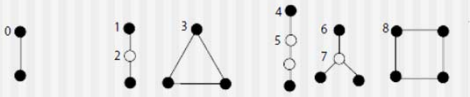
Graflety

[T. Milenkovic i N. Przulj, *Cancer Inform.* 6 (2008) 257–273] – cd.

- Przykład wyliczania wektorów orbit



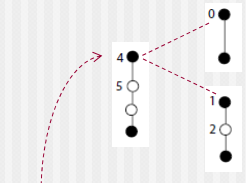
	0	1	2	3	4	5	6	7	8	
w_1	1	2	0	0	3	0	1	0	0	⋯
w_2	3	3	3	0	0	4	1	1	1	⋯
w_3	2	3	1	0	2	1	1	0	1	⋯



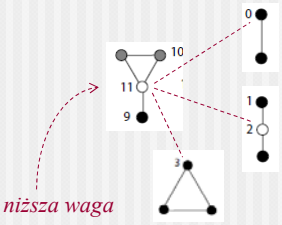
Graflety

[T. Milenkovic i N. Przulj, *Cancer Inform.* 6 (2008) 257–273] – cd.

- Mniejsze graflety zawarte są w większych, a więc to samo miejsce w sieci wpływa na zawartość kilku pozycji wektora. Graflety, które zawierają mniej innych, są bardziej znaczące. Wektor orbit jest przeskalowywany wagami oddającymi tę właściwość



wyższa waga



niższa waga

Graflety

- Graflety o rozmiarze do 5 wierzchołków wydają się najbardziej odpowiednie do analizy sieci — zapewniają dostateczną wartość informacyjną i akceptowalną złożoność obliczeniową
- Analiza sieci z użyciem grafletów może służyć różnym celom:
  - grupowaniu węzłów mających zbliżone powiązania w swoich sąsiedztwach
  - identyfikowaniu podobnych podstruktur w sieciach
  - porównywaniu całych sieci
- Do reprezentacji większych części sieci wystarcza krótszy wektor wystąpień grafletów
- Poza topologią, nie bierze się tu pod uwagę innych informacji: wag krawędzi, etykiet węzłów, itp.

Graflety

- Nie zawsze podobne wektory orbit mogą być przesłanką do stwierdzenia, że dane węzły są bliskie funkcjonalnie

⋯	0	0	1	0	0	0	0	0	0	0	⋯
⋯	0	0	1	0	0	0	0	1	0	0	⋯

vs.

⋯	0	0	21	0	614	0	0	0	0	0	⋯
⋯	0	0	21	0	614	0	0	1	0	0	⋯

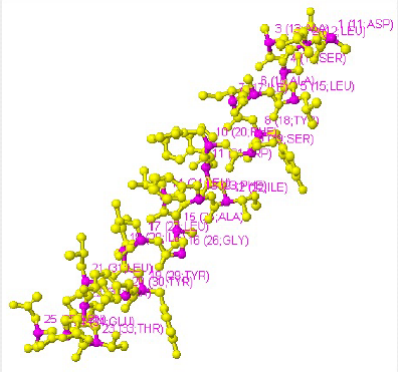
- Miejsca typu „środek długiej ścieżki bez bocznych odgałęzień” czy „krawędzie izolowane” będą miały identyczne wektory, lecz znaczenie takiego podobieństwa zazwyczaj nie jest duże
- Wierzchołki mające bardzo liczne sąsiedztwo wnoszą wiele wystąpień w wektorze, ale niekoniecznie istotnych dla analizy

Struktura przestrzenna biomolekuł

- Cząsteczki biologiczne, np. RNA i białka, poprzez swoją strukturę przestrzenną wpływają na procesy zachodzące w organizmach. Rozpoznanie ich kształtu jest więc kluczowe, jeśli chcemy zrozumieć pełnioną przez nie funkcję
- Analiza struktury przestrzennej biomolekuł może polegać np. na dopasowywaniu całych struktur w celu odkrycia podobieństwa funkcjonalnego lub ich fragmentów w celu zidentyfikowania możliwych kompleksów cząsteczek
- Struktura przestrzenna RNA i białek jest efektem zwijania się liniowych polimerów: łańcucha rybonukleotydów (RNA) lub łańcucha aminokwasów (białko), pod wpływem oddziaływań międzyatomowych

Struktura przestrzenna biomolekuł

Biomolekuły mogą być zwizualizowane jako sieć atomów z odwzorowanymi wiązaniami pomiędzy nimi, gdzie każdy atom stanowi węzeł o pewnych przestrzennych koordynatach



[M. Antczak, rozprawa doktorska (2013)]

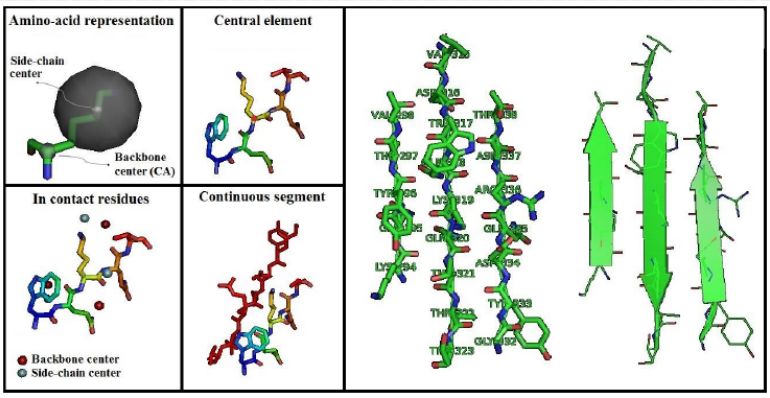
Deskryptory

[T.R. Hvidsten i in., *Bioinformatics* 19 suppl. 2 (2003) ii81–ii91]

- Deskryptor jest swego rodzaju motywem przestrzennym w sieci strukturalnej. Pojęcie zostało zdefiniowane początkowo dla białek, ale może zostać rozszerzone na inne cząsteczki
- Deskryptor identyfikowany jest w cząsteczce białkowej jako strukturalne sąsiedztwo pewnego aminokwasu uznawanego za centrum deskryptora. W odróżnieniu od wcześniej rozważanych motywów strukturalnych, nie musi tworzyć jednej spójnej formy
- Dla zadanej odległości od centrum deskryptora wykreślana jest sfera o takim promieniu i wszystkie aminokwasy wewnątrz sfery wchodzą w skład deskryptora, wraz z ich sekwencyjnym sąsiedztwem o zadanej długości (zwykle po dwa aminokwasy z każdej strony)

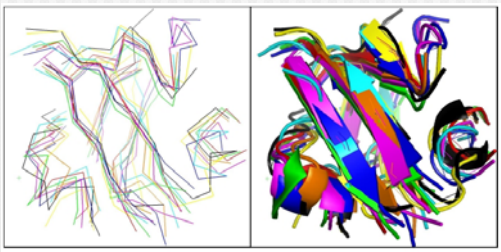
Deskryptory

Schemat tworzenia deskryptora białkowego



[M. Antczak i in., *BMC Bioinformatics* 17 (2016) 383]

Deskryptory



Deskryptory jako zbiory segmentów. Każdy segment powstaje przez złożenie elementów, a one przez rozszerzanie o sekwencyjne sąsiedztwo aminokwasów objętych sferą wokół centrum.

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) 383]

Miara odległości motywów

- Uogólniając do dowolnych sieci strukturalnych z węzłami o koordynatach przestrzennych, można przy konstruowaniu motywu pominąć odwołanie do sekwencji polimeru, motyw byłby wtedy podsiecią mieszczącą się wewnątrz pewnej sfery
- Miara RMSD (ang. *root-mean-square deviation*) służy ocenie odległości obiektów opisanych wartościami w *n*-elementowych wektorach

$$RMSD(x, y) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}}$$

- W odniesieniu do cząsteczek, miara RMSD zlicza odległości w przestrzeni pomiędzy zbiorami atomów po wcześniejszym ich przyporządkowaniu (uszeregowaniu w wektorach)

Miara odległości motywów

- Odległość pomiędzy motywami w dowolnych sieciach przestrzennych musiałaby uwzględnić dodatkowo różnice w obecności krawędzi pomiędzy danymi parami węzłów
- Samo przyporządkowanie węzłów dwóch sieci przestrzennych charakteryzuje się bardzo dużą złożonością obliczeniową. Złożoność jest mniejsza, jeśli założy się obecność pewnych konserwatywnych strukturalnie fragmentów w motywach, których odległość RMSD będzie bardzo mała
- Takie założenie obowiązuje w porównywaniu deskryptorów, gdzie klasyfikuje się jako niepodobne deskryptory o odległych w sensie RMSD elementach centalnych. Pozostałe deskryptory dużo łatwiej dopasować w przestrzeni na podstawie dopasowania strukturalnego ich elementów centralnych

Porównywanie deskryptorów

[M. Antczak i in., BMC Bioinformatics 17 (2016) 383]

- Pierwsze zastosowanie modelu kombinatorycznego do rozwiązania tego problemu biologicznego
- Porównywane są pary deskryptorów po wstępnym odsianiu deskryptorów niepodobnych:
  - o odległych strukturalnie elementach centralnych (wartość RMSD większa niż 1,2 Å)
  - o znacznej różnicy w liczbie elementów (większej niż 4:5)
- W celu ustabilizowania względem siebie dwóch deskryptorów w przestrzeni wykorzystana została koncepcja *dupleksów*, które są parą elementów zawierającą zawsze element centralny

Porównywanie deskryptorów

[M. Antczak i in., BMC Bioinformatics 17 (2016) 383] – cd.

- Problem największego strukturalnego dopasowania dwóch deskryptorów został sformułowany jako odmiana problemu przydziału (ang. *maximum-size assignment*), w którym maksymalizowana jest liczba dopasowanych dupleksów tych deskryptorów przy zachowaniu ograniczeń:
  - dupleks może zostać dopasowany w rozwiązaniu do co najwyżej jednego dupleksu drugiego deskryptora
  - suma kosztów dopasowania par dupleksów, określonych jako ich odległość RMSD, nie może przekroczyć pewnego określonego limitu

Porównywanie deskryptorów

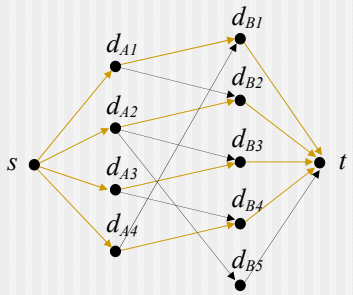
[M. Antczak i in., BMC Bioinformatics 17 (2016) 383] – cd.

- Ostatecznie rozwiązania problemu największego strukturalnego dopasowania są weryfikowane pod względem całościowej wartości RMSD, gdyż suma wartości dla poszczególnych dupleksów jest jedynie wskazówką w poszukiwaniu rozwiązania
- Problem dopasowania deskryptorów w podanym sformułowaniu może zostać rozwiązany z zastosowaniem np. metody węgierskiej albo poprzez sprowadzenie do problemu przepływu w sieci opartej na grafie dwudzielnym

Porównywanie deskryptorów

Deskryptory A i B reprezentowane są zbiorami dupleksów.
W macierzy kosztów wartości RMSD dla poszczególnych par dupleksów.

	d_{B1}	d_{B2}	d_{B3}	d_{B4}	d_{B5}
d_{A1}	2,1	1,9	M	M	M
d_{A2}	M	1,8	1,5	M	2,1
d_{A3}	M	M	2,6	2,2	M
d_{A4}	3,0	M	M	1,9	M



Sieć przepływowa ma wszystkie pojemności równe 1.
Koszty łuków incydentnych z *s* i *t* równe 0.