Algorytmy kombinatoryczne w bioinformatyce

wykład 8: podsumowanie

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

Obecność błędów a złożoność problemów

- Część z zaprezentowanych modeli grafowych odnosi się do tych problemów biologii obliczeniowej, w których zakłada się brak błędów eksperymentalnych. W przypadku danych obarczonych błędami albo konstrukcja takich grafów nie jest możliwa, albo towarzyszący modelowi dokładny algorytm wielomianowy przestaje działać
- Przykładowo, mapowanie przez hybrydyzację do unikalnych próbek w przypadku niepoprawnych danych skutkuje grafem, który może nie być interwałowy
- W przypadku sekwencjonowania przez hybrydyzację i błędów, oba grafy Lysova i Pevznera mogą zostać zbudowane i transformacja pomiędzy nimi ciągle jest możliwa, ale wtedy grafy nie zawierają (w ogólności) ścieżki Hamiltona/Eulera

Obecność błędów a złożoność problemów

- W praktyce rezultaty eksperymentów biologicznych prawie zawsze niosą za sobą błędy. Niestety, odpowiednie problemy kombinatoryczne zazwyczaj nie są łatwe obliczeniowo. Prawidłowość ta potwierdza się w przypadku poruszonych na wykładzie problemów związanych z pierwszorzędową strukturą DNA
- Klasyczne sekwencjonowanie DNA przez hybrydyzację jest rozwiązywalne w czasie wielomianowym jedynie dla instancji pozbawionych błędów. Problem jest silnie NP-trudny nawet po ograniczeniu błędów do tylko negatywnych albo tylko pozytywnych. Nawet warianty problemu z dodatkową wiedzą o unikalności rozwiązania są trudne obliczeniowo

Obecność błędów a złożoność problemów

- Złożoność problemu sekwencjonowania z ograniczeniem do błędów pochodzących jedynie z powtórzeń oligonukleotydów w sekwencji oryginalnej jest zagadnieniem otwartym
- W izotermicznym sekwencjonowaniu DNA mamy sytuację analogiczną do klasycznego sekwencjonowania: problem bez błędów jest łatwy obliczeniowo, podczas gdy oba warianty przeszukiwania z błędami tylko negatywnymi albo tylko pozytywnymi są silnie NP-trudne
- Następnym etapem rozpoznawania genomów jest asemblacja. Problem jest silnie NP-trudny nawet w bardzo uproszczonej, teoretycznej wersji (patrz problem najkrótszego wspólnego superciągu)

Obecność błędów a złożoność problemów

- Fragmenty DNA mogą być porządkowane na etapie mapowania. Mapowanie może zostać zrealizowane poprzez hybrydyzację do unikalnych próbek albo z użyciem enzymów restrykcyjnych
- Mapowanie przez hybrydyzację jest łatwe przy braku błędów i silnie NP-trudne w przeciwnym przypadku. Mapowanie restrykcyjne pod postacią DDP lub SPDP jest silnie NP-trudne nawet bez błędów. PDP bez błędów jest znanym otwartym problemem także w obszarze geometrii obliczeniowej, gdzie ma swojego odpowiednika. Dowody silnej NP-zupełności problemu PDP z błędami zostały przeprowadzone

Rozmiar instancji a złożoność problemów

- Na złożoność obliczeniową problemu może mieć także wpływ ograniczenie rozmiaru instancji
- Przykładem może być problem konstrukcji idealnego drzewa filogenetycznego na podstawie macierzy znakowej — łatwy obliczeniowo, gdy mamy ograniczoną liczbę stanów lub ograniczoną liczbę cech
- Złożoność algorytmu programowania dynamicznego dopasowującego sekwencje w sposób dokładny rośnie z liczbą sekwencji — liczba sekwencji jest w wykładniku funkcji złożoności obliczeniowej algorytmu. Problem optymalnego dopasowania wielu sekwencji jest silnie NP-trudny

Modelowanie problemów biologicznych

- Niektóre problemy kombinatoryczne częściej niż inne wykorzystywane są do modelowania problemów biologicznych. Ich przydatność wynika ze struktury instancji idealnie oddającej zależności pomiędzy obiektami biologicznymi, za częścią z nich dodatkowo przemawia ich niska złożoność obliczeniowa
- Przykładem przydatnego i łatwego obliczeniowo problemu jest problem poszukiwania ścieżki Eulera w grafie skierowanym. Jest on wykorzystywany do zamodelowania prostszych problemów i rozwiązania ich w sposób dokładny, ale także do heurystycznego rozwiązania problemów trudniejszych
  - SBH bez błędów [Pevzner, 1989]
  - dopasowanie wielu sekwencji [Zhang i Waterman, 2003]
  - asemblacja [np. Medvedev i in., 2011]

Modelowanie problemów biologicznych

- Przykłady problemów trudnych obliczeniowo użytecznych w modelowaniu problemów biologicznych
  - problem komiwojażera
    - SBH z błędami [Błażewicz i in., 1999]
    - mapowanie przez hybrydyzację z błędami [Alizadeh i in., 1995]
  - problem kolorowania wierzchołkowego grafu
    - konstrukcja drzew filogenetycznych dla macierzy znakowych [Kannan i Warnow, 1992]
    - konstrukcja drzew konsensusowych [Bonnard i in., 2006]
  - problem poszukiwania maksymalnej kliki (klik) w grafie
    - poszukiwanie motywów [Boucher i in., 2007]
    - konstrukcja drzew konsensusowych [Bonnard i in., 2006]

Operacje na sekwencjach znaków

- Rozważane na wykładach problemy bioinformatyczne odnosiły się do zagadnień związanych z poznawaniem i analizą pierwszorzędowej struktury DNA. Powiązane z nimi algorytmy operowały więc głównie na sekwencjach znaków
  - wyszukiwanie podciągów
    - motywy
    - dopasowanie lokalne
  - konstruowanie superciągu
    - sekwencjonowanie
    - asemblacja
  - porównywanie sekwencji
    - dopasowanie globalne
    - faktoryzacja LZ w konstrukcji drzew filogenetycznych

Podsumowanie

- Zagadnienia zaprezentowane na wykładach pokrywają jedynie niewielki obszar bioinformatyki / biologii obliczeniowej. Szersze spektrum problemów zawarte jest w treściach innych przedmiotów
- Także od strony metodyki zaprezentowane podejścia zostały ograniczone — do modelowania problemów rzeczywistych na gruncie kombinatorycznym. Praktyka pokazuje, że takie ograniczenie nie przeszkadza osiągać trafnych rozwiązań
- Znane problemy kombinatoryczne są użyteczne m.in. dlatego, że zazwyczaj istnieje już dla nich szeroki wachlarz algorytmów dokładnych i/lub przybliżonych. Jeśli nie pasują do problemu w 100%, mogą posłużyć do jego rozwiązania w sposób heurystyczny lub jako wskazówka do nowego algorytmu

Podsumowanie

- Konstruując algorytm rozwiązujący nowy problem biologiczny, należy zachować właściwą kolejność zadań:
  - formalny zapis problemu,
  - zbadanie jego złożoności obliczeniowej (w miarę możliwości),
  - dobranie optymalnego schematu algorytmu do złożoności problemu i rodzaju rozwiązania, które zamierzamy osiągnąć (dokładne, przybliżone)
- Problemy o nieokreślonej złożoności obliczeniowej (problemy otwarte) rozwiązuje się jak problemy obliczeniowo trudne