Wstęp do Biologii Obliczeniowej

Zagadnienia na kolokwium

Bartek Wilczyński

5. czerwca 2018

Sekwencje DNA i grafy

- Sekwencje w biologii, DNA, RNA, białka, alfabety, transkrypcja DNA→RNA, translacja RNA→białko, komplementarność sekwencji, replikacja DNA, kod genetyczny
- Widmo k-merowe sekwencji, sekwencjonowanie przez hybrydyzację, podejście Eulerowskie i Hamiltonowskie, Grafy deBruijna
- Mikromacierze, problem projektowania unikalnych sond dla mikromacierzy, temperatura topnienia DNA

Ewolucja sekwencji biologicznych

- Koncepcja jednego drzewa filogenetycznego łączącego wszystkie organizmy żywe jako liście
- Replikacja DNA, mutacje, selekcja
- Parsymoniczne modele ewolucji, odwracalność czasu, rekonstrukcja sekwencji przodka
- Modele Markowa ewolucji DNA (JC69, K80, F81), parametryzacja, estymacja odległości ewolucyjnej na podstawie mutacji, estymacja parametrów modelu
- Modele ewolucji sekwencji białkowych: PAM, BLOSSUM, macierze log-odds

Porównywanie 2 sekwencji

- Odległość Hamminga i zliczanie mutacji, rodzaje mutacji (kodujące i nie-kodujące), Macierze identyczności
- Rodzaje błędów w replikacji DNA – insercje, delecje, substytucje i sposoby ich powstawania
- Odległość edycyjna, scenariusze edycyjne, optymalne scenariusze, problemy z wyliczaniem
- Uliniowienia: definicja, relacja ze scenariuszem edycyjnym, przestrzeń uliniowień, uliniowienia lokalne i globalne, ocena podobieństwa sekwencji
- Algorytmy do znajdowania uliniowień globalnych i lokalnych (Needleman-Wunsh, Smith-Waterman)
- Kary za przerwę: stałe, afiniczne

Konstrukcja drzew filogenetycznych

- Macierze odległości vs. Macierze podobieństwa
- Drzewa filogenetyczne, binarne i gwiaździste, ukorzenione i nieukorzenione, ukorzenianie, przestrzeń drzew
- Macierze odległości vs. Drzewa filogenetyczne, ultrametryczność, algorytm UPGMA (klastrowanie hierarchiczne)
- Hipoteza zegara ewolucyjnego, drzewa nie-ultrametryczne, algorytm neghbor-joining

Uliniowienia wielu sekwencji

- Niezgodności w estymacji sekwencji przodka na podstawie wielu porównań parami
- Multi-uliniowienie – definicja jako uogólnienie uliniowienia 2 sekwencji, brak interpretacji edycyjnej
- Miara sumy par (SP) dla multi-uliniowień, naiwny algorytm (rozwinięcie programowania dynamicznego), heurystyczne poprawki (Carillo-Lipmann), NP-trudność problemu
- Uliniowienie progresywne – pojęcie profilu i uliniawiania profilów, algorytm CLUSTALW,
- problemy z uliniowieniem progresywnym, heurystyki T-Coffee i MUSCLE

Ukryte modele Markowa

- Ukryte Modele Markowa (HMM) z czasem dyskretnym, macierz przejścia, macierz emisji, symulacje przebiegów(trajektorii), ciąg stanów, ciąg emisji
- Znajdowanie przebiegu na podstawie ciągu emisji (algorytm Viterbiego)
- Znajdowanie prawdopodobieństw przejścia przez stan w danym kroku (metoda forward-backward)
- Estymacja macierzy przejścia i emisji (algorytm Bauma-Welcha)

Modele Markowa do uliniowień

- Modele wyższego rzędu do reprezentacji generatywnej sekwencji (np. wyspy CpG)
- Modele interpolowane (IMM) do reprezentacji złożonych sekwencji
- Ukryte Modele Markowa do profilów sekwencyjnych (HMMER)
- Modele o zmiennym rzędzie (VOM)

Wyszukiwanie sekwencji podobnych

- Ogromne podobieństwo sekwencji genowych pomiędzy bardzo odległymi gatunkami
- Bazy danych sekwencji, cele i zasady działania
- Problem wyszukiwania sekwencji: krótka sekwencja zapytania, bardzo duża baza, konieczność wykonywania wielu zapytań
- Problemy z uliniowieniem, heurystyczne podejście, algorytm FASTA
- Alforytm BLAST: metoda, model statystyczny oceny trafień, testowanie hipotez, rozkład wartości ekstremalnych

Relacje homologii i funkcje genów

- Duplikacje i specjacje – jako model ewolucji funkcji genów
- Szybsza ewolucja paralogów i neo-funkcjonalizacja
- Onologi i ksenologi
- Różne rodzaje rodzin genów w zależności od tolerancji dla paralogów
- Wyszukiwanie potencjalnych homologów: dwustronne trafienia BLAST i klastry genów ortologicznych
- Ontologie funkcjonalne – Gene Ontology
- Testy statystyczne nadreprezentacji funkcji: test Fishera i Gene-Set Enrichment Analysis (GSEA)

Uzgadnianie drzew

- Drzewa genów i drzewa gatunków, scenariusze ewolucyjne, uzgodnienie drzew, relacja scenariusza ewolucyjnego i relacji para- orto-logii
- Różne funkcje kosztów uzgadniania: duplikacje, straty, koalescencja
- Mapowanie LCA i jego relacja z kosztami (D,L, DL, DC)
- Znajdowanie optymalnego uzgodnienia w sensie DL
- Horyzontalne transfery genów i sieci filogenetyczne
- Przykłady zastosowania uzgadniania do rzeczywistych danych

Motywy sekwencyjne

- Rola niekodującego DNA i wiązanie czynników transkrypcyjnych
- Miejsca wiązania i ich uliniowienia bez spacji
- Macierze zliczeń, częstości i log-odds (PWM)
- Zawartość informacyjna pozycji i logo motywu
- Bazy danych motywów
- Problem wyszukiwania motywów, algorytmy AlignACE i MEME

Mapowania odczytów NGS

- Metody sekwencjonowania – Sangera i nowej generacji: wady i zalety
- Problem mapowania krótkich sekwencji na znany genom
- Drzewa sufiksowe, macierze sufiksowe, transformata Burrowsa-Wheelera
- Efektywność indeksów opartych na BWT (indeks Ferraginy-Manziniego)
- Różnorodność genetyczna w populacji, problem błędów, metoda dzielenia odczytów
- Szybkie metody kwantyfikacji odczytów: STAR, Kallisto