Spotkanie 5

Filogeneza z odległości i drzewa w formacie Newick

Autorzy

Mateusz Strzelecki 188692

Paweł Cichowski 184465

Sekcja 1

Teoria

Jak rekonstruujemy historię ewolucyjną?

Drzewo filogenetyczne to drzewo ważone, które pokazuje, jak pewne gatunki lub geny rozgałęziały się w czasie. Budujemy je przez stworzenie macierzy odległości, gdzie każda komórka ma wartość oznaczającą jak każda para stworzeń różni się od pozostałych.

W idealnym przypadku dane spełniałyby zasadę addytywności metryki - odległości macierzy odpowiadałyby sumie wag krawędzi drzewa. Realne dane biologiczne są często nieuporządkowane i rzadko sumują się idealnie, przez co należy użyć heurystyk takich jak metoda najmniejszych kwadratów.

Wynik zazwyczaj zapisuje się w formacie Newick, w którym nawiasy i przecinki opisują strukturę drzewa i długości gałęzi.

Drzewa porównuje się, aby sprawdzić czy wyniki działania są stabilne i czy różne algorytmy i heurystyki dają zbliżony wynik.

Użycie AI w tej sekcji

AI pomogło uporządkować wątki (drzewo, addytywność, Newick, porównywanie drzew)

Sekcja 2

Pojęcia

Minisłowniczek

Drzewo filogenetyczne

Graf opisujący hipotezę pokrewieństwa: liście to obserwowane taksony, a węzły wewnętrzne to hipotetyczni przodkowie. W filogenezie jest to główny obiekt, z którego odczytujemy grupy siostrzane i kolejność rozdzielenia linii.

Macierz odległości

Kwadratowa macierz, w której element (i,j) to odległość między taksonami i oraz j, policzona z danych (np. z sekwencji). Jest wejściem dla metod dystansowych i pozwala traktować filogenezę jak problem rekonstrukcji z metryk.

Metryka addytywna

Własność odległości mówiąca, że istnieje drzewo z długościami gałęzi, które dokładnie odtwarza wszystkie dystanse między liśćmi jako sumy po ścieżkach. Jeśli macierz jest addytywna, rekonstrukcja drzewa z odległości ma jednoznaczniejsze i stabilniejsze rozwiązania.

Heurystyka dystansowa

Algorytm, który buduje drzewo przybliżone na podstawie macierzy odległości, bo pełne przeszukanie przestrzeni drzew jest obliczeniowo nieopłacalne. W praktyce heurystyki (NJ, UPGMA) są standardem, a ich jakość ocenia się błędem dopasowania do macierzy.

UPGMA

Metoda klasteryzacji hierarchicznej, która łączy najbliższe klastry i tworzy drzewo ultrametryczne (zakłada stałe tempo ewolucji, tzw. zegar molekularny). Daje proste w implementacji drzewa, ale może zniekształcać relacje, gdy tempa ewolucji różnią się między liniami.

Neighbor Joining (NJ)

Heurystyka budująca drzewo, która iteracyjnie wybiera parę taksonów minimalizującą kryterium oparte o skorygowane odległości. Często lepiej radzi sobie niż UPGMA, bo nie wymaga założenia ultrametryczności i zwykle daje dobre przybliżenie drzewa addytywnego.

Format Newick

Tekstowy zapis drzewa za pomocą zagnieżdżonych nawiasów: rozgałęzienia koduje struktura, a długości gałęzi dopisuje się po dwukropku. Pozwala łatwo przenosić drzewa między narzędziami i porównywać wyniki różnych rekonstrukcji.

Użycie AI w tej sekcji

AI wykorzystano do wygenerowania pierwszych wersji definicji, następnie hasła zostały przeczytane ze zrozumieniem i nieco przeredagowane

Sekcja 3

Python

Eksperymentacja w Pythonie - Budowa i analiza drzewa filogenetycznego

Opis danych

Użyłem 10 krótkich sekwencji DNA (56 bp) zapisanych w formacie FASTA. Zamiast pobierania z zewnętrznej bazy, wklejam je w kod jako tekst, aby eksperyment był w pełni odtwarzalny w jednym pliku. Odległości policzyłem jako p-distance (odsetek niedopasowanych pozycji), czyli prostą metrykę „na surowych znakach” bez modelu substytucji.

Dlaczego wybór NJ?

Wybrałem Neighbor Joining, bo nie zakłada ultrametryczności (stałego tempa ewolucji) jak UPGMA. W metodach dystansowych NJ często lepiej przybliża drzewo addytywne, gdy tempa zmian między liniami są nierówne.

Kod źródłowy

Dependencje: Python, Biopython, NumPy, Matplotlib

```
from io import StringIO
from Bio import SeqIO

records = list(SeqIO.parse(StringIO(fasta_text), "fasta"))
names = [r.id for r in records]
seqs  = [str(r.seq) for r in records]

# distance matrix
import numpy as np
n, L = len(seqs), len(seqs[0])
D = np.zeros((n, n), float)
for i in range(n):
    for j in range(i+1, n):
        mism = sum(a != b for a, b in zip(seqs[i], seqs[j]))
        D[i, j] = D[j, i] = mism / L

# neighbor joining
from Bio.Phylo.TreeConstruction import DistanceMatrix, DistanceTreeConstructor
lower = [[float(D[i, j]) for j in range(i+1)] for i in range(n)]
dm = DistanceMatrix(names, lower)
tree = DistanceTreeConstructor().nj(dm)

T = np.zeros((n, n), float)
for i, a in enumerate(names):
```

```
    for j, b in enumerate(names):
        if i < j:
            dist = tree.distance(a, b)
            T[i, j] = T[j, i] = dist

mse = float(np.mean((T - D)**2))
l2  = float(np.linalg.norm(T - D))
```

MSE - średni błąd kwadratowy

0.00013022

między macierzą D i dystansami w drzewie T

Błąd l^2

0.114112

globalna miara niezgodności T i D

Maksymalna różnica |T − D|

0.033460

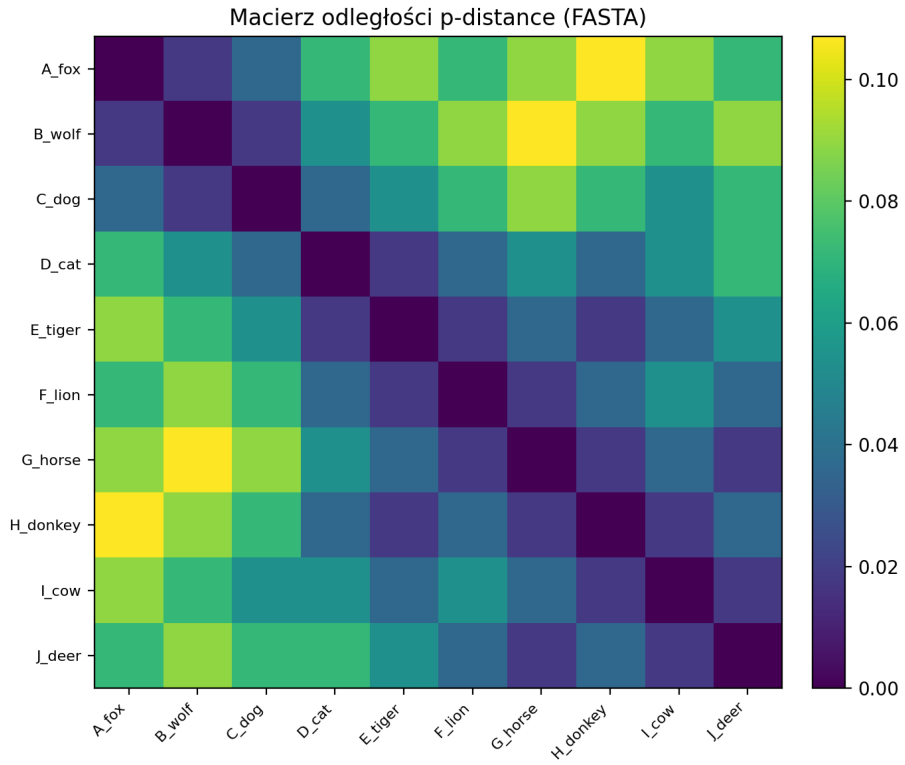
najgorsza para taksonów

Fragment macierzy odległości (p-distance)

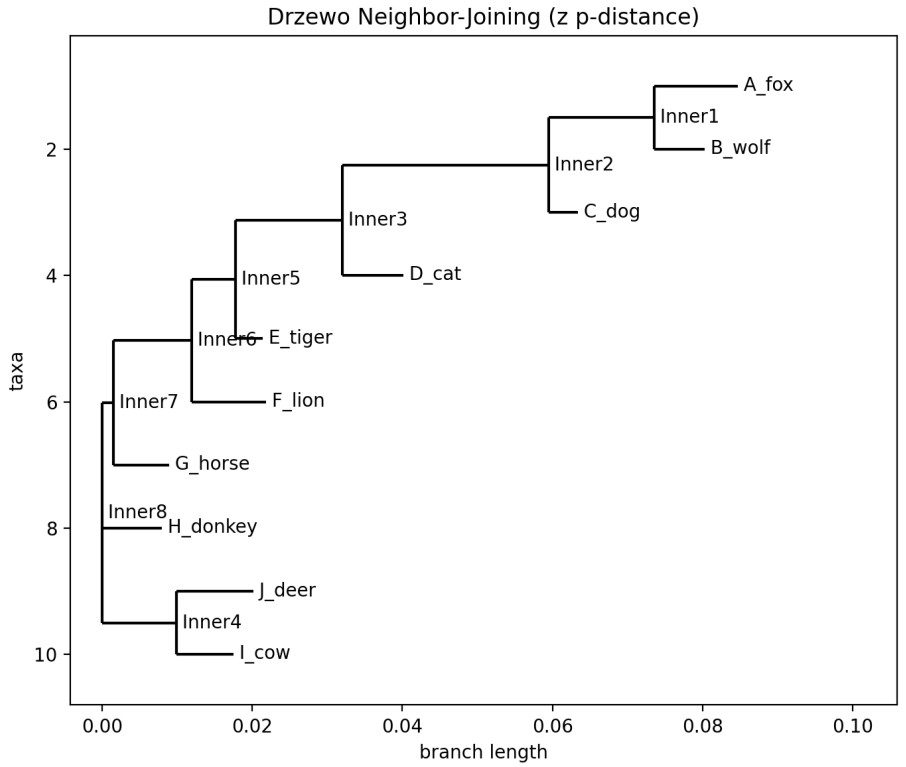
pierwsze 6×6, zaokrąglone do 3 miejsc

	A_fox	B_wolf	C_dog	D_cat	E_tiger	F_lion
A_fox	0.000	0.018	0.036	0.071	0.089	0.071
B_wolf	0.018	0.000	0.018	0.054	0.071	0.089
C_dog	0.036	0.018	0.000	0.036	0.054	0.071
D_cat	0.071	0.054	0.036	0.000	0.018	0.036
E_tiger	0.089	0.071	0.054	0.018	0.000	0.018
F_lion	0.071	0.089	0.071	0.036	0.018	0.000

Wykres 1: Heatmap macierzy odległości



Wykres 2: Drzewo NJ



Interpretacja

Otrzymane drzewo NJ grupuje sekwencje o najmniejszych p-distance jako sąsiadujące liście, co widać też na heatmapie (ciemniejsze „bloki" małych odległości). Wartości MSE i l2 są małe, ale nie zerowe, co oznacza, że drzewo nie odtwarza idealnie wszystkich odległości z macierzy. Największe rozbieżności pojawiają się, gdy prosta p-distance „gubi" informację o wielokrotnych podstawieniach lub gdy różne linie mają inne tempa zmian, więc jedna długość gałęzi nie pasuje naraz do wielu par.

Dlaczego błąd aproksymacji (l2) sugeruje nieaddytywność danych i jak to wpływa na wiarygodność rekonstrukcji historii ewolucyjnej?

Jeśli dane byłyby addytywne, istniałoby drzewo, dla którego T = D (błąd l2 równy zero), więc każda para liści miałaby dystans dokładnie jako sumę wag po ścieżce. Gdy l2 > 0, znaczy to, że nie da się jednocześnie dopasować wszystkich par odległości jedną strukturą drzewa i jednym zestawem długości gałęzi, więc część relacji musi być „kompromisem" algorytmu. Taki kompromis obniża wiarygodność wniosków o kolejności rozdzielania linii: niektóre rozgałęzienia mogą być artefaktem metryki lub szumu, a nie stabilnym sygnałem ewolucyjnym.

2–3 zdania: co rozumiem z kodu (anti black-box)

Wiem, że p-distance liczy tylko niedopasowania znaków w wyrównanych sekwencjach i nie modeluje wielokrotnych substytucji, więc może zaniżać „prawdziwe” odległości dla bardziej odległych par. Rozumiem też, że NJ konstruuje drzewo iteracyjnie, wybierając pary do połączenia na podstawie skorygowanego kryterium, a wynikowe długości gałęzi są potem używane do obliczania dystansów w drzewie (sumy po ścieżkach).

Użycie AI w tej sekcji

AI pomogło mi ułożyć strukturę eksperymentu (kolejność kroków, metryki MSE i l2) oraz wygenerować szkielet kodu do rekonstrukcji NJ i eksportu wykresów. Kod przejrzałem linia po linii, dopisałem komentarze i sprawdziłem, co oznaczają obliczane macierze D oraz T i dlaczego ich różnica jest miarą nieaddytywności.