Ewolucja sekwencji

Program zaliczeniowy nr 2 z przedmiotu *Algorytmy i Struktury Danych*

semestr letni 2018/2019

Wstęp

Elementarnymi operacjami edycyjnymi nazwiemy następujące modyfikacje tekstu:

- *substytucja* – zmiana pojedynczego symbolu na inny,
- *delecja* – usunięcie pojedynczego symbolu,
- *insercja* – wstawienie pojedynczego symbolu.

Odległość edycyjna pomiędzy dwoma tekstami to minimalna liczba elementarnych operacji edycyjnych potrzebnych do przekształcenia jednego tekstu w drugi. Odległość edycyjna bywa używana jako miara odległości ewolucyjnej pomiędzy homologicznymi sekwencjami biologicznymi, np. genami lub białkami.

Zadanie

Należy napisać moduł *ImieNazwisko.py* zawierający implementację następujących obiektów:

- `EditDistance(sequence1,sequence2)` – funkcja zwracająca odległość edycyjną sekwencji `sequence1` i `sequence2`.
- `PhylTree`, `PhylNode` – klasy implementujące drzewa filogenetyczne dla sekwencji biologicznych oraz ich wierzchołki etykietowane sekwencjami.

  Klasa `PhylNode` powinna zawierać następujące metody:

  - `__init__(distance=None, sequence=None, children=None)` – utwórz wierzchołek z dziećmi z listy `children` (domyślnie bez dzieci), etykietowany sekwencją `sequence`, odległy o `distance` od swojego rodzica w drzewie
  - `get_children()` – zwróć listę dzieci
  - `get_distance()` – zwróć odległość wierzchołka od rodzica (lub `None`, jeśli ta nie została ustalona)
  - `get_sequence()` – zwróć sekwencję wierzchołka, jeśli ją posiada; w przeciwnym razie zwróć `None`
  - `set_distance(distance)` – zmień odległość wierzchołka od jego rodzica na `distance`
  - `set_sequence(sequence)` – nadaj wierzchołkowi etykietę `sequence`

  Klasa `PhylTree` powinna zawierać następujące metody:

  - `__init__(node)` – utwórz drzewo z wierzchołkiem `node` w korzeniu
  - `root()` – zwróć korzeń drzewa
  - `distance_sum()` – zwróć sumę wszystkich odległości w drzewie (łącznie z odległością korzenia od jego rodzica)
  - `get_sequences()` – zwróć listę sekwencji ze wszystkich wierzchołków drzewa, uporządkowaną prefiksowo (najpierw dany wierzchołek, potem pierwszy syn z potomkami, drugi syn z potomkami itd.)
  - `calculate_distances(dist_function=EditDistance)` – nadaj wszystkim wierzchołkom odległości od rodziców wyznaczone za pomocą funkcji `dist_function` (domyślnie `EditDistance`; funkcję należy aplikować do etykiet wierzchołków, czyli sekwencji); korzeniowi należy nadać odległość 0 od rodzica
- `BuildTree(sequences, dist_function=EditDistance)` – funkcja zwracająca drzewo filogenetyczne dla sekwencji z listy `sequences` z odległościami wyznaczonymi za pomocą `dist_function`; drzewo powinno być tak skonstruowane, żeby suma wszystkich odległości była jak najmniejsza.

Pamięć potrzebna do analizy danych składających się z kilkunastu sekwencji o długości kilkadziesiąt symboli nie powinna prekraczać 0.5GB, a czas wykonania poszczególnych operacji na przeciętnym laptopie nie powinien przekraczać:

- 1 minuty dla `BuildTree`,
- 10 sekund dla `calculate_distances`,
- 1 sekundy dla pozostałych operacji.

Rozwiązanie zadania powinno zawierać kod programu z komentarzami.

Ocena

Za pełne rozwiązanie można otrzymać 12 pkt., w tym:

4 pkt. implementacja funkcji `EditDistance`

4 pkt. implementacja klas `PhylNode` i `PhylTree`

4 pkt. implementacja funkcji `BuildTree`