Statystyczne metody oznaczania niepewności w problemach klasteryzacji danych biologicznych

Joanna Całkiewicz

W przedstawionej pracy podjęłam próbę oceny możliwości zastosowania metod badania istotności statystycznej gałęzi wewnętrznych wykorzystywanych w analizie filogenetycznej do analizy drzew otrzymanych dla innego rodzaju danych biologicznych.

Testowanie gałęzi wewnętrznych jest ostatnim etapem analizy filogenetycznej. Wraz z rozwojem technik budowy drzew, w tym technik opartych na odległościach oraz technik związanych z przeszukiwaniem przestrzeni drzew, rozwijano oprogramowanie badające jakość otrzymywanych dendrogramów czyli kwestie wsparcia statystycznego. Część z nich opiera się na pseudopróbkowaniu z oryginalnego zestawu danych, a następnie ponownej klasteryzacji (metoda Felsensteina, metoda AU), część polega na zbadaniu z użyciem powtórzeń macierzy pseudoodległości, jak często rekonstruowane są drzewa z gałęziami o zerowej długości, które świadczą o możliwości innego grupowania (metoda Dopazo oraz WLS-LRT).

W innych dziedzinach biologii, takich jak ekologia, gdzie większość badań polega na oznaczaniu i zliczaniu gatunków w próbach środowiskowych lub jak biologia molekularna, gdzie wykorzystując płytkę mikromacierzową, możliwe jest określenie jednorazowo poziomów ekspresji wielu genów, do wyznaczania i obrazowania zależności podobieństwa między analizowanymi obiektami, również stosowane są drzewa. Otrzymywane są one z użyciem różnych technik klasteryzacji hierarchicznej. W analizach tych rzadko jednak dotyka się kwestii wsparcia dla gałęzi wewnętrznych w uzyskiwanych dendrogramach.

Wyniki przedstawione w rozprawie wskazują, że zastosowane przeze mnie metody oryginalnie wykorzystywane w analizie filogenetycznej mogą być również z powodzeniem wykorzystywane do badania istotności statystycznej gałęzi wewnętrznych dendrogramów - reprezentujących innego rodzaju dane biologiczne niż dane sekwencyjne - w szczególności testy oparte na analizie długości wewnętrznych gałęzi w drzewach.

Badania przeprowadziłam dla kilku typów danych. Były to zbiory danych ekspresji genów osób chorych na białaczkę szpikową lub limfoblastyczną (zbiór „*Golub*" - przedstawione w Rozdziale 5.3.2.), dwa zespoły danych pacjentów chorych na raka dwunastnicy (zbiory „*Alon*" i „*Notterman*" przedstawione odpowiednio w Rozdziałach 5.3.3 oraz 5.3.4) oraz cztery zespoły danych składu i liczebności flory i fauny rejonu Spitsbergenu

(zbiory: „*Kongsfjorden*", „*Outer*", „*Deep*", „*Marinok*" przedstawione w Rozdziałach od 5.2.1. do 5.2.4.).

W przypadku danych liczebności gatunków podjęłam próbę weryfikacji, czy obarczone najmniejszą niepewnością gałęzie w drzewach utworzonych dla powyższych danych odpowiadają istotnym podziałom badanych obiektów, poprzez konsultację z osobami bezpośrednio zaangażowanymi w projekty badawcze. W przypadku danych poziomów ekspresji genów jakość klasyfikacji wyznaczyłam za pomocą wskaźników: dokładności, czułości i specyficzności klasyfikacji. Dodatkowo, porównałam wyniki prostych klasyfikatorów - opartych na metodach analizy skupień i badaniu wsparcia dla gałęzi - do rezultatów, które osiągają bardziej złożone klasyfikatory.

Oprócz analiz na podstawie powyżej wymienionych danych: dostępnych publicznie danych mikromacierzowych dla pacjentów z chorobami onkologicznym oraz danych ekologicznych, również porównywałam metody badania wsparcia gałęzi w drzewach konstruowanych na podstawie danych sekwencyjnych - zbiorów „*DnaC*" oraz „*OxyR*" (przedstawione odpowiednio w Rozdziałach 5.1.2. i 5.1.3.), danych rearanżacji genomów bakteryjnych (przedstawione w Rozdziale 5.1.4.) oraz danych mikromacierzowych wykorzystanych do analizy liczebności transpozonów w genomach rybich – zbiór „*Tc1*" (przedstawione w Rozdziale 5.3.1.). Część wyników uzyskanych dla danych sekwencyjnych - w tym dla zbioru „*DnaC*" złożonego z bakteryjnych i wirusowych sekwencji o potencjalnej funkcji ładowaczy helikazy - została zamieszczona w pracy Słomiński i wsp. (2007), której jestem jednym z współautorów. Praca dotyczyła hipotezy o wirusowym pochodzeniu bakteryjnych ładowaczy helikazy, które należą do najważniejszych białek replikacyjnych bakterii *Escherichia coli*. W celu weryfikacji powyższej hipotezy przeprowadziła pełną analizę filogenetyczną, która obejmowała wyszukanie sekwencji, konstrukcję drzew oraz testy Felsensteina, Dopazo i WLS-LRT. Również uzyskane przeze mnie wyniki, związane z zespołem danych sekwencyjnych „*OxyR*" - sekwencje aminokwasowe białka represorowego fagów lambdoidalnych - zostały ujęte w pracy Glinkowska i wsp. (2010). Celem badań było sprawdzenie (w tym *in silico*), czy białka OxyR i CI wiążą się do tego samego rejonu regulatorowego fagów lambdoidalnych. W tym przypadku, między innymi, konstruowałam drzewa na podstawie wyszukanych sekwencji białka represora fagów oraz badałam istotność statystyczną gałęzi wewnętrznych uzyskanych drzew. Natomiast w pracy Wenne i wsp. (2011) zamieszczono wyniki, które uzyskałam dla zespołu danych „*Tc1*". Celem badań było sprawdzenie, czy technikę mikromacierzy można potraktować jako narzędzie do identyfikacji elementów repetytywnych oraz czy dane występowania i przybliżonej liczebności

transpozonów w genomach niosą informację, która pozwala na wyznaczenie pokrewieństwa między rybami.