

## Studium zdarzeń duplikacji w genomie.

Historię ewolucji można reprezentować za pomocą drzewa gatunków. Jest ono podobne w strukturze i przeznaczeniu do drzewa genealogicznego rodziny, stworzonego tylko dla jej członków płci męskiej o tym samym nazwisku. Aby uzyskać jego strukturę, można wykorzystać porównanie genomów poszczególnych gatunków (uliniowanie sekwencji złożonych z A,C,G,T).

Gen jest to fragment sekwencji genomu, zazwyczaj kodujący białko. W analogiczny sposób porównując sekwencje możemy podzielić geny na rodziny i dla każdej rodziny utworzyć drzewo. Założmy chwilowo, że dla dowolnych dwóch elementów, przez ich odległość w drzewie rozumiemy liczbę krawędzi pomiędzy nimi.

Hipotetycznie może się zdarzyć, że w drzewie gatunków dla człowieka, małpy i świni, odległość między genomem człowieka a genomem małpy jest mniejsza niż między genomami człowieka i świni. Z drugiej strony dla drzewa genów kodujących hemoglobinę, założmy, że genowi człowieka bliżej do genu świni, niż do genu małpy. W celu wyjaśnienia tych niezgodności stosuje się proces zwany uzgadnianiem.

Z uzgadnianiem związane jest zjawisko duplikacji genu i usytuowania jego wystąpień w ten sposób, aby było ich jak najmniej, oraz aby wyjaśniały różnice między drzewami genów a drzewem gatunków.

Problem EC to pogrupowanie tych duplikacji w jak najmniejszą liczbę miejsc na drzewie gatunków. W wyniku rozwiązania tego problemu otrzymujemy miejsca na drzewie gatunków, które utożsamiamy z epizodami duplikacji. Pojedynczy epizod zazwyczaj jest związany z wielokrotnymi zdarzeniami duplikacji, które indukują zbiór poddrzew w drzewie genów. Wysokość najwyższego poddrzewa determinuje liczbę zdarzeń w danym epizodzie. Problem ME to pogrupowanie duplikacji w jak najmniejszą liczbę zdarzeń. Pogrupowanie duplikacji w minimum miejsc nie zawsze daje minimum zdarzeń i odwrotnie.

Dla ustalenia uwagi rozważamy drzewa binarne. Drzewo ukorzenione posiada korzeń, dla drzewa gatunków reprezentuje on praprzodka, od którego pochodzą wszystkie gatunki wymienione w drzewie. Badając powiązania genów, bardziej naturalne jest przedstawienie ich jako drzewa nieukorzonego (bez wyróżnionego elementu, od którego pochodzą wszystkie inne). Wiele narzędzi do rekonstrukcji drzew z sekwencji zwraca wyniki w postaci nieukorzonych drzew genów. Problemy EC i ME mają liniowe rozwiązania dla wersji, w której drzewa genów są ukorzone. Dla nieukorzonej wersji problemu EC zaproponowaliśmy rozwiązanie. Natomiast kwestie złożoności problemu EC i problem ME w wersji dla nieukorzonych drzew genów to problemy otwarte, nie posiadające rozwiązań. Celem naszego projektu jest znalezienie rozwiązań tych problemów.

Dla nieukorzonej wersji problemu EC zaproponowaliśmy rozwiązanie w postaci algorytmu FPT sprowadzającego ten problem do instancji problemów EC w wersji ukorzonej. Artykuł z wynikami został zaakceptowany w czasopiśmie BMC Genomics. Wykorzystaliśmy teorie nieukorzonego uzgadniania i może to stanowić fundament do dalszych prac w ramach projektu.

Naszym celem jest rozwiązanie problemu otwartego ME w wersji nieukorzonej. Naszą strategią jest zbadanie klasy złożoności problemów EC i ME w wersji nieukorzonej. Planujemy opracowanie algorytmów i ich implementację. W wyniku tego projektu powstaną nowe narzędzia, które nie tylko będą w stanie generować lepsze wyniki, ale dodatkowo będą mogły być wykorzystane w przypadkach, których obecnie istniejące narzędzia nie potrafią rozwiązać. Nasze narzędzia będą niezależne od faktu, czy wejściowe drzewa genów są ukorzone, czy nie. W genomice porównawczej dane są przetwarzane czasem kolejno przez wiele różnych aplikacji. Planujemy, że nasze narzędzia, nie obciążone ograniczeniami na format wejścia, będą mogły współpracować ze wszystkimi istniejącymi aplikacjami. Ponadto, doświadczenia przeprowadzone na rzeczywistych danych biologicznych z wykorzystaniem implementacji naszego algorytmu FPT wykazały, że możemy poprawić znane wyniki dotyczące duplikacji genomowych dla rzeczywistych zbiorów danych, jeśli wykorzystamy rozwiązanie dla nieukorzonej wersji problemu EC.

Naszym celem jest również przeprowadzenie doświadczeń na rzeczywistych zbiorach danych biologicznych. Wierzmy, że mogą prowadzić one do nowych wyników w obszarach wielokrotnych duplikacji genów i duplikacji całego genomu.

Podsumowując, nasze badania wzbogacą dziedzinę genomiki porównawczej. Aby wymienić kilka praktycznych zastosowań tej dziedziny, wspomnę o tym, że drzewa gatunków można wykorzystać w całym spektrum zagadnień jak przykładowo w procesie wytwarzania pestycydów, przewidywania epidemii chorób zakaźnych czy badania dynamiki progresji raka, w celu dostosowania terapii.