ZESZYTY NAUKOWE WYDZIAŁU ELEKTRONIKI,
TELEKOMUNIKACJI I INFORMATYKI POLITECHNIKI GDAŃSKIEJ
Nr 9 Seria: ICT Young 2011

TRANSLACJA DRZEW FILOGENETYCZNYCH.

Piotr Płoński[1)], Jan Radomski[2)]

[1)] [1]Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Nowowiejska 15/19 00-665 Warszawa
[2)] Interdyscyplinarne Ctr. Modelowania Matematycznego i Komputerowego UW, Pawińskiego 5A, 02-106 Warszawa

Streszczenie

Skuteczną metodą wizualizacji zależności biologicznych pomiędzy organizmami jest przedstawianie ich za pomocą drzew filogenetycznych. Drzewa te pokazują sekwencje jako wierzchołki grafu, a zdarzenia mutacji jako krawędzie pomiędzy nimi. Istnieje wiele metod do rekonstrukcji drzew filogenetycznych, aczkolwiek większość z nich zakłada binarną postać budowanego drzewa, oraz przedstawia wszystkie analizowane sekwencje tylko wyłącznie jako liście. Założenia te w analizie organizmów szybko mutujących oraz organizmów, dla których wyizolowano dużo sekwencji mogą być nieprawdziwe. W pracy przedstawiono opis algorytmu translacji drzew filogenetycznych. Drzewa po translacji nie mają ograniczenia na binarną strukturę, a sekwencje są prezentowane na drzewie jako węzły wewnętrzne albo liście w zależności od ich charakteru rodzinnego. Algorytm do translacji wykorzystuje informację zawartą zarówno w topologii jak i długość gałęzi drzewa wejściowego. Wyjściowe drzewo umożliwia uzyskanie ścieżek ewolucyjnych w prosty sposób. Prezentowana metoda została przetestowane na sztucznych zbiorach generowanych za pomocą symulacji Monte Carlo, oraz na zbiorze rzeczywistym *Trypansoma cruzi*, dla którego prawdziwa topologia jest znana.

1. WSTĘP

Wraz z rozwojem technik komputerowych pojawiły się nowe możliwości analizy sekwencji genetycznych. Komputery o dużej pojemności pamięci pozwoliły na gromadzenie w bazach danych znacznych ilości informacji genetycznych. Wraz ze wzrostem ilości przechowywanych sekwencji wzrasta zapotrzebowanie na efektywne narzędzia umożliwiające analizowanie zależności między sekwencjami *in silico*. Najczęściej używanym sposobem reprezentacji zależności ewolucyjnych są drzewa filogenetyczne. Sposób ten został wprowadzony przez Karola Darwina, który jako pierwszy użył drzewa do przedstawienia ewolucji pomiędzy organizmami, do wyjaśnienia procesu doboru naturalnego [3].

Drzewo filogenetyczne to spójny graf acykliczny przedstawiający zależności ewolucyjne pomiędzy sekwencjami. Można je przyrównać do rodowodowego drzewa genealogicznego. Wierzchołki drzewa przedstawiają sekwencje, natomiast gałęzie opisują zdarzenia między sekwencjami. Zazwyczaj gałęzie mają przypisaną wagę, co odpowiada odległości genetycznej (liczbie mutacji) pomiędzy sekwencjami. Wierzchołki w drzewie

można podzielić na wierzchołki wewnętrzne, których stopień, czyli liczba przyległych krawędzi, jest większy od jednego, oraz wierzchołki końcowe – liście, które mają stopień równy jeden [5]. Ma to interpretację w genetyce, gdzie wierzchołki wewnętrzne przedstawiają sekwencje z potomstwem, a liście - sekwencje bez dzieci. Po wskazaniu w drzewie węzła, który jest korzeniem, czyli najstarszym wspólnym przodkiem wszystkich węzłów, można na drzewie przedstawić kierunek ewolucji. Wyznaczenie ścieżki ewolucyjnej w drzewie to przejście z liścia do korzenia drzewa z zapamiętaniem wszystkich mijanych wierzchołków wewnętrznych [2].

Konstrukcja prawdziwego drzewa filogenetycznego jest problemem NP-trudnym [4], dlatego do jego budowy używa się różnych metod heurystycznych. Można je generalnie podzielić na 4 grupy. Pierwszy sposób to nieparametryczna metoda Maximum Parsimony [6], która stara się zbudować drzewo tak by odległości między węzłami były jak najkrótsze. Metoda ta jest jednak czuła na różne współczynniki szybkości mutacji na gałęziach. Odporne na ten czynnik są algorytmy typu Maximum Likelihood (ML) [7]. Zakładają one z góry model zmian pomiędzy sekwencjami i budują zbiór różnych topologii drzew, które są oceniane na podstawie prawdopodobieństwa wystąpienia w rzeczywistości. Dużą wadą tej metody jest koszt obliczeniowy. Aby ominąć sprawdzanie wszystkich topologii stosuje się różne techniki heurystyczne [12]. Uogólnieniem podejścia ML jest metoda bayesowska [10]. Drzewo wynikowe jest całką ważoną prawdopodobieństwem wystąpienia drzewa po wszystkich sprawdzonych topologiach. Metody ML i bayesowskie mają duży koszt obliczeniowy i stosuje się je do analizy małych zbiorów sekwencji – do 30-50 sekwencji. Alternatywą dla nich są metody bazujące na macierzy odległości, które są szybkie. Metody te są to algorytmy klastrujące. Najpopularniejsze z nich to algorytmy Neighbor-Joining (NJ) [13] oraz UPGMA [15]. Spośród wymienionych metod, algorytm NJ można nazwać klasycznym w dziedzinie budowy drzew, ponieważ jest on jednym z najczęściej używanych algorytmów do budowy drzew [8], został dokładnie zbadany [1] i stał się de facto benchmarkiem w testowaniu nowych algorytmów [11].

Algorytmy budujące drzewa mają dużą skuteczność w rekonstrukcji prawdziwych topologii, wszystkie jednak wymienione algorytmy mają poważne ograniczenia jeżeli idzie o wynikową topologie budowanego drzewa. Zakładają one a priori iż drzewo rekonstruowane jest z sekwencji, które pojawić mogą się tylko na liściach. Co więcej przyjmują, że szukane drzewo jest drzewem binarnym. Założenia te upraszczają działanie algorytmów rekonstrukcji aczkolwiek wprowadzają ograniczenia na biologiczną interpretację drzewa.

Celem pracy jest przedstawienie algorytmu do translacji drzew filogenetycznych. Przetłumaczone drzewo zawiera wszystkie węzły poprawnie etykietowane i nie ma ograniczenia na stopień wierzchołków w grafie. Uzyskanie ścieżek ewolucyjnych z takiego drzewa jest proste.

2. WADY ALGORYTMÓW REKONSTRUKCJI DRZEW

Opis wad rekonstrukcji drzew zostanie przeprowadzony dla ustalenia uwagi na algorytmie NJ, aczkolwiek wady te tyczą się większości algorytmów. W trakcie działania algorytmu NJ, zawsze łączone są w parę dwa węzły, a z nich powstaje jeden nowy. Narzuca to ograniczenie na topologię rekonstruowanego drzewa – jest to topologia binarna. Można to interpretować jako możliwość rekonstrukcji tylko drzew gdzie przodek ma zawsze tylko dwóch potomków. Zmienia to biologiczny sens drzewa ponieważ natura nie ogranicza liczby potomków. Algorytm NJ radzi sobie z tym problemem używając

pomocniczych wierzchołków wewnętrznych, które używane są tylko po to by zapisać większą niż dwa liczbę przodków. Węzły pomocnicze można łatwo rozpoznać ponieważ odległość między węzłem pomocniczym a węzłem wewnętrznym dąży do zera. Używanie węzłów pomocniczych które nie mają znaczenia genetycznego stwarza problem przy wyznaczaniu ścieżek ewolucyjnych ze względu na niejednoznaczość w kolejności umieszczania węzłów pomocniczych. Kolejnym ograniczeniem topologicznym jaki powstaje w czasie konstrukcji drzewa za pomocą algorytmu NJ to etykietowanie zawsze tylko liści - niezależnie od tego czy sekwencja ma potomstwo w analizowanym zbiorze czy też nie. Rozpatrzmy przykład topologii prawdziwego drzewa podany na Rys.1b. zbudowanego dla zbioru z wyizolowanymi wszystkimi sekwencjami oraz zrekonstruowane drzewo pozyskane za pomocą algorytmu NJ Rys.1a. Na przykład sekwencja "3" ma dwoje potomstwa, sekwencje "9" i "5", aczkolwiek sekwencja "9" ma do pokonania mniejszą liczbę krawędzi do korzenia niż sekwencja "3". Co więcej sekwencja "3" mimo iż ma potomstwo, przedstawiana jest jako liść. Drzewo na Rys.1a. ma 18 wierzchołków chociaż zbudowane jest z 10 sekwencji. W przedstawionym syntetycznym przykładzie analizowany zbiór zawierał wszystkie sekwencje, dlatego liczba wierzchołków w drzewie powinna być równa liczbie sekwencji. W analizie rzeczywistych zbiorów sekwencji zazwyczaj obserwuje się jakąś część populacji, a brakujące sekwencje pokazywane są jako wierzchołki wewnętrzne bez etykiet.

3. OPIS ALGORYTMU QUICK PATHWAY FINDING (QPF)

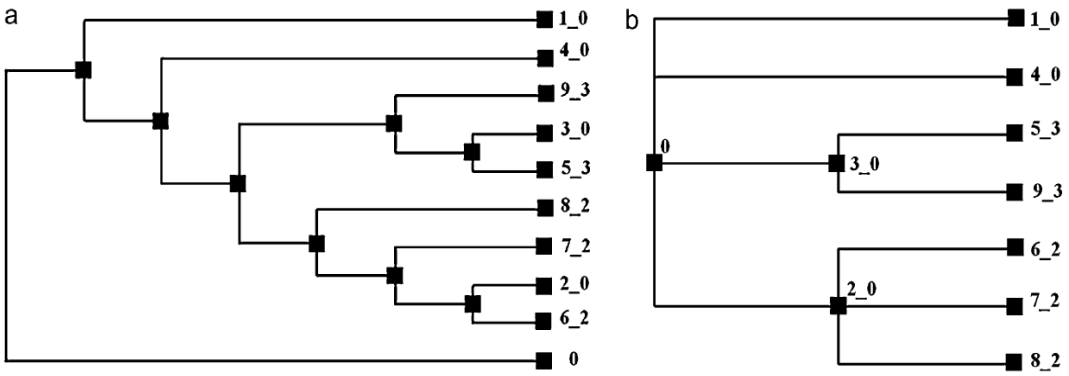
Algorytm QPF tłumaczy każdy wierzchołek drzewa zaczynając od wierzchołków najbardziej oddalonych od korzenia. W każdym wierzchołku sprawdzny jest charakter rodzinny wierzchołków które są przez niego trzymane, a do translacji dochodzi jeżeli wśród sprawdzanych jest jeden wierzchołek z charakterem rodzica. Decyzja o charakterze podejmowana jest na podstawie zakumulowanych długości krawędzi. Jako próg decyzyjny w algorytmie przyjmuje się połowę długości jednostkowej pomiedzy sekwencjami[1]. Poniżej przedstawiono opis algorytmu w krokach.

1. Wczytaj drzewo T=G(V,E), dla każdego wierzchołka v z V, wierzchołki które z niego wychodzą oznaczone jako C, wierzchołek który jest jego rodzicem oznaczony jako P, i długość krawędzi prowadzącą do rodzica D; Utwórz puste drzewo Q; Ustal wartość progu h;
2. Użyj przeszukiwania grafu wszerz (Breadth First Search) do wyznaczenia liczby krawędzi oznaczonej jako Ri między wierzchołkami grafu a korzeniem drzewa;
3. Posortuj wierzchołki grafu według malejącej liczby Ri, dodaj je to kolejki priorytetowej L;
4. Powtarzaj kroki 5, 6, 7 dopóki kolejka L nie jest pusta;
5. Weź z kolejki L wierzchołek v z największym Ri, usuń go z kolejki L;
6. Jeżeli wierzchołek v zawiera etykietę, dodaj wierzchołek q do drzewa Q z etykietą i krawędzią prowadzącą do rodzica taką samą jak w v;
7. Jeżeli v nie ma etykiety, dla każdego wierzchołka wychodzącego z niego ze zbioru C wyznacz charakter rodzinny, tak jak w kroku 8:

[1] Długość jednostkowa to odległość miedzy dwoma sekwencjami, które różnia się o jedną mutację.

a) jeżeli pośród wierzchołków ze zbioru C istnieje jeden z charakterem rodzica oznacz go jako A, dla wszystkich innych wierzchołków przypisz jako rodzica wierzchołek A; do krawędzi prowadzącej A dodaj krawędź prowadzącą z v;
b) jeżeli pośród wierzchołków ze zbioru C nie istnieje żaden z charakterem rodzica, przenieś wszystkie wierzchołki z C do zbioru wierzchołków C wierzchołka trzymającego P węzeł v; dla wszystkich przeniesionych wierzchołków zwiększ długość krawędzi prowadzącej do rodzica o wartość D węzła v;

8. Wyznaczanie charakteru rodzinnego wierzchołka v:
   a) jeżeli krawędź prowadząca jest mniejsza niż h, ustaw charakter wierzchołka jako rodzic;
   b) w przeciwnym wypadku ustaw charakter wierzchołka jako dziecko;



Rys. 1. Panel (a) – kladogram drzewa uzyskany za pomocą algorytmu NJ; panel (b) – kladogram drzewa prawdziwego. Oba drzewa przedstawiają zależności dla tego samego zbioru złożonego z 10 sekwencji. Każda etykieta sekwencji składa się z numeru sekwencji oraz po znaku '_' numeru sekwencji matki, korzeń drzewa oznaczony jest numerem "0" .

3. WYNIKI

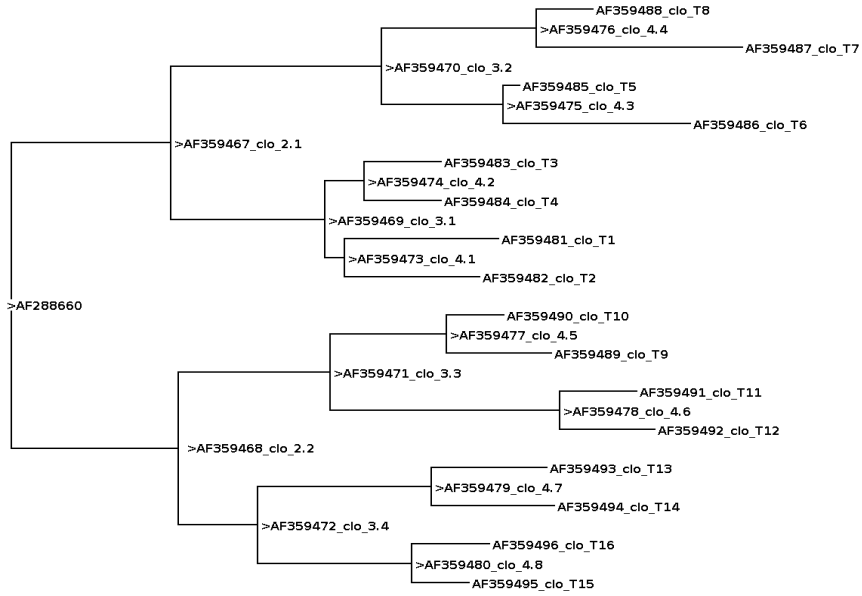
Aby przetestować algorytm, zostały wygenerowane zbiory sztucznych sekwencji za pomocą symulacji Monte Carlo. Wszystkie sekwencje zakodowane są za pomocą binarnych łancuchów. Mutacja pomiędzy dwoma sekwencjami zakłada zmianę z „0” na „1”, przy czym nie może być mutacji dwa razy na tej samej pozycji w całym zbiorze – dla takiego zbioru danych prawdziwe drzewo jest napewno rekonstruowalne [9]. Sekwencja korzeń zakodowana jest za pomocą samych „0”. Zostało wygenerowanych 8 zbiorów o wielkości N={10, 50, 100, 200, 500, 1000, 2000, 5000}. Dla wygenerowanych zbiorów zrekonstruowano drzewa za pomocą algorytmu NJ, a następnie przetłumaczono je za pomocą algorytmu QPF żeby uzyskać ścieżki ewolucyjne. Dla wszyskich wygenerowanych uzyskano 100% poprawnych ścieżek. W tablicy 1 przedstawiono czas działania algorytmu w zależności od liczby sekwencji przeprowadzony na komputerze z procesorem 1,5 GHz. Czas działania algorytmu jest w przybliżeniu liniowy i w porównaniu z czasem rekosntrukcj drzewa jest znikomy.

Tablica 1

Czas działania algorytmu

Liczba sekwencji	10	50	100	200	500	1000	2000	5000
Czas działania [sekundy]	0,0224	0,0185	0,0378	0,0615	0,1326	0,2523	0,5474	1,2466

Testy przeprowadzono również na zbiorze rzeczywistym *Trypansoma cruzi*, który został uzyskany w laboratorium, dzieki czemu jego prawdziwa topologia jest znana [14]. Przy użyciu wszystkich sekwencji zbudowano drzewo za pomocą algorytmu FastTree [12]. Następnie po doborze odpowiedniego progu dokonano translacji drzewa. Wynik zaprezentowano na Rys.2.



Rys.2. Drzewo

Wnioski do drzewla

4. ZAKOŃCZENIE

Algorytm QPF naprawia główne wady algorytmów rekonstrukcji drzew filogenetycznych, dając wynikowe drzewo o jasnej interpretacji biologicznej. Algorytm działa w czasie zbliżonych do liniowego. Z wynikowego drzewa można w prosty sposób zzyskać ścieżeki ewolucyjne, przechodząc z liścia do korzenia i zapamiętajuąc etykiety mijanych wierzchołków. Algorytm może działać z dowolną metodą rekonstrukcji drzew filogenetycznych, która buduje drzewo ewolucyjne. Jakość odczytanych ścieżek zależy od

jakości wejściowego drzewa. Autorzy pracują nad metodą rekonstrukcji drzewa produkującą odrazu poprawną topologię drzewa.

BIBLIOGRAFIA

[1] Atteson, K., 1999. The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25, 251–278.
[2] Baum, D., 2008. Reading a phylogenetic tree: the meaning of monophyletic groups. Nat. Educ. 1, http://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956.
[3] Darwin C., Carroll, J.T., 2003. On the origin of species by means of natural selection. Peterborough, Ont.: Broadview Press, 2003. ISBN 1-55111-337-6.
[4] Day, W.H.E., 1986. Computational complexity of inferring phylogenies from dissimilarity matrices. Bulletin of Mathematical Biology 49, 461-7.
[5] Deo, N., 1974. Graph Theory with Applications to Engineering and Computer Science. Prentice-Hall, Inc.
[6] Felsenstein, J., 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.
[7] Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.
[8] Gascuel, O., Steel, M., 2006. Neighbor-Joining Revealed. Mol. Biol. Evol. 23 (11), 1997-2000. doi: 10.1093/molbev/msl072
[9] Gusfield, D., 1991. Efficient algorithms for inferring evolutionary trees. Networks 21, 19–28.
[10] Larget, B., Simon, D.L., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16, 750–759.
[11] Nakhleh, L., Roshan, U., St. John, K., Sun, J., Warnow, T., 2001. Designing Fast Converging Phylogenetic Methods Bioinformatics 1, 1-9.
[12] Price, M.N., Dehal, P.S., Arkin, A.P., 2009. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. Mol. Biol. Evol. 26, 1641-1650.
[13] Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
[14] Sanson, G.F.O., Kawashita, S.Y., Brunstein, A., Briones, M.R.S., 2002. Experimental Phylogeny of Neutrally Evolving DNA Sequences Generated by a Bifurcate Series of Nested Polymerase Chain Reactions. Mol. Biol. Evol. 19, 170–178.
[15] Sokal, R., Michener, C., 1958. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, 38, 1409-1438.