ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2000
Seria: AUTOMATYKA z. 132 Nr kol. 1477

Zbigniew STAROSOLSKI

REKONSTRUKCJA DRZEW GENEALOGICZNYCH W PRZYPADKU ZMIENNEJ W CZASIE LICZEBNOŚCI POPULACJI

Streszczenie. W pracy porównano wyniki rekonstrukcji drzew genealogicznych uzyskane dwiema metodami: największej oszczędności oraz największej wiarygodności, przy dodatkowym założeniu, iż drzewo, będące wzorcem dla drzew rekonstruowanych, powstało w warunkach zmiany liczebności populacji w czasie. Przyjęto trzy typy przebiegów liczebności populacji: dwustanowy, wielostanowy, wykładniczy. Przy tworzeniu drzewa wzorcowego została zastosowana funkcja gęstości prawdopodobieństwa procesu koalescencji.

RECONSTRUCTION OF THE GENEALOGICAL TREES IN CASE OF NOT CONSTANT POPULATION SIZE

Summary. In this work we compare the results of reconstruction of genealogical trees using two methods: maximum parsimony and maximum likelihood. An additional condition on pattern tree for trees reconstructed, were built in case of varying population size. Three types of courses population size: two-step, multiple-step, exponential were tested. For generation of the true tree probability function the process coalescence is used.

Wstęp

Zagadnienie odtwarzania (oceny) topologii oraz metryki drzew genealogicznych na podstawie obserwowanych różnic pomiędzy osobnikami (losowanymi) ze studiowanej populacji ma duże znaczenie w takich dziedzinach, jak paleontologia, antropologia, genetyka, dynamika populacyjna, taksonomia. Problem ten ma także znaczenie w teorii estymacji, statystyce oraz teorii algorytmów z uwagi na wykorzystywane przy jego rozwiązywaniu metody i modele matematyczne.

Do dnia dzisiejszego dla sekwencji DNA opracowano wiele metod służących do rekonstrukcji drzew genealogicznych. Wielkością wejściową tych metod jest zbiór sekwencji kodu, np. róż-

nych gatunków lub przedstawicieli ras ludzkich. Liczba tych sekwencji odpowiada liczbie gatunków lub ras. Długość sekwencji jest uzależniona od długości posiadanych zasobów DNA badanych osobników [1-6],[15].

W niniejszej pracy rozważane jest zagadnienie odtwarzania genealogii, przy założeniu że obserwuje się różnice pomiędzy sekwencjami kwasu DNA u losowanych z populacji osobników. Spośród wielu opracowanych metod najistotniejsze znaczenie praktyczne mają tylko trzy, które są poniżej krótko omówione w kolejności ich wzrastającej złożoności obliczeniowej.

Metoda łączenia sąsiadów (neighbor joining method). Metoda ta identyfikuje najmniej różniące się pary w populacji w ten sposób, aby minimalizować całkowitą długość gałęzi drzewa. Pierwszym krokiem jest utworzenie formacji gwiazdy, na ramionach której umieszcza się sekwencje wejściowe. Kolejnym krokiem jest policzenie różnic pomiędzy wszystkimi sekwencjami i wybór tych dwóch, które dają najmniejsze różnice. Następnie wybrane dwie sekwencje łączy się w jeden węzeł i powtarza procedurę aż do otrzymania trzech węzłów końcowych, z których jest możliwe utworzenie tylko jednego drzewa wolnego [14].

Metoda największej oszczędności (maksimum parsimony method). Metody te testują prawdopodobne topologie drzew pod kątem dopasowania do drzewa różnic pomiędzy sekwencjami wejściowymi. Wybrane zostają te drzewa, które minimalizują liczbę niezbędnych zmian w tych sekwencjach. Metoda wykorzystuje algorytm heurystyczny, na podstawie którego dopasowuje do drzewa przebieg mutacji [1,2],[7].

Metoda największej wiarygodności (maximum likelihood method) wykorzystuje modele statystyczne zamian nukleotydów, użytkuje informację o częstotliwościach występowania nukleotydów w danych sekwencjach wejściowych. Metoda ta estymuje stały współczynnik zdarzenia mutacji. Na podstawie modelu i zgromadzonych informacji dopasowuje do danych sekwencji przebieg mutacji, budując drzewo genealogiczne [5].

Z uwagi na złożoność zagadnienia metody rekonstrukcji genealogii opierane są na modelach statystycznych, opisie słownym lub na specjalnych metrykach. Istotne jest zatem porównywanie wyników uzyskiwanych różnymi metodami.

W literaturze można znaleźć wiele przykładów takich porównań [1-2,8-10,16-17]. Zakładano, że liczebność analizowanej populacji w toku ewolucji nie ulegała zmianie. Założenie to jest oczywiście tylko grubym przybliżeniem rzeczywistości; wiadomo, że większość populacji zmienia swą liczebność z upływem czasu. Częstotliwość procesu mutacji także nie jest wielkością

stałą, lecz podlega zmianom. Kunher, zakładając zmienną częstotliwość procesu mutacji, wykazała dla drzewa o stosunkowo małej liczbie osobników wyższość metody największej wiarygodności nad pozostałymi metodami. Podobne wyniki otrzymał Yang badając grupę 10 osobników. Yang za pomocą metody największej wiarygodności otrzymał lepsze wyniki rekonstrukcji topologii drzewa. Jednak dla stałej częstotliwości procesu wyniki otrzymane metodą największej oszczędności były lepsze.

Problem wpływu zmiany liczebności populacji w czasie nie był rozważany w literaturze. Dlatego interesujące wydaje się przebadanie metod odtwarzania genealogii w warunkach zmiany liczebności populacji w czasie. Sprawdzano trzy typy przebiegów liczebności w czasie: dwustanowy, wielostanowy i wykładniczy. Zmiana liczebności populacji w czasie jest zawarta w modelu matematycznym określonym w dalszej części pracy. Opracowano metodę symulacyjną generowania topologii i metryki drzew genealogicznych i dalej na ich podstawie sekwencji pseudokodów DNA (pod terminem pseudokod DNA rozumie się losowo wygenerowany ciąg nukleotydów, które nie mają odniesienia w rzeczywistych sekwencjach obserwowanych w DNA). Pseudokody służą do testowania algorytmów odtwarzania genealogii. Wykorzystując dane otrzymane za pomocą opracowanego algorytmu porównano dwie metody: największej oszczędności i największej wiarygodności.

Dla metody oszczędności do rekonstrukcji drzewa genealogicznego na podstawie pseudokodu mtDNA użyto programu DNApars, a dla metody największej wiarygodności- program DNAmlk. Oba te programy wchodzą w skład dostępnego publicznie w sieci Internet pakietu PHYLIP służącego do analizy filogenetycznej.

W celu określenia miary zgodności drzewa rekonstruowanego z drzewem wzorcowym zmodyfikowano metodę porównywania Robinsona [12].Wyniki z tak przeprowadzonych symulacji przedstawiono na zbiorczych wykresach (rys.3-10).

Generacja drzewa genealogicznego wzorcowego – reprezentacja, podstawowe procesy

Procedura generacji drzewa wzorcowego składa się z kolejnych etapów:

I. Topologia drzewa – losowanie próbki

Poniżej zostanie omówiony sposób budowy drzewa genealogicznego na podstawie łańcucha kodu genetycznego hipotetycznego wspólnego przodka, czyli sekwencji przypisanej do korzenia drzewa genealogicznego.

Na starcie procedury mamy zbiór *K*-elementowy reprezentujący próbkę wybraną spośród populacji osobników żyjących obecnie. W następnym etapie generowania pseudokodów każdemu elementowi tego zbioru zostanie przypisana odpowiednia sekwencja pseudokodu DNA. Występuje tutaj założenie dotyczące łączenia z sobą elementów zbioru: elementy łączymy w pary w sposób przypadkowy - jedna para w każdym kroku procedury. Wylosowany *j*-ty element spośród *K-1* elementów łączy się z elementem o indeksie *j-1*.W ten sposób po *K-1* krokach powstaje *K-1* połączeń. Połączenia uzyskane w powyższy sposób reprezentują węzły budowanego drzewa.

II. Matematyczny model wystąpienia mutacji

W trakcie ewolucji kod genetyczny ulegał zmianom (mutacjom) na skutek działania różnego rodzaju przyczyn, takich jak niekorzystne czynniki środowiska, czy też niewłaściwa replikacja. Mutacje te są z pokolenia na pokolenie następne. Zjawisko to podlega jednemu z procesów Markowa. Jest to proces Poissona, który ma przyrosty niezależne. Proces ten określa prawdopodobieństwo k zdarzeń mutacji na przedziale czasu o długości t jako zależność [4]:

$$p(k;\lambda t)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}, \tag{1}$$

gdzie $\lambda>0$ jest to częstotliwość procesu .

Wielkość λ jest dobierana doświadczalnie, w wypadku ewolucji częstotliwość procesu mutacji została ona określona na $1*10^6$.

III. Procedura generowania czasu połączenia sekwencji wejściowych

Model matematyczny wykorzystany w celu generowania czasu połączeń powstał przy następujących założeniach [7]:

1. Populacja jest diploidalna, w chwili t_0 ma liczność N_0. Czas t_0 oznacza chwilę rozpoczęcia procedury.

2. Wielkość populacji zmienia się w czasie i w chwili τ wynosi N_τ.

Prawdopodobieństwo zdarzenia, że wspólny przodek losowo wybranej pary ze zbioru *K* osobników znajdował się w generacji odległej od chwili obecnej o czas τ (czas liczymy wstecz), przedstawia wzór (2)

$$P_\tau = \frac{\binom{K}{2}}{2N_\tau} \prod_{\sigma=0}^{\tau-1} \left[1 - \frac{\binom{K}{2}}{2N_\sigma} \right] \tag{2}$$

Zakładając, że rozmiar populacji jest wystarczająco duży, możemy przejść z dyskretnej skali czasu do ciągłej. Po tych założeniach wzór (2) przekształcamy do (3):

$$p(\tau) = \frac{\binom{K}{2}}{2N(\tau)} e^{-\int_0^\tau \frac{\binom{K}{2}}{2N(\sigma)} d\sigma} \tag{3}$$

Funkcja $p(\tau)$ jest gęstością prawdopodobieństwa procesu koalescencji [11].

We wzorze (3) czas rozpisano na przedziały $\tau \in \langle \tau_i, \tau_{i+1} \rangle$, co prowadzi do następującego wzoru:

$$P(\tau_n) = \frac{\binom{K}{2}}{2N(\tau_1)} \cdot e^{-\binom{K}{2}(sum_{n-1})} e^{-\frac{\binom{K}{2}(\tau-\tau_{n-1})}{2N(\tau_n)}}, \tag{4}$$

gdzie: $sum_{n-1} = \frac{\tau_1}{2N(\tau_1)} + \frac{\tau_1 - \tau_2}{2N(\tau_2)} + \cdots + \frac{\tau_n - \tau_{n-1}}{2N(\tau_{n-1})}$.

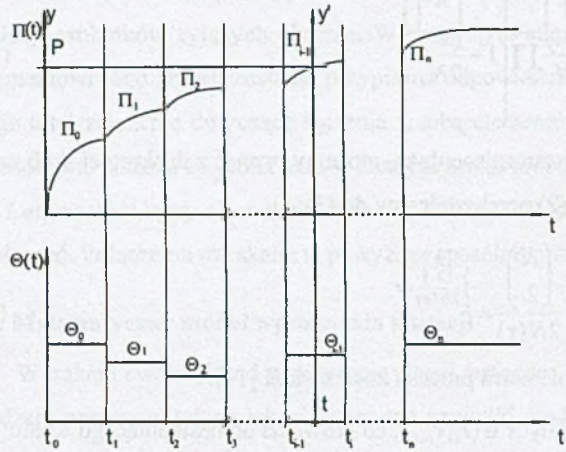
Ze wzoru (4) po krótkich przekształceniach określamy czas τ:

$$\tau = \tau_{n-1} - 2N(\tau_n) \left(sum_{n-1} - \frac{1}{\binom{K}{2}} \ln \frac{\binom{K}{2}}{2N(\tau_n) \cdot P(\tau_n)} \right). \tag{5}$$

W przypadku badań ewolucji przedstawienie czasu τ jest niewygodne w normalnej skali czasu, dlatego wzór (5) najczęściej przekształca się do skali mutacyjnej. W tym celu do wzoru (3) podstawiamy $P\left(\frac{t}{2\lambda}\right) = \Pi(t)$, $N(t) = \frac{1}{4\lambda}\Theta(t)$ i otrzymujemy:

$$t = t_{n-1} - \Theta(t_n) \left(sum_{n-1} - \frac{1}{\binom{K}{2}} \ln \frac{2\lambda \binom{K}{2}}{\Pi(t_n) \cdot \Theta(t_n)} \right), \tag{6}$$

gdzie: $sum_{n-1} = \frac{t_1}{\Theta(t_1)} + \frac{t_1 - t_2}{\Theta(t_2)} + \cdots + \frac{t_n - t_{n-1}}{\Theta(t_{n-1})}$.



Rys.1. Przebieg przykładowych funkcji Π(t),Θ(t) oraz opis wyznaczania czasu

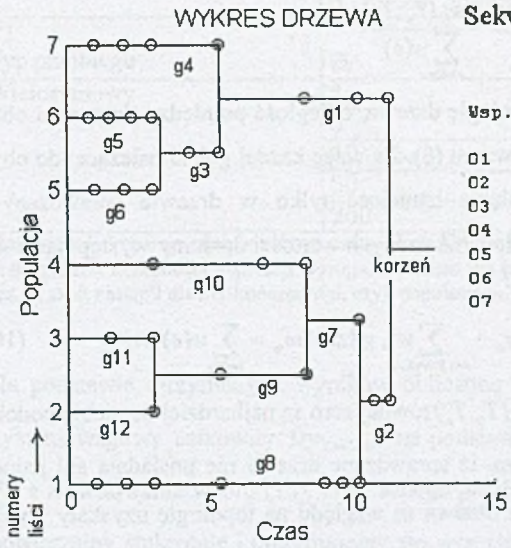
Fig.1. Course of example of function Π(t),Θ(t) and discription of determinimg time

Wzór (6) jest podstawowym wzorem, na podstawie którego będzie się generować czas połączeń. Występują w nim dwie funkcje Θ(t) oraz Π(t), których wartości liczbowe muszą być dane. Pierwszą z nich jest zadany przebieg liczebności populacji w czasie. Wartości drugiej funkcji wyznacza się w sposób opisany w dalszej części artykułu. Przyjęto trzy typy przebiegu liczebności populacji Θ(t): wykładniczy, dwustanowy, wielostanowy (dowolnego kształtu i złożony z dowolnej liczby elementów). Wyznaczenie czasu połączenia, czyli czasu znalezienia wspólnego przodka dla dwóch osobników wybranych losowo z próbki, może sprawiać trudności obliczeniowe. Algorytm został tak napisany, aby można było wyznaczyć nieznany czas połączenia. Funkcje Π(t) i Θ(t) są zależne od czasu. Aby wyjaśnić, jak w tym przypadku jest wyznaczany czas połączenia, warto rozpatrzyć przykład, gdy funkcja Θ(t) jest wielostanowa. Opis pokazano na przykładzie z rys.1. Losujemy liczbę z przedziału (0,1); nazwijmy ją P. Wartość wylosowanej liczby będzie następnie określać prawdopodobieństwo tego, że osobnik populacji z $\Theta(t_0)$ ma wspólnego przodka, oddalonego o czas t, a czas ten jest nieznany. Dla danych czasów $t_0, t_1, t_2, .., t_{i-1}, t_i, ..., t_n$ obliczamy wartości funkcji Π(t) ze wzoru (6). Wartości t_i są to czasy, w których funkcja Θ(t) zmienia skokowo swoją wartość. Korzystając z własności funkcji rosnącej, jaką jest Π(t), można wyznaczyć, w którym przedziale (t_{i-1}, t_i) leży argument funkcji Π(t). Mając w ten sposób wyznaczoną wartość czasu t_{i-1}, można otrzymać szukane rozwiązanie, gdzie jedyną niewiadomą jest t. Następnym krokiem procedury jest przeliczenie wartości $t_0, t_1, t_2, ..., t_{i-1}, , t_i, ..., t_n$. Można to sobie wyobrazić jako przesunięcie osi y o wektor $-t$, wskutek czego otrzymujemy oś y'. Przedstawione zostało to na rys.1. Wynikiem tej operacji są nowe wartości: $t_0' = t$, $t_1' = t_1 - t$, $t_2' = t_{i+1} - t$ itd. Procedurę powtarzamy n-1-krotne, gdzie n oznacza liczbę liści generowanego drzewa. Otrzymane czasy połączeń określają długości gałęzi w generowanym drzewie genealogicznym wzorcowym.

IV. Wpływ topologii drzewa na zmiany w kodzie genetycznym

Przedstawiono powyżej, w jaki sposób otrzymać topologię drzewa wzorcowego. Następnie należy wygenerować pseudokod DNA zgodnie z topologią wcześniej zbudowanego drzewa wzorcowego.

Opis procedury zilustrowano na rys.2. Generujemy pseudokod o długości zadanej d zgodnie z modelem substytucji HKY85 [8]. Następnie zapisujemy wygenerowanym kodem macierz D o wymiarach d x K, gdzie K liczebność próbki. Po zakończeniu generacji macierz D będzie zawierała K sekwencji kodu. Przyjmujemy, że każdy wiersz zawierający kod jest przypisany do jednego liścia. Rozpatrując topologię postępujemy od korzenia w stronę liści. Każda mutacja zaznaczona na rys.2 jako kółko oznacza zamianę nukleotydu na inny nukleotyd lub na ten sam. Miejsce zamiany jest każdorazowo losowane spośród *d* możliwości, gdzie *d* długość sekwencji wejściowej. Gałąź g1 ma trzy kółeczka, oznacza to, że w sekwencjach przypisanych do liści o numerach 5,6,7 zostaną zapisane trzy jednakowe zamiany. Dwie mutacje w gałęzi g2 spowodują zapisanie dwóch zamian w sekwencjach liści 1,2,3,4, następne dwie mutacje w g3 spowodują zapisanie kolejnych dwóch zamian w sekwencjach liści 5,6, cztery mutacje w g4 zostaną zapisane jako cztery zamiany w sekwencji liścia 7. Procedurę tę powtarzamy tak długo, aż przedstawimy wszystkie mutacje jako zamiany w sekwencjach liści.



Sekwencje wygenerowane przez program

```
Wsp.przodek AGGACGCGUGGGAUCAAGUAGAUACGUCGU

01          AGUACACGUGAGAUAAAGUCGAUGCAUUCU
02          AGUACGCGUAGGAUCAACUAGAUACGCACU
03          AGUACGCGUGGGAUCAAGUAGAUACCCACU
04          AGUACCCGGAGGAUCAAGUAGAUUCGUCCU
05          CCAACGCGUGGGAUCUACUAGAUUCGUCGU
06          CCGACCCGUGGGAUCCAGUAGAUACGUCAG
07          ACGAUGCGUGGGAUCAGAUAGACACGUCGU
```

Rys.2. Generacja pseudokodu DNA dla d=30, K=7
Fig.2. Generation output-code DNA for d=30, K=7

Sposób porównywania drzew

W celu porównania drzew rekonstruowanych z drzewem wzorcowym użyto metody zgodnej z ideą Robinsona [9],[10], lecz dostosowaną do potrzeb jej zastosowania. Miara odległości pomiędzy drzewami T_w (wzorcowym) i T_s (sprawdzanym- zrekonstruowanym), czyli dystans wagowy całkowity Dw_C, wyrażona jest wzorem:

$$Dw_C(T_w,T_s) = \sum_{e\in(Ew\cap Es)} |w_w(e) - w_s(e)| + \sum_{e\in Ew\setminus(Ew\cap Es)} w(e) + \sum_{e\in Es\setminus(Ew\cap Es)} w(e) \tag{8}$$

Wzór ten jest sumą trzech składników. Pierwszy z nich jest sumą modułów różnic wag (mutacji) dla tych samych gałęzi w obu drzewach. Drugi składnik jest to suma wag (mutacji) gałęzi nie występujących w drzewie badanym, a istniejących we wzorcowym. Trzeci składnik jest to suma wag (mutacji) gałęzi występujących w drzewie badanym, a nie istniejących we wzorcowym. Drzewo T_s jest identyczne z drzewem T_w, jeżeli dystans wagowy całkowity $Dw_C(T_w,T_s)$ równy jest 0.

Suma mutacji w drzewie wzorcowym nie jest stała, zależy od wartości parametrów przebiegów liczebności populacji w czasie. Dla dużej liczebności populacji liczba mutacji jest większa od liczby mutacji dla małej liczebności populacji. W celu przedstawienia miary uniezależnionej od całkowitej liczby mutacji w drzewie wzorcowym wzór (8) unormowano:

$$Dw_{CNORM}(T_w,T_s) = \frac{Dw_C(T_w,T_s)}{\sum_{e\in Ew} w(e)} \tag{9}$$

Dla przypadku gdy rozpatrujemy jedynie topologię drzewa, odległość pomiędzy drzewami obliczamy ze wzoru (10), który jest modyfikacją wzoru (8). Za wagę każdej gałęzi należącej do obydwu drzew przyjmujemy 1, natomiast gałęzie istniejące tylko w drzewie sprawdzanym i tylko w drzewie wzorcowym nie powodują dodania żadnych wartości do sumy występującej we wzorze (10):

$$D_t(T_w,T_s) = Sm_w - \sum_{e\in(Ew\cap Es)} w \text{ , gdzie } Sm_w = \sum_{e\in Ew} w(e) \text{ ,w=1} \tag{10}$$

Zgodnie z tą koncepcją, drzewa o wartości $D_t(T_w,T_s)$ równej zero są najbardziej do siebie podobne. Dla drzew o N liściach D_t=2N-3 oznacza, iż sprawdzane drzewa nie posiadają ani jednej jednakowej gałęzi. Oznacza to, że rozważane drzewa ze względu na topologię uzyskały wynik minimalny.

Dobór parametrów symulacji

Sprawdzono trzy typy przebiegów liczebności populacji w czasie. Należy pamiętać, iż w trakcie generacji drzewa wzorcowego czas był liczony do tyłu. Wartości parametrów przedstawiono na zbiorczej tablicy1:

Tablica 1

Parametry modelowanych przebiegów funkcji liczebności populacji w czasie

Typ przebiegu	Θ_i	γ lub t_i *)
Wykładniczy $\Theta(t) = \Theta_0 e^{-\gamma t}$	200	γ : 1; 0.5; 0.25; 0.125; 0.0625
	50	
Dwustanowy $\Theta(t) = \begin{cases} \Theta_1 & dla \quad t < t_1 \\ \Theta_2 & dla \quad t > t_1 \end{cases}$	200, 100	t_1=1; 5; 10; 20; ∞
	200, 50	
	200, 10	
	50, 10	
Wielostanowy $\Theta(t) = \begin{cases} \Theta_1 & dla \quad t < t_1 \\ \Theta_2 & dla \quad t_2 > t > t_1 \\ \vdots \\ \Theta_n & dla \quad t_n > t > t_{n-1} \end{cases}$	200, 50,10	t_i:1, 2.5, 5, ∞
	50, 10, 5	
	10, 50, 10	
	10, 200, 10	
	200, 180, 150, 90, 50, 10	t_i: 1, 2.5, 5, 6.25, 7, 10, ∞
	50, 45, 35, 20, 10, 5	
	10, 50, 90, 150, 190, 200	
	5, 10, 20, 35, 45, 50	

Typ przebiegu	Θ_i	γ lub t_i *)
Wielostanowy	1	$t_1=\infty$
	5	
	10	
	50	
	200	

*)oznaczenie t_i opisuje czas, w którym nastąpiła skokowa i-ta zamiana liczebności populacji, tak więc t_2=5 oznacza, iż drugi skok liczebności populacji wystąpił dla czasu t=5 (czas liczony do tyłu), nietypowe oznaczenie $t_1=\infty$ oznacza, iż skok nastąpił dla nieskończoności, czyli populacja miała stały przebieg .

Na podstawie otrzymanych wyników obliczono wielkości: dystans wagowy całkowity Dw_C, dystans wagowy całkowity Dw_{CNORM}, na podstawie wzorów (8), (9) oraz dystans topologiczny D_t z zastosowania wzoru (10). Dla każdego przebiegu liczebności populacji w czasie symulację powtarzamy stukrotnie i otrzymujemy sto wartości Dw_C, Dw_{CNORM}, D_t dla każdej badanej me-

tody. Końcowym wynikiem dla jednego zadanego przebiegu są ich wartości średnie otrzymane na podstawie stu symulacji.

Porównanie wyników metod: największej oszczędności i największej wiarygodności

Porównano dwie metody rekonstrukcji drzew genealogicznych: metodę oszczędności i metodę największej wiarygodności. Sprawdzono wpływ trzech typów przebiegów liczebności populacji w czasie na jakość estymacji drzew genealogicznych: wykładniczego, dwustanowego i wielostanowego. Na wykresach przyjęto, iż etykiety z oznaczeniem NO dotyczą wyników otrzymanych metodą największej oszczędności oraz z oznaczeniem NW metodą największej wiarygodności.

I. Przebieg wykładniczy

Zebrane wyniki, dla przebiegu wykładniczego liczebności populacji w czasie, zobrazowano wykresami. Rysunek 3 przedstawia zależność Dw_C od wielkości gamma. Z jego przebiegu wynika, iż przy tych samych wartościach zadanego przebiegu populacji w czasie, metodą największej wiarygodności otrzymano dokładniejsze wyniki. Dla $\Theta_0 = 50$ jest widoczna wyraźna różnica pomiędzy krzywymi uzyskanymi metodą największej wiarygodności a metodą największej oszczędności. Natomiast dla $\Theta_0 = 200$ różnica ta nie jest już tak znaczna. Ponadto wyniki uzyskane metodą największej wiarygodności są obarczone mniejszym odchyleniem standardowym (dane nie prezentowane). Zależność Dw_{CNORM} od parametru Θ_0 (teta0) oraz od γ (gamma) przedstawia wykres z rys.4 Na jego podstawie można stwierdzić, iż wartości liczbowe Dw_{CNORM} można uznać za stałe dla badanego zakresu gamma. Zależą one jedynie od parametru Θ_0. Wartości Dw_{CNORM} uzyskane za pomocą metody największej wiarygodności są dokładniejsze od wartości uzyskanych metodą największej oszczędności. Metoda największej wiarygodności charakteryzuje się mniejszą zależnością od sumarycznej liczby mutacji rekonstruowanego drzewa.

Zdolność obu metod do rekonstrukcji jedynie topologii bez uwzględnienia zgodności mutacji z drzewem rzeczywistym przedstawia rys. 5. Im D_t (dystans topologiczny) jest mniejsze, tym otrzymany wynik jest lepszy. Z rys. 5 wynika, że metoda największej wiarygodności jest nieznacznie gorsza. Jedynie dla Θ_0=50 wyniki D_t otrzymane obiema metodami są prawie jednako-

we. Dla przebadanych przebiegów wykładniczych metoda największej wiarygodności uzyskała lepszy wynik Dw_{CNORM} (rys.4), ponieważ wartości D_t uzyskane tą metodą są tylko nieznacznie gorsze od drugiej badanej metody, można stwierdzić, iż dla przebiegu wykładniczego liczebności populacji w czasie o wzorze $\Theta(t) = 50e^{-0.0625t}$ metoda największej wiarygodności dała jednoznaczny, najlepszy rezultat.

II. Przebieg dwustanowy

Wyniki otrzymane dla przebiegu dwustanowego zilustrowano na osobnych wykresach dla każdej z metod. Wyjątek stanowi rys.6, na którym umieszczono krzywe uzyskane z wyników otrzymanych dwiema badanymi metodami. Pozostałe wykresy przedstawiają zależności uzyskanych wartości Dw_C, D_t od czasu skoku t_i oraz od wartości, jaką przyjmuje przebieg po skoku, jest to wartość teta. Na rys.6 krzywa o etykiecie „100teta NW" w odróżnieniu od innych nie zachowuje monotoniczności w całym przedziale czasu skoku t. Przyczyną tego zjawiska jest wysoka wartość odchylenia standardowego dla punktów o czasie skoku 5 oraz 10.

Z porównania rys.7 i rys.8 wynika, iż uzyskano lepsze wyniki metodą największej wiarygodności dla skoku wartości Θ z 200→50 oraz 200→10. Wyniki dla tych wartości skoku cechują się również mniejszymi wartościami odchylenia standardowego. Natomiast dla skoku wartości Θ z 200→100 oraz Θ stałe równe 200 otrzymano metodą największej wiarygodności większe odchylenia standardowe. Na rys.9 i rys.10 zestawiono wartości D_t uzyskane obu metodami. Dla dwustanowego przebiegu liczebności populacji w czasie metodą największej oszczędności uzyskano lepiej zrekonstruowaną topologię generowanego drzewa.

III. Przebieg wielostanowy

Za pomocą przebiegu wielostanowego sprawdzono, jaki wpływ na jakość rekonstrukcji miały zadawane nietypowe przebiegi liczebności populacji. Testowano przebiegi o kształtach hipotetycznych, niespotykane w rzeczywistym toku ewolucji.

Zbadano przypadek, gdy Θ zmienia się 10→200→10. Ten skok liczebności populacji w czasie spowodował wzrost wartości Dw_C dla metody największej oszczędności z 63.78 na 81.15, a

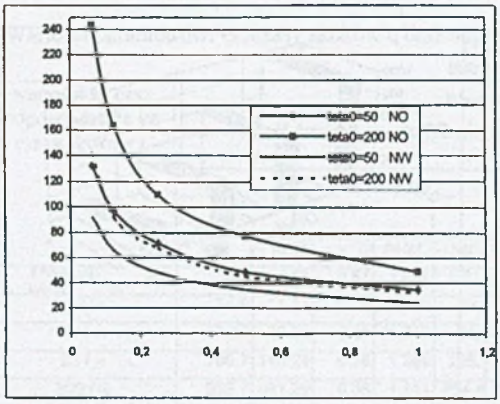
dla metody największej wiarygodności z 51.5 na 64.2. Gdy Θ zmienia się 10→50→10, również zaobserwowano wzrost Dw_C dla metody największej oszczędności z 63.78 na 83.18, a dla metody największej wiarygodności Dw_C wzrosło z 51.5 na 68.4. Skok liczebności populacji nie spowodował wyraźnych zmian w wartościach D_t uzyskanych za pomocą obydwu metod.

Jakość rekonstrukcji obu metod była gorsza, gdy przebieg liczebności populacji w czasie zmieniał się gwałtownie. W przypadku takiej zmiany wartość Dw_{CNORM} dla obu metod wzrosła. Zjawisko to jest widoczne przy porównaniu wyników, jakie uzyskano dla przebiegów: Θ zmienia się 200→180→150→90→50→20→10 oraz 200→50→10. W przypadku małych skoków dla metody największej oszczędności Dw_{CNORM}=1.091, a dla dużych skoków wartości Θ Dw_{CNORM}=1.266. Ten sam efekt zaobserwowano dla metody największej wiarygodności dla małych skoków Θ Dw_{CNORM}=0.675, a dla dużych skoków Θ Dw_{CNORM}=0.843.

Dla przebiegów stałych Θ = 1, 5, 10, 50, 200 za pomocą obu metod otrzymano zbliżone wyniki Dw_C (zamieszczone w tablicy 2.). Dla obu metod dla wykresu Θ stałe równe 200 otrzymano najlepsze zgodności topologii D_t.

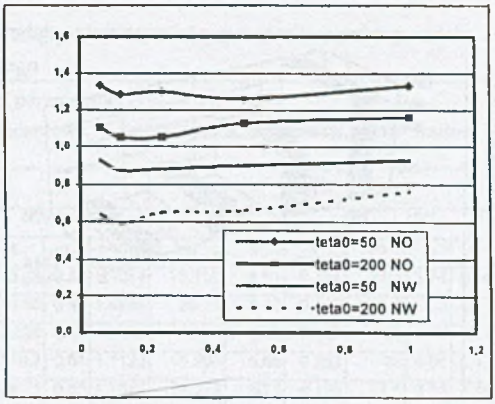
Drzewa zrekonstruowane przy użyciu metody oszczędności, dla wszystkich przebadanych przebiegów populacji, uzyskały lepsze wyniki biorąc pod uwagę jedynie topologię. Natomiast za pomocą metody największej wiarygodności otrzymano lepsze wyniki, uwzględniające mutacje w całym drzewie.

Dla przebiegów czysto hipotetycznych, czyli takich, w których liczebność populacji wzrasta wraz z cofaniem się w czasie, wyniki otrzymane za pomocą obu metod cechowały się najgorszą zgodnością z drzewami wzorcowymi.



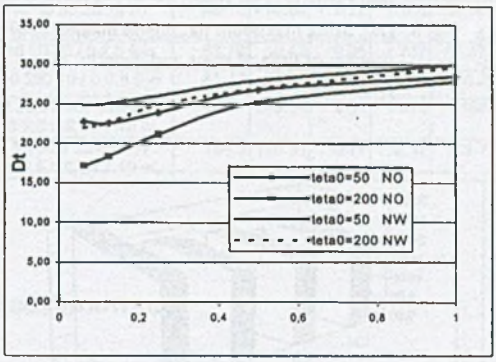
Rys.3. Zależność Dw_C od gamma obu metod

Fig.3. Dw_C as a functon of gamma for both methods



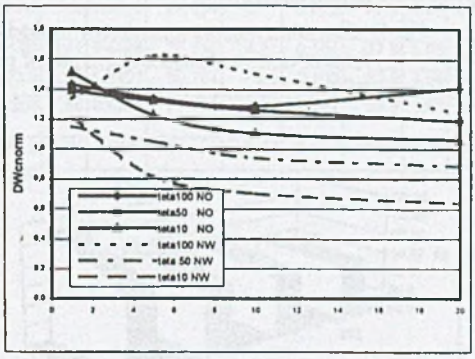
Rys.4. Zależność Dw_{CNORM} od gamma obu metod

Fig.4. Dw_{CNORM} as a functon of gamma for both methods



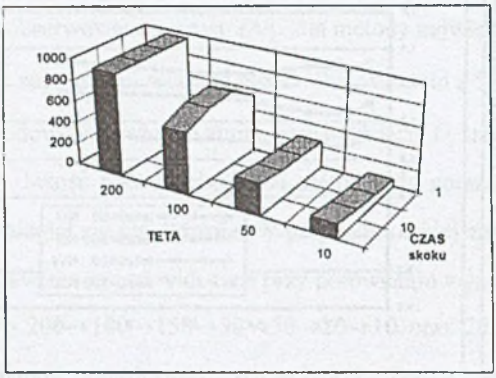
Rys.5. Zależność D_t od gamma dla obu metod

Fig.5. D_t as a functon of gamma for both methods



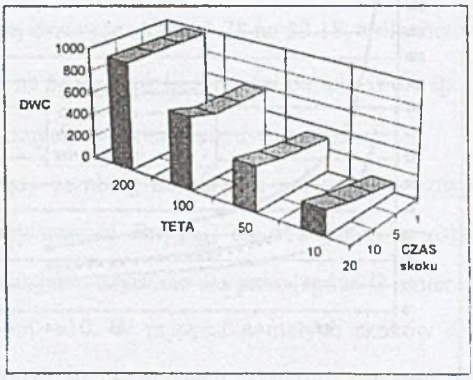
Rys.6. Zależność Dw_{CNORM} od czasu skoku dla obu metod

Fig.6. Dw_{CNORM} as a functon of time of jump for both methods



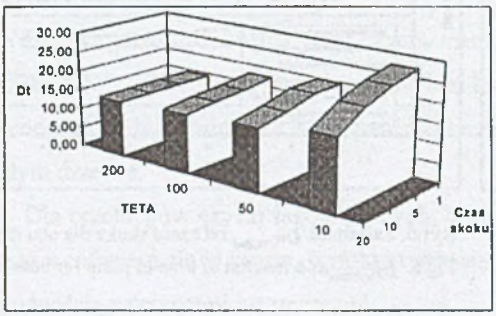
Rys.7. Zależność Dw_C uzyskane dla metody największej oszczędności od czasu skoku oraz wartości teta, jaką przyjmuje liczebność populacji po skoku

Fig. 7. Dw_C using maximum parsimony method as a functon of time of steps and quantity of populatin size teta



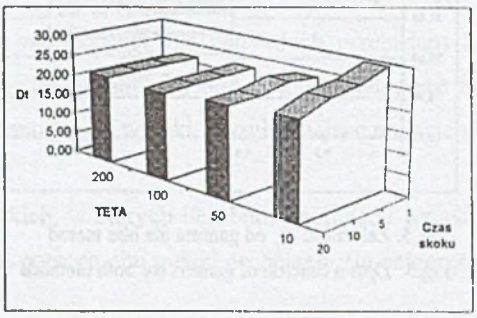
Rys.8. Zależność Dw_C uzyskane dla metody największej wiarygodności od czasu skoku oraz wartości teta, jaką przyjmuje liczebność populacji po skoku

Fig. 8. Dw_C using maximum likelihood method as a functon of time of steps and quantity of populatin size teta



Rys.10. Wartości D_t uzyskane dla metody największej wiarygodności jako funkcje czasu skoku i liczebności populacji

Fig.10. D_t using maximum likelihood method as a functon of time of steps and quantity of populatin size teta



Rys. 9. Wartości D_t uzyskane dla metody największej oszczędności jako funkcje czasu skoku i liczebności populacji.

Fig. 9. D_t using maximum parsimony method as a functon of time of steps and quantity of populatin size teta.

Tablica 2

Wartości parametrów: Θ ,czasy skoków t_i oraz odpowiadające im wartości Dw_C i D_l dla przebiegu wielostanowego

wartości Θ_i oraz odpowiadające im czasy skoków t_i	Wartości parametrów Dw_C i D_l xsr- wartość średnia, s- odchylenie standardowe, v- współczynnik zmienności, norm- wartość unormowana													
	Dw_C								D_l					
	MO				NW				MO			NW		
	xsr	s	v	norm	xsr	s	v	norm	xsr	s	v	Xsr	s	v
1 / ∞	6.6	4.58	0.67	1.651	5.615	3.48	0.62	1.4037	15.156	4.91	0.324	21.3	3.74	0.18
5 / ∞	34.9	15.61	0.45	1.785	29.2	13.69	0.47	1.4938	29.71	3.48	0.117	30.7	3.3	0.11
10 / ∞	63.78	26.38	0.41	1.613	51.5	22.98	0.45	1.3021	28.78	3.51	0.122	29	3.63	0.13
50 / ∞	290.7	110.24	0.38	1.369	236.1	119.7	0.51	1.1122	19.996	5.65	0.283	23	4.01	0.17
200 / ∞	959.7	331.26	0.35	1.111	954.4	579.25	0.62	1.7544	14.373	60.3	4.195	21.6	3.55	0.16
200 50 10 / 0 2,5 5 ∞	115.6	31.76	0.27	1.266	77.01	27.18	0.35	0.8428	24.662	4.34	0.176	26.6	3.59	0.14
200 180 150 90 50 10 / 0 2,5 5 6,25 7 10 ∞	148.9	33.53	0.23	1.091	92.09	35.6	0.39	0.675	21.209	4.95	0.233	24.7	3.65	0.15
50 10 5/0 2,5 5 ∞	63.63	17.96	0.28	1.218	44.25	14.58	0.33	0.8468	25.133	4.18	0.166	27	3.88	0.14
50 45 35 20 10 5 / 0 2,5 5 6,25 7 10 ∞	88.18	19.34	0.22	1.165	58.38	17.46	0.3	0.7714	21.834	4.71	0.216	25.3	4.24	0.17
10 50 10 / 0 2,5 5 ∞	83.78	30.43	0.36	1.737	68.43	28.02	0.41	1.4188	26.135	4.21	0.161	27.9	3.98	0.14
10 200 10 / 0 2,5 5 ∞	81.15	28.38	0.35	1.656	64.2	23.2	0.36	1.3102	30.324	3.1	0.102	30.8	3.54	0.12
10 50 90 150 190 200 / 0 2,5 5 6,25 7 10 ∞	760.7	389.54	0.51	1.816	1184	2203	1.86	2.8269	27.738	4.08	0.147	28.5	3.56	0.12
5 10 20 35 45 50 / 0 2,5 5 6,25 7 10 ∞	154.3	118.55	0.77	2.32	161.7	146.9	0.91	2.4321	28.131	3.83	0.136	29.2	3.83	0.13

Podsumowanie

Rozpatrzono problem dokładności rekonstrukcji drzew genealogicznych, w zależności od wpływu zmieniającej się w czasie liczebności populacji przy stałej częstotliwości procesu mutacji. Testowane drzewa powstały na podstawie kodów genetycznych dla przyjętych hipotetycznych dwudziestu osobników. Badano trzy typy przebiegu liczebności populacji w czasie: wykładniczy, dwustanowy oraz wielostanowy.

Dla każdej z metod wygenerowano 3900 drzew, uzyskując metodą największej oszczędności 319940 rekonstrukcji drzew, a metodą największej wiarygodności 3900. Na podstawie przeprowadzonej symulacji zbadano związki pomiędzy przebiegiem liczebności populacji a zdolnością badanych metod do prawidłowej rekonstrukcji drzewa genealogicznego.

1. Dla wszystkich przebadanych przebiegów metodą największej oszczędności otrzymano lepsze wyniki rekonstrukcji .
2. Otrzymano lepsze wyniki całkowitego dystansu wagowego Dw_C metodą największej wiarygodności dla wszystkich przebadanych przebiegów.
3. Wartości D_t, otrzymane obiema metodami, dla przebiegu liczebności populacji o stałej liczności Θ=200 okazały się być najlepsze spośród sprawdzanych. Należy to tłumaczyć w następujący sposób: duża liczebność populacji w czasie powoduje, że w gałęziach drzewa występuje duża liczba mutacji, a w związku z tym sekwencje pseudokodu przypisanego do liści drzewa wzorcowego różnią się znacznie. Duże różnice pomiędzy sekwencjami wejściowymi ułatwiają proces rekonstrukcji drzewa. Jednak gdy liczebność populacji byłaby tak znaczna, iż mutacje w pseudokodzie nakładałyby się wielokrotnie, spowodowałoby to taki sam skutek jak za mała ilość różnic pomiędzy sekwencjami.
4. Zaobserwowano zależność pomiędzy kształtem przebiegu liczebności populacji a współczynnikami zmienności uzyskanych wyników. Im przebieg ma większe skoki, tym współczynniki zmienności są większe. Oznacza to, że analizowane metody największej oszczędności i największej wiarygodności dają więcej wyników o wartościach skrajnych.

W pracy tej rozpatrywano drzewa dla 20 osobników. Uzyskane wyniki są zbliżone do wyników często spotykanych w literaturze dla przypadku stałej liczebności populacji w czasie.

Wykorzystane oprogramowanie

Program generujący drzewa wzorcowe oraz program porównujący drzewo estymowane ze wzorcowym powstały w środowisku MATLAB. Programy DNApars oraz DNAmlk są częścią pakietu pod nazwą PHYLIP. Twórcą tego pakietu jest Joseph Felsenstein z Department of Genetics, University of Washington, Seattle. PHYLIP jest dostępny za pomocą sieci Internet pod adresem: fttp://evolution.genetics.washington.edu./phylip.html

LITERATURA

1. Cavalli-Sworza L.L., Menozzi P. and Piazza A.: The history and geography of human genes. Princeton University Press, vol.376, 1994, pp.30-39.
2. Clifford W. Cunningham, Kevin E. Omland, Todd H. Oakley.: Reconstruction ancestral character states: a critical reappraisal. Tree vol.13 no. 9 september 1998.
3. Farris J. S.: Methods for computing Wagner trees. Syst. Zool. vol. 19, 1970, pp.83-92.
4. Feller W.: Wstęp do rachunku prawdopodobieństwa i jego zastosowań. PWN, Warszawa 1960.
5. Felsenstein J.: Maximum likelihood and minimum steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. vol. 22, 1973, pp.240-249.
6. Fitch W.M., Margoliash E.: Contruction of phylogenic trees. Science vol. 155, 1967, pp.279-284.
7. Fitch W.M.: Toward defining the course of evolution minimum change for a specific tree topology. Syst. Zool. vol. 20, 1971, pp. 406-416.
8. Hasegawa M, Kishino H,Yano T.: Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol Evol. vol. 22, 1985, pp.160-174.
9. Jeffrey M. Koshi, Richard A. Goldstein: Probablic reconstruction of ancestral protein sequences. J. Mol. Evol. vol. 42, 1996, pp.313-320.
10. Mary K.: Kunher and Joseph Felsenstein. A simulatin comparision of phylogeny algoritms under Equal and unequal evolutionary rates. Mol. Biol. Evol. vol.11(3), 1994, pp.459-268.
11. Polanski A., Kimmel M., Chakraborty R.: Application of time-dependant coalescence process for inferring the history of population size changes from DNA sequences data. Proc. Natl. Acad. Sci. USA vol. 95, 1998, pp. 5456-5461.
12. Robinson D.F.: Comparison of Weighted Labelled Trees. Combinational Mathematics VI, Lecture Notes in Matematiccs 748, Springer, Berlin 1979, pp. 119-126.
13. Robinson D.F., Foulds L.R.: Comparison of phylogenetic Trees. Math. Biosc. vol. 53, 1981, pp.131-147.
14. Siatou N, Nei N.: The neighbor joining method: a new method for reconstructing phylogenetic trees. Amol. Biol. Evol vol. 4, 1987, pp.406-425.
15. Weir B.S.: Genetic data analysis II. 1996.

16. Ziheng Yang.:Phylogenic analysis using parsimony and likelihood methods. J. Mol Evol vol. 42, 1996, pp.294-307.

17. Yoshio Tateno: A method for phylogeny construction by use of nucleotide sequence data. J.Mol.Evol. vol. 30, 1990, pp.85-93.

Recenzent: Dr hab.inż. Jacek Waniewski

Wpłynęło do Redakcji 12 stycznia 2000 r.

Abstract

The problem of the topology reconstruction based on the differences among random sample of individuals from population and genealogical trees metrics have large meaning in domains like paleontology, anthropology, genetics, taxonomy. From mathematical methods and mathematical modeling point of view of this problem has also a great importance in the estimation theory, in the statistics and in the theory of algorithms.

There are many methods of reconstruction of the genealogical trees for DNA sequence developed till present day. The input data for these methods is set of sequences of DNA, for example of the different species or of representatives of human races. Number of these sequences is equal to number of species or of races. Length of the sequence is dependent of amount of collected data [5],[7].

In the present work we consider a problem of reconstruction of the genealogy based on observing differences among sequences of DNA acid, drawn randomly from population of individuals.

Maximum parsimony method: it counts the minimal number of mutations necessary for determining an observed difference between sequences of DNA. The trees, which minimize the number of required changes in the input sequences are selected. They do not explicate models of evolution [2].

Maximum likelihood method uses the statistical models of nucleotide changes and their frequencies of occurrence. On this to base course of mutation is fitted to given sequence, building the genealogical tree. This method estimates constant coefficient of the mutation event [1].

The problem of reconstructing genealogical tree can be solved using two kinds of methods: statistical or heuristic. It is essential to compare results obtained using these different methods. A number of such the comparisons one may found in literature [7]. However the essential assumption, which was taken in the mentioned works was that in the evolution of the population size remains constant. This assumption is of course only rough approximation of the reality. It is known that most of populations change their size during evolution. Therefore examining of the methods of recovering genealogy in case of varying population size during evolution seems interesting. This problem is researched in the present work.