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

Ewa ŁOBOS

ZMIANY W ROZMIARZE POPULACJI A ROZKŁAD RÓŻNICY LICZBY MUTACJI W n-ELEMENTOWYCH GRUPACH Z TEJ POPULACJI

Streszczenie. W artykule tym zostały wyprowadzone wzory rekurencyjne na funkcje tworzące liczby mutacji w próbce n-elementowej (n=2,3,...), przy założeniu że rozmiar populacji jest funkcją czasu. Dokładne obliczenia przeprowadzono dla n=3 i porównano je z danymi doświadczalnymi dla czterech modeli rozwoju populacji.

CHANGES IN POPULATION SIZE AND NUMBER OF SEGREGATING SITES IN n-INDIVIDUAL SAMPLE

Summary. This paper presents recurrent formulae for probability generating function of number of mutations in n-individuals sample (n=2,3,...) with assumption that population size is changing in time. Exact calculations are for n=3 and we compare them with experimental results for four models of population evolution.

1.Wstęp

Na podstawie badań nad sekwencjami zasad w łańcuchach DNA w różnych populacjach stwierdzono występowanie różnic pomiędzy osobnikami. Różnice te są spowodowane działaniem przypadkowych mutacji. Od lat sześćdziesiątych tworzono modele matematyczne rozwoju populacji (np. [1], [2]), które wykorzystywano m.in. do oszacowania intensywności procesu mutacji w różnych regionach ludzkiego genomu.

A.R.Rogers i H.C.Harpending w pracy [6] postawili hipotezę, że w rozkładzie liczby różnic między pewnymi homologicznymi sekwencjami DNA dla par elementów danej populacji jest zawarta informacja o historii zmian rozmiaru tej populacji. Jeżeli rozkład ten otrzymujemy na

podstawie takiej sekwencji mtDNA[1], w której zachodzą tylko mutacje selektywnie neutralne[2], to im większy jest rozmiar danej generacji, tym większej liczby nowych mutacji możemy oczekiwać w następnym pokoleniu. W pracach [6] i [7] różnymi metodami oszacowano zmiany w rozmiarze populacji ludzi na podstawie danych dotyczących rozkładu różnic między parami mtDNA (opublikowanych w [4]). Otrzymano różne wyniki. Biorąc pod uwagę niestabilność rozwiązań tego problemu (zob. [7]), różnice te są niewielkie.

Celem tej pracy jest wyprowadzenie wzorów na funkcje tworzące liczby mutacji między n elementami próbki, przy założonym zmiennym rozmiarze populacji. Wzory te mają wygodną rekurencyjną postać i sprowadzają się do znanych wzorów Wattersona ([1]) w przypadku populacji o stałym rozmiarze. Dla n=3 przeprowadzono dokładne obliczenia, które zastosowano do czterech modeli rozwoju populacji: dwóch teoretycznych (skokowego i wykładniczego) oraz dwóch dotyczących ludzkiej populacji (podanych w pracach [6] i [7]). Wyniki tych obliczeń porównano z danymi uzyskanymi drogą symulacji i z danymi doświadczalnymi.

Opis modelu

Rozważamy populację, która w chwili obecnej składa się z $2N_0$ osobników, jedno pokolenie wcześniej liczyło $2N_1$ osobników i, ogólnie, τ pokoleń wcześniej rozmiar tej populacji wynosił $2N_\tau$. Zakładamy, że populacja w każdej chwili τ jest na tyle duża, że $\frac{1}{(2N_\tau)^2} \approx 0$, a na polimorfizm ustalonego odcinka DNA wpływają tylko zdarzenia neutralnych mutacji (w każdej generacji niezależne od siebie, o rozkładzie Poissona z parametrem ν, model *infinite sites*) oraz dryft genetyczny (model Wrighta-Fishera ze zmiennym rozmiarem populacji). W wyniku dryftu genetycznego następuje utrata różnorodności alleli w genomie osobników danej populacji, natomiast mutacje, jako losowe zmiany w kolejności nukleotydów, tę różnorodność zapewniają.

[1] mitochondrialne DNA jest przekazywane w zasadzie tylko przez matkę i rzadko podlega rekombinacji
[2] tzn. mutacje, które nie mają wpływu na wielkość populacji

2. Rozkład liczby mutacji w próbce *n*-elementowej

2.1. Prawdopodobieństwa koalescencji

Oznaczmy przez $p_n(\tau_1,\tau_2)$ prawdopodobieństwo zajścia sytuacji takiej, że w chwili τ_1 próbka liczy n osobników, natomiast w chwili τ_2 dwóch z nich ma wspólnego przodka (zachodzi koalescencja) i próbka składa się już z n-1 osobników. Wówczas

$$p_n(\tau_1,\tau_2)=\frac{\binom{n}{2}}{2N_{\tau_2}}\prod_{\sigma=\tau_1+1}^{\tau_2-1}\left(1-\frac{\binom{n}{2}}{2N_\sigma}\right). \tag{1}$$

Wzór (1) jest łatwym uogólnieniem wzoru (5) z [3].

Dla ułatwienia opisu matematycznego uważamy dyskretną zmienną τ za zmienną ciągłą (wówczas rozmiar populacji jest funkcją $2N(\tau)$); po zastosowaniu przybliżenia $1-\frac{1}{2N(\tau)}\approx e^{-\frac{1}{2N(\tau)}}$ otrzymujemy

$$p_n(\tau_1,\tau_2)=\frac{\binom{n}{2}}{2N(\tau_2)}\exp\left(-\int_{\tau_1}^{\tau_2}\frac{\binom{n}{2}}{2N(\sigma)}d\sigma\right). \tag{2}$$

Funkcję $p_n(\tau_1,\tau_2)$ nazywamy funkcją intensywności koalescencji dla próbki n-elementowej.

Oznaczmy $p_n(\tau)=p_n(0,\tau)$ i $P_n(\tau)=\int_\tau^\infty p_n(\sigma)d\sigma$. Wówczas

$$p_n(\tau_1,\tau_2)=\frac{p_n(\tau_2)}{P_n(\tau_1)} \tag{3}$$

oraz

$$p_n(\tau_1,\tau_2)=\binom{n}{2}\frac{p_2(\tau_2)}{P_2(\tau_1)}\left(\frac{P_2(\tau_2)}{P_2(\tau_1)}\right)^{\binom{n}{2}-1}.$$

Z ostatniego wzoru wynika, że do obliczenia prawdopodobieństwa koalescencji próbki n-elementowej wystarczy znajomość prawdopodobieństw koalescencji dla par.

2.2. Funkcja tworząca liczby mutacji w próbce *n*-elementowej

Oznaczmy przez $\alpha_n(s,\tau_0)$ funkcję tworzącą liczby mutacji, które pojawiają się w próbce liczącej w chwili τ_0 n osobników (tzn. od chwili pojawienia się ich najbliższego wspólnego przodka do τ_0). Liczba mutacji w drzewie o n gałęziach na odcinku między τ_0 i τ ma rozkład Poissona z parametrem $n\nu(\tau-\tau_0)$ i ma funkcję tworzącą $e^{n\nu(\tau-\tau_0)(s-1)}$. Ponieważ zdarzenia mutacji są niezależne, a funkcja tworząca sumy niezależnych zmiennych losowych jest iloczynem funkcji tworzących tych zmiennych, otrzymujemy

$$\alpha_n(s,\tau_0)=\int_{\tau_0}^{\infty} e^{n\nu(\tau-\tau_0)(s-1)}\alpha_{n-1}(s,\tau)p_n(\tau_0,\tau)d\tau .$$

Korzystając ze wzoru (3) mamy:

$$\alpha_n(s,\tau_0)=\frac{e^{-n\nu\tau_0(s-1)}}{P_n(\tau_0)}\int_{\tau_0}^{\infty} e^{n\nu\tau(s-1)}\alpha_{n-1}(s,\tau)p_n(\tau)d\tau . \tag{4}$$

Funkcję tworzącą liczby mutacji w drzewie genealogicznym próbki n-elementowej otrzymamy podstawiając $\tau_0=0$:

$$\alpha_n(s)=\int_0^{\infty} e^{n\nu\tau(s-1)}\alpha_{n-1}(s,\tau)p_n(\tau)d\tau , \tag{5}$$

co w porównaniu ze wzorem (4) z [6]:

$$\alpha_2(s,0)=\int_0^{\infty} e^{2\nu\tau(s-1)}p_2(\tau)d\tau \tag{6}$$

i jego konsekwencją

$$\alpha_2(s,\tau_0)=\frac{e^{-2\nu\tau_0(s-1)}}{P_2(\tau_0)}\int_{\tau_0}^{\infty} e^{2\nu\tau(s-1)}p_2(\tau)d\tau ,$$

daje rekurencyjną zależność na funkcje tworzące $\alpha_n(s,\tau_0)$, n=2,3,... .

2.3. Populacja o stałym rozmiarze

Załóżmy, że rozmiar populacji nie zmienia się w czasie, tzn. $2N(\tau)=2N$ dla każdego $\tau\geq 0$. Wówczas rozkład liczby mutacji nie zależy od chwili początkowej τ_0, a zależy jedynie od liczebności próbki i długości gałęzi. Mamy więc $\alpha_n(s,\tau)=\alpha_n(s)$ i we wzorze (5) otrzymujemy:

$$\alpha_n(s)=\alpha_{n-1}(s)\int_0^{\infty} e^{n\nu\tau(s-1)}p_n(\tau)d\tau . \tag{7}$$

W przypadku populacji o stałym rozmiarze $p_n(\tau)=\frac{\binom{n}{2}}{2N}e^{-\frac{\binom{n}{2}}{2N}\tau}$, więc we wzorze (7) mamy

$$\alpha_n(s) = \alpha_{n-1}(s) \cdot \frac{1}{1 - \frac{\theta(s-1)}{n-1}}, \text{ gdzie } \theta = 4N\nu. \tag{8}$$

Łatwo sprawdzić, że $\alpha_2(s) = \frac{1}{1-\theta(s-1)}$, co w porównaniu z (8) daje:

$$\alpha_n(s) = \prod_{k=1}^{n-1}\left[1 + \frac{\theta(1-s)}{k}\right]^{-1}. \tag{9}$$

Wzór (9) jest wzorem (1.3a) podanym przez G.A.Wattersona ([1]).

3. Rozkład liczby mutacji dla trójek

Dla n=3 wzór (5) przyjmuje postać:

$$\alpha_3(s) = 3\int_0^{\infty}\left(\int_r^{\infty} e^{\nu(s-1)(\tau+2\tau_1)} p_2(\tau)P_2(\tau)p_2(\tau_1)d\tau_1\right)d\tau.$$

Rozwijając $e^{\nu(s-1)(\tau+2\tau_1)}$ w szereg względem zmiennej s, otrzymamy

$$\alpha_3(s) = \sum_{k=0}^{\infty} s^k q_k,$$

gdzie

$$q_k = \frac{3}{k!}\int_0^{\infty}\left(\int_r^{\infty} e^{-\nu(\tau+2\tau_1)}\left(\nu(\tau+2\tau_1)\right)^k p_2(\tau)P_2(\tau)p_2(\tau_1)d\tau_1\right)d\tau$$

jest prawdopodobieństwem wystąpienia k mutacji w próbce 3-elementowej.

Ponieważ w praktyce rzadko znamy wartość parametru ν, wprowadzamy zamianę zmiennych: $t = 2\nu\tau, t_1 = 2\nu\tau_1$, przy czym oznaczamy $\frac{1}{2\nu}p_2\left(\frac{t}{2\nu}\right) = \pi_2(t)$ i $P_2\left(\frac{t}{2\nu}\right) = \Pi_2(t)$, natomiast rozmiar populacji charakteryzowany jest przez funkcję $\theta(t) = 4\nu N\left(\frac{t}{2\nu}\right)$. Zmieniając kolejność całkowania, otrzymujemy

$$q_k = \frac{3}{k!}\int_0^{\infty} e^{-t_1}\pi_2(t_1)\left(\int_0^{t_1}\left(\frac{1}{2}t + t_1\right)^k e^{-\frac{t}{2}}\pi_2(t)\Pi_2(t)dt\right)dt_1. \tag{10}$$

4. Wyniki obliczeń

Rozważmy cztery populacje:

1) populację A, której rozmiar zmienia się skokowo

$$\theta(t) = \begin{cases} \theta_0, \text{dla } t \le t_s \\ \theta_1, \text{dla } t > t_s \end{cases}, \qquad (11)$$

gdzie $\theta_0 = 100, \theta_1 = 1, t_s = 10$,

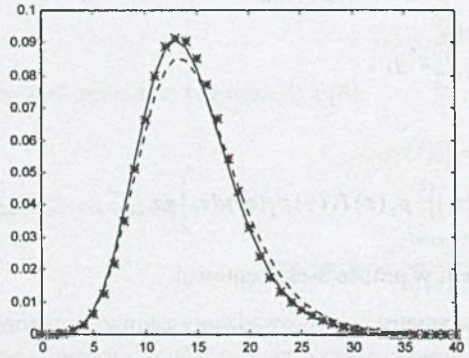
2) populację B, której rozmiar zmienia się wykładniczo

$$\theta(t) = \theta_0 e^{-\gamma t},$$

gdzie $\theta_0 = 650, \gamma = 0{,}5$,

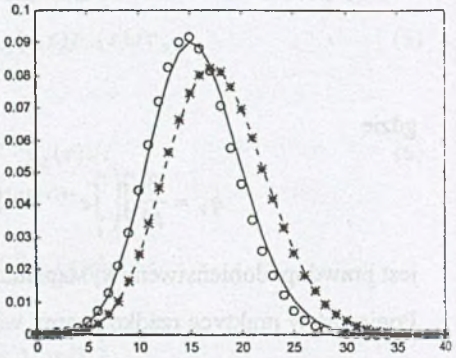
3) populację C, której rozmiar zmienia się jak na rys.2a w [7],

4) populację D, której rozmiar zmienia się skokowo, zgodnie ze wzorem (11) przy $\theta_0 = 410{,}69$, $\theta_1 = 2{,}44$, $t_s = 7{,}18$ (por.[6], s.555).



Rys.1. Prawdopodobieństwa q_k wystąpienia k=0,1, ...40 różnych mutacji między trójkami, przy założeniach że rozmiar populacji ludzkiej zmienia się jak w [6] (linia ciągła) i jak w [7] (linia przerywana) oraz rzeczywiste częstości występowania tych różnic uzyskane na podstawie danych z [4] (gwiazdki)

Fig1. The probabilities q_k of k=0,1,...40 segregating sites in triplet calculated for human population, if its size is taken from [6] (solid line) and from [7] (dotted line). The experimental data (stars) are based on [4]



Rys.2. Prawdopodobieństwa q_k wystąpienia k=0,1, ...40 różnych mutacji między trójkami dla populacji A (linia ciągła) i populacji B (linia przerywana) oraz wyniki symulacji (odpowiednio kółka i gwiazdki)

Fig.2. The probabilities q_k of k=0,1,...40 segregating Sites in triplet calculated for population A (solid line) and B (dotted line), and the results of simulations (open circles and stars)

Na rys.1 i 2 przedstawiono prawdopodobieństwa q_k (k=0,1,...40) wystąpienia k różnic między trójkami dla każdej z tych populacji wyliczone ze wzoru (10). Występujące w tym wzorze funkcje $\pi_2(t)$ i $\Pi_2(t)$ oblicza się następująco: $\pi_2(t) = \frac{1}{\theta(t)} \exp\left(-\int_0^t \frac{d\sigma}{\theta(\sigma)}\right)$, $\Pi_2(t) = \pi_2(t)\theta(t)$.

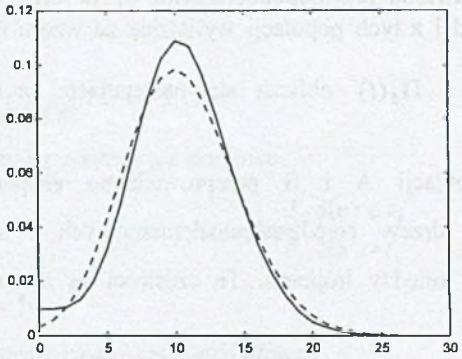
Dodatkowo dla populacji A i B przeprowadzono eksperymenty symulacyjne - wygenerowano po sto drzew pięćdziesięcioelementowych i obliczono częstości p_k występowania k mutacji między trójkami. Te częstości są zaznaczone na rys.2 kółkami i gwiazdkami.

Ponieważ funkcje $\theta(t)$ dla populacji C i D są oszacowaniami rozmiaru populacji ludzi i powstały na podstawie rozkładu różnic między parami mtDNA, warto je porównać z częstościami p_k występowania k różnic między trójkami mtDNA, które można uzyskać na podstawie danych z [4]. Te częstości są zaznaczone na rys.1 gwiazdkami.

Jak widać, jedynie w przypadku C (rys.1) dane doświadczalne wyraźnie odbiegają od obliczeń dokładnych. Jeżeli za miarę „dopasowania" przyjąć sumę $\sum_{k=0}^{40} |q_k - p_k|$, to suma ta dla populacji A, B, C i D wynosi odpowiednio 0,0194, 0,0266, 0,0863 i 0,0433.

5. Podsumowanie

Chociaż wydaje się, że oszacowanie rozmiaru populacji ludzi jest lepsze w pracy [6] niż [7], to warto jednak zwrócić uwagę na fakt, że wykorzystane tu dane doświadczalne dotyczą tylko jednego drzewa (147-elementowego). Ponieważ funkcje $\theta(t)$ dla populacji C i D prowadzą do zbliżonych rozkładów różnic dla par (por. rys.3 w [6] i rys.2b w [7]), więc parametry dla populacji A i B zostały tak dobrane, aby rozkłady różnic dla par w ich przypadku również były do siebie zbliżone. Dla tych populacji rozkłady różnic dla par przedstawia rys.3.



Rys.3. Prawdopodobieństwa q_k występowania k=0,1,...30 różnic między parami dla populacji A (linia ciągła) i B (linia przerywana)
Fig.3. The probabilities q_k of k=0,1,...30 pairwise differences in population A (solid line) and B (dotted line)

Jeżeli przez $q_k^{(2)}(X)$ i $q_k^{(3)}(X)$ oznaczymy odpowiednio prawdopodobieństwo wystąpienia k różnic w parach i w trójkach dla populacji X, to $\sum_{k=0}^{30}\left|q_k^{(2)}(\mathrm{A})-q_k^{(2)}(\mathrm{B})\right|=0{,}1158$, zaś $\sum_{k=0}^{40}\left|q_k^{(3)}(\mathrm{A})-q_k^{(3)}(\mathrm{B})\right|=0{,}3550$.

Można stąd wyciągnąć wniosek, że znajomość rozkładu różnic dla trójek niesie dodatkowe informacje, które pozwolą na dokładniejsze oszacowanie zmian rozmiaru populacji. Wydaje się, że zastosowanie rozkładu różnic dla trójek, czwórek, piątek itd. może dać lepsze metody rozróżniania populacji, które trudno odróżnić na podstawie rozkładu liczby różnych mutacji między parami.

LITERATURA

1. WATTERSON G.A.: On the number of segregating sites in genetical models without recombination, Theor. Popul. Biol. 7, 1975, 256-276.
2. LI W.H.: Distribution of nucleotide differences between two randomly chosen cistrons in a finite population, Genetics **85**, 1977, 331-337.
3. TAJIMA F.: Evolutionary relationship of DNA sequences in finite populations, Genetics **105**, 1983, 437-460.

4. CANN R.L., STONEKING M., WILSON A.C.: Mitochondrial DNA and human evolution, Nature **325**, 1987, 31-36.
5. TAJIMA F.: The effect of change in population size on DNA polymorphism, Genetics **123**, 1983, 597-601.
6. ROGERS A.R., HARPENDING H.C.: Population Growth Makes Waves in the Distribution of Pairwise Genetic Differences, Mol.Biol.Evol.**9**, 1992, 552-569.
7. POLAŃSKI A., KIMMEL M., CHAKRABORTY R.: Application of a time-dependent coalescence process for inferring the history of populations size changes from DNA sequences data, Proc.Natl.Acad.Sci. **95**, 1998, 5456-5461.

Recenzent: Prof. dr hab. Ryszard Tadeusiewicz

Wpłynęło do Redakcji 5 stycznia 1999 r.

Abstract

In this article we consider the number of mutations (segregating sites) which occur in genealogy of n-individual sample taken from population with changing size. For any n=2,3,.. we derive formulae (5)-(6) for probability generating function of number of mutations in such sample. If population size is constant from (5)-(6) we have known Watterson formulae ([1]). From these formulae we obtain probability for number of triple differences (10) and we use it to calculations. We do it for four population models - two theoretical and two based on research concerning human population (from [6] and [7]). Next, we compare our computations with the experimental data - it is shown on Fig.1 and Fig.2.