ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1998
Seria: AUTOMATYKA z.122 Nr kol. 1388

Joanna POLAŃSKA[1)]
Marek KIMMEL[2)]

MODEL MATEMATYCZNY PROCESU REKOMBINACJI CHROMOSOMÓW[3)]

Streszczenie. W pracy przedstawiono model matematyczny ewolucji rozkładów alleli dla dwóch genów typu VNTR (Variable Number of Tandem Repeats). W procesie tworzenia modelu uwzględniono wszystkie możliwe zdarzenia, z jednoczesnym występowaniem trzech oddziaływań genetycznych: mutacji, dryfu genetycznego i rekombinacji. Otrzymane wzory znaleźć mogą z powodzeniem zastosowanie w analizie statystycznej danych eksperymentalnych.

A MATHEMATICAL MODEL OF RECOMBINATION AND MUTATION PROCESS FOR MIKROSATELLITE LOCI

Summary. The mathematical model of two-linked microsatellite loci evolution is presented. It considers all possible events, and includes three genetic processes: mutation, genetic drift, and recombination. The obtained equations could be successfully applied to the stochastic analysis of experimental data.

1. Wstęp

Rekombinacja polega na losowym krzyżowaniu się dwóch ramion chromosomu a następnie utworzeniu połączenia pomiędzy przeciwległymi ramionami. Uważa się, że w ewolucji proces rekombinacji ma znaczenie dla zwiększania różnorodności w populacjach, a przez to zwiększania ich zdolności dostosowawczych. Wykorzystując możliwości współczesnej bio-

[1)] Instytut Automatyki, Politechnika Śląska, ul. Akademicka 16, 44-100 Gliwice email: jpolanska@ia.polsl.gliwice.pl
[2)] Department of Statistics, Rice University, P.O.Box 1892, Houston, TX 77251, USA
[3)] Praca częściowo finansowana przez KBN Grant nr 8T11E01511

chemii dokonuje się obserwacji przypadków rekombinacji z wykorzystaniem technik sekwencjonowania DNA. Proces ten także wyjaśnia się na poziomie molekularnym.

W genetyce populacyjnej tworzy się modele matematyczne, których celem jest ujęcie ilościowe (lub choćby jakosciowe) tego zjwiska z punktu widzenia procesów, które zachodzą, gdy patrzymy na dziedziczenie genów w kolejnych generacjach różnych populacji [15]. Oprócz wartości poznawczej tego typu modele mają ważne zastosowanie. Jest nim możliwość tworzenia tzw. map genów, w których odległości między genami mierzy się częstotliwością przypadków rekombinacji zachodzących między tymi genami przy dziedziczeniu [4] [5] [13] [7] [1] [14]. Im większa częstotliwość, tym większa jest odległość pomiędzy genami (mierzona w cM - centymorganach zdefiniowanych jako 100 x prawdopodobieństwo zajścia zdarzenia rekombinacji pomiędzy genami). W populacji ludzkiej, przy dziedziczeniu, w jednym procesie mejozy zachodzi średnio ok. 30 przypadków rekombinacji. Biorąc pod uwagę wymiar ludzkiego genotypu (ok. 3.5 x 10^9 par nukleotydów) intensywność tego procesu na jedną parę nukleotydów jest bardzo mała. Dlatego do jego badania znaczenie mają metody statystyczne oparte na dużych próbach z populacji oraz modele matematyczne genetyki populacyjnej [13] [2] [12].

Należy podkreślić, że z punktu widzenia modelowania matematycznego zjawisko rekombinacji w procesie dziedziczenia przez kolejne pokolenia jest bardzo złożone - zawiera szereg nieliniowości i wymaga tworzenia skomplikowanych genealogii.

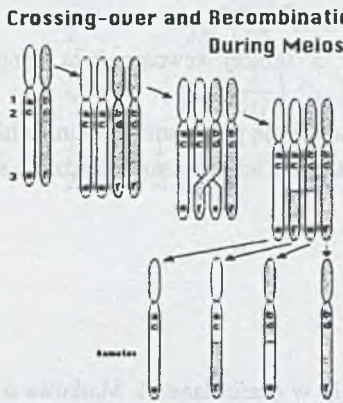
W artykule niniejszym przedstawiony jest model ewolucji rozkładów alleli dla dwóch genów typu VNTR (Variable Number of Tandem Repeats) [11]. Zamiast określenia „geny typu VNTR" używać będziemy terminu „lokusy", który podkreśla neutralność ewolucyjną większości kodów zawierających powtórzenia motywów w postaci di-tri- czy kwardrupletów. Liczba alleli na lokusach tego typu jest zwykle dość duża, różnorodoność może obejmować nawet kilkadziesiąt wariantów. Stąd wynika ich duża użyteczność w badaniach wykorzystujących metody statystyczne. Utworzony model w najbardziej ogólnym przypadku ma postać relacji pomiędzy prawdopodobieństwami przejść dla pewnego łańcucha Markowa. Można mu także nadać postać złożonego równania różniczkowego dla funkcji tworzącej. Model ten jest bardzo skomplikowany z uwagi na uwzględnienie w nim wszystkich możliwych zdarzeń, z jednoczesnym występowaniem trzech oddziaływań genetycznych: mutacji, dryftu genetycznego i rekombinacji. Jednak przy wprowadzeniu odpowiednich uproszczeń do tego modelu udaje się wyprowadzić, jako przypadki szczególne, znane wcześniej w literaturze zależności dla uproszczonych opisów ewolucji z mutacją. Wyprowadzamy także równania ewolucji dla momentów odpowiednich zmiennych losowych. Metoda momentów została zaproponowana w pracy jako sposób tworzenia miary odległości między lokusami. Wyprowadzone wzory mogą być podstawą do analizy szeroko dostępnych danych statystycznych, dotyczących rozkładów alleli liczb powtórzeń na lokusach typu VNTR. Poniżej przedstawiamy postulaty dotyczące zmiennych losowych i procesów statystycznych występujących w modelu.

2. Model populacji

Rozważa się populację ewoluujących w czasie 2N haploidalnych osobników, odpowiadającą ciągłej w czasie wersji modelu Morana [9]. Analizie poddaje się przypadek dwu VNTR lokusów (mikrosatelitów), oznaczanych odpowiednio przez 1 i 2, o przeliczalnym zbiorze alleli indeksowanych przez liczby całkowite.

2.1. Proces narodzin i śmierci, losowanie z puli alleli i rekombinacja

- Każdy osobnik w momencie śmierci jest natychmiast zastępowany przez innego osobnika (następuje jego odrodzenie). Szereg kolejnych narodzin i śmierci danego osobnika tworzy jednorodny proces Poissona o intensywności $\lambda/2$ w zakresie $[0,\infty)$, zwany jego linią czasu lub historią osobniczą. Czas życia (czas trwania jednego cyklu od momentu narodzin do śmierci) jest zatem zmienną losową o rozkładzie wykładniczym z parametrem $\lambda/2$.
- W każdym momencie narodzin materiał genetyczny odradzającego się osobnika jest odtwarzany poprzez proces losowania z puli alleli, przy czym:
  - z prawdopodobieństwem 1 - r para lokusów jest losowana ze zwracaniem spomiędzy par lokusów wszystkich osobników, wliczając w to zmarłego osobnika,
  - z prawdopodobieństwem r każdy z lokusów jest niezależnie losowany spomiędzy odpowiednich lokusów wszystkich oobników (proces rekombinacji).
- Stała r jest nazywana współczynnikiem rekombinacji. Zmienia się on w zakresie od 0 dla bardzo blisko położonych lokusów, do $\frac{1}{2}$ - dla bardzo odległych. Schematycznie proces rekombinacji przedstawia rys.1 [8].

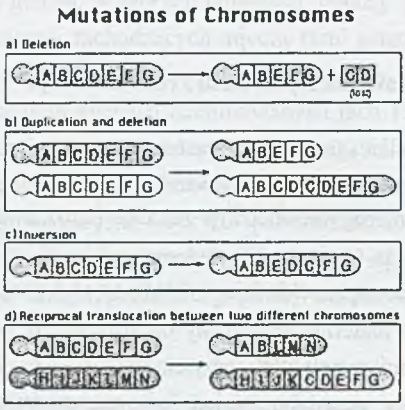


Rys.1. Kolejne etapy procesu rekombinacji
Fig.1. The phases of recombination

2.2. Mutacja

- Linia życia każdego z osobników zawiera mutacje występujące niezależnie na każdym z lokusów. Zdarzenia te definiują niezależne jednorodne procesy Poissona w zakresie $[0, \infty)$ z intensywnościami odpowiednio ν_1 i ν_2.

  Możliwe procesy mutacyjne ideowo przedstawia rys.2 [10].



Rys.2. Obrazowe przedstawienie procesu mutacji
Fig.2. The types of mutation process

- W przypadku wystąpienia mutacji na danym lokusie i (i = 1, 2) następuje zamiana allela o długości X_i przez allel o długości $X_i + U_i$, gdzie U_i jest całkowitoliczbową zmienną losową o następującej funkcji tworzącej:

$$\varphi_i(s) = \sum_{u=-\infty}^{\infty} s^u \varphi_{iu} = \sum_{u=-\infty}^{\infty} s^u \Pr\{U_i = u\} = E[s^{U_i}], \tag{1}$$

zdefiniowanej dla zmiennej zespolonej s leżącej wewnątrz koła jednostkowego na płaszczyźnie zespolonej.

Zdarzenia narodzin i śmierci oraz mutacje są procesami wzajemnie niezależnymi dla każdego z oobników. Są one również niezależne od procesów narodzin i śmierci oraz mutacji pozostałych osobników.

3. Model matematyczny

Przedstawiony proces stanowi ciągły w czasie łańcuch Markowa o przestrzeni stanu S, zawierającej wszystkie możliwe pary alleli dla każdego z 2N osobników, tzn. $E = (Z^{x2})^{x(2N)}$ i momentach zatrzymania zdeterminowanych przez poissonowski proces narodzin i śmierci oraz zdarzenia mutacyjne. Zapisane w sposób jawny macierzy tranzycji może być dla takiego

systemu bardzo skomplikowane, dlatego też w pracy proponuje się wykorzystanie równań perspektywnych do opisu dynamiki populacji.

Rozważa się niskończenie mały przedział czasu (t, t + dt) oraz zakłada się, że prawdopodobieństwa zdarzeń zależne od $(dt)^i$ dla $i > 1$ są pomijalnie małe. Wówczas prawdopodobieństwo śmierci osobnika w przedziale (t, t + dt) wynosi $\frac{\lambda}{2}$ dt, podczas gdy prawdopodobieństwo mutacji na lokusie 1 lub 2 jest równe odpowiednio ν_1 dt i ν_2 dt. Zapis

$$P_{n,m}(t) = P_{(n_1,n_2)(m_1,m_2)}(t) = \Pr[\text{haplotyp}_1 = (n_1, n_2), \text{haplotyp}_2 = (m_1, m_2); \text{dla chwili } t] \quad (2)$$

oznacza prawdopodobieństwo wystąpienia haplotypów (n_1, n_2) i (m_1, m_2) w chwili t dla osobników 1 i 2. Symbol δ_{ij} oznacza, iż dany składnik równania występuje jedynie dla przypadku, gdy i - j (identical by state). Natomiast $\Delta_{i,j}$ określa składnik, który pojawia się w przypadku, gdy chromosony i oraz j są i b.d (identical by descent).

Proponowane równania mają następującą postać:

$$\begin{aligned}
P_{(n_1,n_2)(m_1,m_2)}(t+dt) = {} & [1 - \lambda dt - 2(\nu_1 + \nu_2)\delta t] P_{(n_1,n_2)(m_1,m_2)}(t) + \\
& + \nu_1 dt \left[\sum_{i_1} \varphi_{1i_1} P_{(n_1-i_1,n_2)(m_1,m_2)}(t) + \sum_{j_1} \varphi_{1j_1} P_{(n_1,n_2)(m_1-j_1,m_2)}(t) \right] + \\
& + \nu_2 dt \left[\sum_{i_2} \varphi_{2i_2} P_{(n_1,n_2-i_2)(m_1,m_2)}(t) + \sum_{j_2} \varphi_{2j_2} P_{(n_1,n_2)(m_1,m_2-j_2)}(t) \right] + \\
& + \frac{\lambda}{2} dt (1-r) \frac{1}{2N} \left(\sum_{j_1,j_2} P_{(n_1,n_2)(j_1,j_2)}(t) \right) \delta_{n_1 m_1} \delta_{n_2 m_2} \Delta_{1,2} + \\
& + \frac{\lambda}{2} dt (1-r) \frac{1}{2N} \left(\sum_{i_1 i_2} P_{(i_1,i_2)(m_1,m_2)}(t) \right) \delta_{m_1 n_1} \delta_{m_2 n_2} \Delta_{2,1} + \\
& + \lambda dt (1-r) \left(1 - \frac{1}{2N} \right) P_{(n_1,n_2)(m_1,m_2)}(t) + \\
& + \frac{\lambda}{2} dt r \frac{1}{2N} \frac{1}{2} \left(\sum_{i_1,i_2,k_1} P_{(i_1,i_2)(m_1,m_2)(k_1,n_2)}(t) \right) \delta_{m_1 n_1} \Delta_{2,1} +
\end{aligned}$$

$$
\begin{aligned}
&+\frac{\lambda}{2}dtr\frac{1}{2N}\frac{1}{2}\left(\sum_{i_1,i_2,k_2} P_{(i_1,i_2)(m_1,m_2)(n_1,k_2)}(t)\right)\delta_{m_2n_2}\Delta_{2,1}+\\
&+\frac{\lambda}{2}dtr\frac{1}{2N}\frac{1}{2}\left(\sum_{j_1,j_2,k_1} P_{(n_1,n_2)(j_1,j_2)(k_1,m_2)}(t)\right)\delta_{n_1m_1}\Delta_{1,2}+\\
&+\frac{\lambda}{2}dtr\frac{1}{2N}\frac{1}{2}\left(\sum_{j_1,j_2,k_2} P_{(n_1,n_2)(j_1,j_2)(m_1,k_2)}(t)\right)\delta_{n_2m_2}\Delta_{1,2}+\\
&+\frac{\lambda}{2}dtr\frac{1}{2N}\left(\sum_{k_1,k_2} P_{(n_1,n_2)(m_1,m_2)(k_1,k_2)}(t)\right)\Delta_{1,3}+\\
&+\frac{\lambda}{2}dtr\frac{1}{2N}\frac{1}{2}\left(\sum_{j_2,k_1,k_2} P_{(n_1,n_2)(m_1,j_2)(k_1,k_2)}(t)\right)\delta_{n_2m_2}\Delta_{1,3}+\\
&+\frac{\lambda}{2}dtr\frac{1}{2N}\frac{1}{2}\left(\sum_{j_1,k_1,k_2} P_{(n_1,n_2)(j_1,m_2)(k_1,k_2)}(t)\right)\delta_{n_1m_1}\Delta_{1,3}+\\
&+\frac{\lambda}{2}dtr\frac{1}{2N}\left(\sum_{k_1,k_2} P_{(n_1,n_2)(m_1,m_2)(k_1,k_2)}(t)\right)\Delta_{2,3}+\\
&+\frac{\lambda}{2}dtr\frac{1}{2N}\frac{1}{2}\left(\sum_{i_2,k_1,k_2} P_{(n_1,i_2)(m_1,m_2)(k_1,k_2)}(t)\right)\delta_{m_2n_2}\Delta_{2,3}+\\
&+\frac{\lambda}{2}dtr\frac{1}{2N}\frac{1}{2}\left(\sum_{i_1,k_1,k_2} P_{(i_1,n_2)(m_1,m_2)(k_1,k_2)}(t)\right)\delta_{m_1n_1}\Delta_{2,3}-\\
&-\frac{\lambda}{2}dtr\frac{1}{2N^2}\left(\sum_{i_1,i_2,k_1,k_2} P_{(i_1,i_2)(m_1,m_2)(k_1,k_2)}(t)\right)\delta_{m_1n_1}\delta_{m_2n_2}\Delta_{2,1}\Delta_{2,3}-\\
&-\frac{\lambda}{2}dtr\frac{1}{2N^2}\left(\sum_{j_1,j_2,k_1,k_2} P_{(n_1,n_2)(j_1,j_2)(k_1,k_2)}(t)\right)\delta_{n_1m_1}\delta_{n_2m_2}\Delta_{1,2}\Delta_{1,3}+\\
&+\frac{\lambda}{2}dtr\left[1-\frac{1}{(2N)^2}-3\frac{2N-1}{(2N)^2}\right]\frac{1}{2}\left(\sum_{i_2,k_1} P_{(n_1,i_2)(m_1,m_2)(k_1,n_2)}(t)\right)+\\
&+\frac{\lambda}{2}dtr\left[1-\frac{1}{(2N)^2}-3\frac{2N-1}{(2N)^2}\right]\frac{1}{2}\left(\sum_{i_1,k_2} P_{(i_1,n_2)(m_1,m_2)(n_1,k_2)}(t)\right)+\\
&+\frac{\lambda}{2}dtr\left[1-\frac{1}{(2N)^2}-3\frac{2N-1}{(2N)^2}\right]\frac{1}{2}\left(\sum_{j_2,k_1} P_{(n_1,n_2)(m_1,j_2)(k_1,m_2)}(t)\right)+\\
&+\frac{\lambda}{2}dtr\left[1-\frac{1}{(2N)^2}-3\frac{2N-1}{(2N)^2}\right]\frac{1}{2}\left(\sum_{j_1,k_2} P_{(n_1,n_2)(j_1,m_2)(m_1,k_2)}(t)\right), \qquad (3)
\end{aligned}
$$

gdzie:

- $1-\lambda dt-(2\nu_1+2\nu_2)\,dt$ określa prawdopodobieństwo, że nie wystąpi śmierć żadnego z osobników oraz nie zajdzie mutacja na żadnym z lokusów;

- $\nu_1\varphi_{1,j_1}dt$ jest prawdopodobieństwem, że mutacja o wymiarze i_1 wystąpi na lokusie 1 osobnika 1;
- $\nu_1\varphi_{1,j_1}dt$ jest prawdopodobieństwem, że mutacja o wymiarze j_1 wystąpi na lokusie 1 osobnika 2;
- $\nu_2\varphi_{2,i_2}dt$ jest prawdopodobieństwem, że mutacja o wymiarze i_2 wystąpi na lokusie 2 osobnika 1;
- $\nu_2\varphi_{2,j_2}dt$ jest prawdopodobieństwem, że mutacja o wymiarze j_2 wystąpi na lokusie 2 osobnika 2;
- $\frac{\lambda}{2}dt(1-r)\frac{1}{2N}$ jest prawdopodobieństwem, że jeden z osobników (1 lub 2) umrze, wylosowany haplotyp odradzającego się osobnika będzie taki sam jak tego, który nie umarł (odpowiednio 2 lub 1) oraz nie wystąpi mutacja;
- $\frac{\lambda}{2}dt\frac{1}{2N}r$ jest prawdopodobieństwem, że jeden z osobników (1 lub 2) umrze, wylosowany haplotyp odradzającego się osobnika będzie taki sam jak tego, który nie umarł (odpowiednio 2 lub 1) oraz wystąpi mutacja;
- $\frac{\lambda}{2}dt(1-r)(1-\frac{1}{2N})$ jest prawdopodobieństwem, że jeden z osobników (1 lub 2) umrze, wylosowany haplotyp odradzającego się osobnika będzie różny od tego, który nie umarł (odpowiednio 2 lub 1) oraz nie wystąpi mutacja.

Zdefiniujmy funkcję tworzącą rozkładu łącznego dla wymiaru alleli na obu lokusach dla obu osobników.

$$P(s,u;t) = P(s_1,s_2,u_1,u_2;t) = \sum_{n_1}\sum_{n_2}\sum_{m_1}\sum_{m_2} P_{n,m}(t)s_1^{n_1}s_2^{n_2}u_1^{m_1}u_2^{m_2} = E\left(s_1^{X_1}s_2^{X_2}u_1^{Y_1}u_2^{Y_2}\right), \tag{4}$$

gdzie $[X_1(t), X_2(t)]$ i $[Y_1(t), Y_2(t)]$ określają odpowiednio haplotypy osobników 1 i 2. Argument t będzie w dalszych rozważaniach pomijany tam, gdzie nie spowoduje to jakichkolwiek niejednoznaczności. Sprowadzając równanie (3) do równania różniczkowego w dziedzinie czasu, następnie mnożąc obustronnie przez $s_1^{n_1}s_2^{n_2}u_1^{m_1}u_2^{m_2}$ i sumując po n_1, n_2, m_1, m_2 otrzymujemy następującą zależność:

$$\begin{aligned}\frac{\partial P(s_1,s_2,u_1,u_2,t)}{\partial t} = & -2(\nu_1+\nu_2)P(s_1,s_2,u_1,u_2,t) + \\ & +\nu_1[\varphi_1(s_1)+\varphi_1(u_1)]P(s_1,s_2,u_1,u_2,t) + \\ & +\nu_2[\varphi_2(s_2)+\varphi_2(u_2)]P(s_1,s_2,u_1,u_2,t) + \\ & +\left[\frac{\lambda}{2N}(1-r)-2r\frac{\lambda}{(2N)^2}\right]P(s_1u_1,s_2u_2,1,1,t) + \\ & +\left[\lambda r\left(\frac{1}{N}-1\right)-\frac{\lambda}{2N}\right]P(s_1,s_2,u_1,u_2,t) + \\ & +\frac{\lambda}{2}r\frac{1}{2N}[P(1,s_2,s_1u_1,u_2,t)+P(s_1,1,u_1,s_2u_2,t)] +\end{aligned}$$

$$+\frac{\lambda}{2}r\frac{1}{2N}[P(s_1u_1.s_2,1,u_2,t)+P(s_1,s_2u_2.u_1,1,t)]+$$
$$+\frac{\lambda}{2}r\left[1-\frac{1}{(2N)^2}-3\frac{2N-1}{(2N)^2}\right][P(s_1.1.u_1,u_2,1,s_2.t)+$$
$$+P(s_1,s_2,u_1,1,1,u_2,t)] \tag{5}$$

Ze względu na skomplikowaną postać powyższego równania oraz zależność jego rozwiązania od znajomości rozkładów trójek chromosomów w analizowanej populacji nie można w świetle dzisiejszej wiedzy literaturowej podać jawnej postaci funkcji tworzącej P (s_1, s_2, u_1, u_2, t). Jednakże możliwe jest przeanalizowanie przypadków szczególnych, wyjaśniających niektóre z aspektów badanego procesu.

4. Przypadki szczególne

Poprzez odpowiednie podstawienie zmiennych do równania (5) otrzymuje się wyrażenia określające dynamikę zmian w czasie interesujących wielkości. Istotnym, z punktu widzenia genetyki populacyjnej, zagadnieniem jest analiza dynamiki zmian zarówno rozkładu alleli (a tym samym liczby powtórzeń sekwencji postawowej) w badanej populacji, jak i różnic wielkości alleli dla odpowiednich lokusów.

4.1. Rozkłady graniczne

- **Pojedynczy chromosom, pojedynczy lokus [X_1]**
  Podstawienie $s_2 = u_1 = u_2 = 1$ oraz $s_1 = s$ umożliwia podanie równania dynamiki funkcji tworzącej rozkładu alleli na pojedynczym chromosomie. Ma ono postać:

$$\frac{\partial P(s_1,1,1,1.t)}{\partial t}=\frac{\partial P_1(s.t)}{\partial t}=-\nu_1[1-\varphi_1(s)]P_1(s,t) \tag{6}$$

  Uzyskana tą drogą funkcja tworząca jest następująca:

$$P_1(s,t)=P_1(s,0)e^{-\nu_1[1-\varphi_1(s)]t} \tag{7}$$

- **Para chromosomów, pojedynczy lokus [X_1, Y_1]**
  Funkcja tworząca rozkładu par [X_1, Y_1] uzyskana może być poprzez podstawienie w równaniu (5) $s_2 = u_2 = 1$ oraz $s_1 = s$ i $u_1 = u$. Wówczas:

$$\frac{\partial P(s_1,1,u_1,1,t)}{\partial t}=\frac{\partial P_2(s,u,t)}{\partial t}=$$
$$\left\{-2\nu_1\left[1-\frac{1}{2}(\varphi_1(s)+\varphi_1(u))\right]-\frac{\lambda(N-r)}{2N^2}\right\}P_2(s,u,t)+$$
$$+\frac{\lambda(N-r)}{2N^2}P(s_1u_1,1,1,1,t) \tag{8}$$

Znajomość funkcji tworzącej dla pojedynczego lokusa na jednym chromosomie P_1 (s, u, t) pozwala na wyliczenie funkcji tworzącej rozkładu par. Podstawiając zależność (7) do równania (8) otrzymujemy:

$$\frac{\partial P_2(s,u,t)}{\partial t} = \left\{-2\nu_1\left[1-\frac{1}{2}(\varphi_1(s)+\varphi_1(u))\right]-\frac{\lambda(N-r)}{2N^2}\right\}P_2(s,u,t)+ \\ +\frac{\lambda(N-r)}{2N^2}P_1(su,0)e^{-\nu_1[1-\varphi_1(su)]t} \tag{9}$$

Rozwiązując powyższe równanie uzyskuje się następującą postać funkcji P_2

$$P_2(s,u,t) = P_1(su,0)e^{-2\nu_1\left[1-\frac{\varphi_1(s)+\varphi_1(u)}{2}\right]t}e^{-\frac{\lambda}{2N^*}t}+ \\ +P_1(su,0)\frac{\lambda}{2N^*}\frac{e^{-2\nu_1\left[1-\frac{\varphi_1(s)+\varphi_1(u)}{2}\right]t}e^{-\frac{\lambda}{2N^*}t}}{-\nu_1[1-\varphi_1(su)]+2\nu_1\left[1-\frac{\varphi_1(s)+\varphi_1(u)}{2}\right]+\frac{\lambda}{2N^*}} \\ \cdot\left(e^{-\nu_1[1-\varphi_1(su)]t}e^{2\nu_1\left[1-\frac{\varphi_1(s)+\varphi_1(u)}{2}\right]t}e^{\frac{\lambda}{2N^*}t}-1\right), \tag{10}$$

przy czym:

$$\psi_1(s) := \frac{\varphi_1(s)+\varphi_1(s^{-1})}{2} \tag{11}$$

oznacza funkcję tworzącą prawdopodobieństwa zmiennej losowej odpowiadającej symetryzowanej wielkości mutacji na 1 lokusie. Natomiast

$$N^* = \frac{N^2}{N-r} \approx N \ (\ dla\ N \gg 1) \tag{12}$$

określa tzw. uogólnioną liczebność populacji (pojęcie wprowadzone między innymi przez J. Crow i M. Kimura w [3]).

- **Para lokusów, pojedynczy chromosom [X_1, X_2]**

Funkcję tworzącą takiego rozkładu uzyskać można w wyniku podstawienia $u_1 = u_2 = 1$ do równania (5). Po niewielkich przekształceniach otrzymuje się:

$$\frac{\partial P(s_1,s_2,1,1,t)}{\partial t} = \frac{\partial P_H(s_1,s_2,t)}{\partial t} = -(\nu_1+\nu_2)P_H(s_1,s_2,t)+ \\ +[\nu_1\varphi_1(s_1)+\nu_2\varphi_2(s_2)]P_H(s_1,s_2,t)- \\ -\frac{\lambda}{2}r\left(1-\frac{1}{2N}+\frac{1}{2N^2}\right)P_H(s_1,s_2,t)+ \\ +\frac{\lambda}{2}r\left(1-\frac{1}{2N}+\frac{1}{2N^2}\right)P_D(s_1,s_2,t), \tag{13}$$

gdzie:

$$P_D(s_1,s_2,t) = P(s_1,1,1,s_2,t) \tag{14}$$

jest funkcją tworzącą rozkładu par skrośnych [X_1, Y_2].

Sprzężone do równania (13) równanie dla funkcji tworzącej par skrośnych ma postać:

$$\begin{aligned}\frac{\partial P(s_1,1,1.s_2,t)}{\partial t} &= \frac{\partial P_D(s_1,s_2,t)}{\partial t} = -(\nu_1+\nu_2)P_D(s_1,s_2,t) + \\ &+[\nu_1\varphi_1(s_1)+\nu_2\varphi_2(s_2)]P_D(s_1,s_2,t) - \\ &-\frac{\lambda}{2}\left(r+\frac{1}{N}-3r\frac{1}{2N}-r\frac{1}{2N^2}\right)P_D(s_1,s_2,t) + \\ &+\frac{\lambda}{2}\left(r+\frac{1}{N}-3r\frac{1}{2N}-r\frac{1}{2N^2}\right)P_H(s_1,s_2,t)\end{aligned} \tag{15}$$

Rozwiązanie układu równań (13) (15) daje w efekcie funkcje tworzące rozkładu par zarówno wzdłużnych, jak i skrośnych. Mają one następującą postać:

$$\begin{aligned}P_D(s_1,s_2,t) &= P_D(s_1,s_2,0)e^{-(A_1+A_2)t}\left[B_1e^{-\lambda r(1-\frac{1}{2N})t}e^{-\lambda(1-r)\frac{1}{2N}t}-B_2\right]+ \\ &+P_H(s_1,s_2,0)e^{-(A_1+A_2)t}\left[B_1-B_1e^{-\lambda r(1-\frac{1}{2N})t}e^{-\lambda(1-r)\frac{1}{2N}t}\right]\end{aligned} \tag{16}$$

$$\begin{aligned}P_H(s_1,s_2,t) &= P_D(s_1,s_2,0)e^{-(A_1+A_2)t}\left[B_2e^{-\lambda r(1-\frac{1}{2N})t}e^{-\lambda(1-r)\frac{1}{2N}t}-B_2\right]+ \\ &+P_H(s_1,s_2,0)e^{-(A_1+A_2)t}\left[B_1-B_2e^{-\lambda r(1-\frac{1}{2N})t}e^{-\lambda(1-r)\frac{1}{2N}t}\right]\end{aligned} \tag{17}$$

gdzie

$$B_1 = \frac{-\lambda(1-r)\frac{1}{2N}-\frac{\lambda}{2}r(1-\frac{1}{2N})+\lambda r\frac{1}{(2N)^2}}{-\lambda r(1-\frac{1}{2N})-\lambda(1-r)\frac{1}{2N}} \tag{18}$$

$$B_2 = \frac{\frac{\lambda}{2}r\left[1-\frac{1}{2N}+\frac{1}{2N^2}\right]}{-\lambda r(1-\frac{1}{2N})-\lambda(1-r)\frac{1}{2N}} \tag{19}$$

$$A_1 = \nu_1[1-\varphi_1(s_1)] \qquad A_2 = \nu_2[1-\varphi_2(s_2)] \tag{20}$$

4.2. Zależności funkcyjne

4.2.1. Różnica liczby powtórzeń na poszczególnych lokusach

- **Pojedynczy lokus**

Zdefinujmy funkcję tworzącą rozkładu różnic $[X_1 - Y_1]$ jako:

$$R_1(s,t) = P(s_1,s_2,u_1,u_2,t)\,|_{s_1=s,s_2=1,u_1=s^{-1},u_2=1}, \tag{21}$$

wówczas:

$$\frac{\partial R_1(s,t)}{\partial t} = -2\nu_1[1-\Psi_1(s)]R_1(s,t) - \frac{\lambda}{2N^*}R_1(s,t) + \frac{\lambda}{2N^*} \tag{22}$$

Równanie to jest dla r = 0 zgodne z modelem uzyskanym przy analiie procesu bez udziału rekombinacji [6].

$$R_1(s,t) = e^{-2\nu_1[1-\psi_1(s)]t}e^{-\frac{\lambda}{2N^*}t}\left\{R_1(s,0) - \frac{\lambda}{2N^*}\frac{1}{2\nu_1[1-\psi_1(s)]+\frac{\lambda}{2N^*}}\right\} + \frac{\lambda}{2N^*}\frac{1}{2\nu_1[1-\psi_1(s)]+\frac{\lambda}{2N^*}} \tag{23}$$

W analogiczny sposób definuje się funkcję;

$$R_2(u,t) = P(s_1,s_2,u_1,u_2,t)\,|_{s_1=1,s_2=u,u_1=1,u_2=u^{-1}} \tag{24}$$

- **Para lokusów**

Funkcję tworzącą rozkładu różnic $[X_1 - Y_1, X_2 - Y_2]$ definuje się jako:

$$R(s_1,s_2,t) = P(s_1,s_2,s_1^{-1},s_2^{-1},t) = \sum_{n_1}\sum_{n_2}\sum_{m_1}\sum_{m_2} P_{n,m}(t)s_1^{n_1-m_1}s_2^{n_2-m_2} = E\left(s_1^{X_1-Y_1}s_2^{X_2-Y_2}\right) \tag{25}$$

Stosując powyższe podstawienie do równania (5) otrzymuje się:

$$\begin{aligned}\frac{\partial R(s_1,s_2,t)}{\partial t} = {} & -2\{\nu_1[1-\psi_1(s_1)]+\nu_2[1-\psi_2(s_2)]\}R(s_1,s_2,t) + \\ & + \left[\lambda r\left(\frac{1}{N}-1\right)-\frac{\lambda}{2N}\right]R(s_1,s_2,t) + \\ & + \lambda r\frac{1}{2N}[R_1(s_1,t)+R_2(s_2,t)] + \frac{\lambda}{2N}(1-r) - 2r\frac{\lambda}{(2N)^2} + \\ & + \frac{\lambda}{2}r\left[1-\frac{1}{(2N)^2}-3\frac{2N-1}{(2N)^2}\right]P(s_1,1,s_1^{-1},s_2^{-1},1,s_2,t) + \\ & + \frac{\lambda}{2}r\left[1-\frac{1}{(2N)^2}-3\frac{2N-1}{(2N)^2}\right]P(s_1,s_2,s_1^{-1},1,1,s_2^{-1},t)\end{aligned} \tag{26}$$

Rozwiązanie równania (26) wymaga znajomości pewnych granicznych rozkładów trójek chromosomów i nie może być znalezione w sposób bezpośredni.

4.2.2. Suma liczby powtórzeń na określonym lokusie

Problem poszukiwania rozkładu sum $[X_1 + Y_1]$ traktować można jako zagadnienie poszukiwania następującej funkcji:

$$S_1(s,t) = P(s,1,s,1,t)$$

Korzystając z równania (5) oraz podstawienia $s_1 = s$, $s_2 = 1$, $u_1 = s$, $u_2 = 1$ otrzymujemy:

$$\frac{\partial S_1(s,t)}{\partial t} = -2\nu_1[1-\varphi_1(s)]S_1(s,t) + \left[-\frac{\lambda}{2N} + \lambda r\frac{1}{2N^2}\right]S_1(s,t) + \\ + \left[\frac{\lambda}{2N} - \lambda r\frac{1}{2N^2}\right]P_1(s^2,t) \tag{27}$$

Dzięki obliczonej wcześniej funkcji $P_1(s,t)$ możliwe jest podanie rozwiązania powyższego równania różniczkowego, które przyjmuje postać:

$$S_1(s,t) = S_1(s,0)e^{-2\nu_1[1-\varphi_1(s)]t}e^{-\frac{\lambda}{2N^2}t} + P_1(s^2,0)\frac{e^{-\nu_1[1-\varphi_1(s^2)]t}e^{-2\nu_1[1-\varphi_1(s)]t}e^{-\frac{\lambda}{2N^2}t}}{\nu_1 + \nu_1\varphi_1(s^2) - 2\nu_1\varphi_1(s) + \frac{\lambda}{2N^2}} \tag{28}$$

Tak zdefiniowana zmienna losowa ma istotne znaczenie w procesie szacowania współczynnika rekombinacji r. Ważną cechę tej zmiennej losowej podaje następujące twierdzenie:

Twierdzenie 1

Rozkład graniczny zmiennej standaryzowanej Z_1, której funkcja tworząca dana jest wzorem:

$$Z_1(s,t) = s^{-\frac{M_1(t)}{\sigma}}S_1(s,t), \tag{29}$$

gdzie:

$$M_1(t) = \frac{dS_1(s,t)}{ds}\Big|_{\substack{t=0\\ s=1}} = 2M_1(0) + 2\nu_1\varphi_1'(1)t \tag{30}$$

oraz dla $t \to \infty$ wariancja

$$V_1(t) \sim \sqrt{t} \tag{31}$$

jest przy $t \to \infty$ rozkładem normalnym

$$Z_1(t) \to N(0, 4\nu_1\left[\varphi_1'(1) + \varphi_1''(1)\right]) = N(0, 4\nu_1\psi_1''(1)) \tag{32}$$

Słuszność powyższego twierdzenia pokazuje następujące rozumowanie. Dokonując standaryzacji zmiennej losowej $S_1 = [X_1 + Y_1]$ otrzymujemy następującą zależność w dziedzinie funkcji tworzących:

$$Z_1(s,t) = s^{-\frac{2M_1(0)+2\nu_1\varphi_1'(1)t}{\sqrt{t}}}\{S_1(s^{\frac{1}{\sqrt{t}}},0)e^{-2\nu_1[1-\varphi_1(s^{\frac{1}{\sqrt{t}}})]t}e^{-\frac{\lambda}{2N^2}t} + \\ + P_1(s^{\frac{2}{\sqrt{t}}},0)\frac{\lambda}{2N^2}\frac{e^{-\nu_1[1-\varphi_1(s^{\frac{2}{\sqrt{t}}})]t}}{\nu_1[1+\varphi_1(s^{\frac{2}{\sqrt{t}}}) - 2\varphi_1(s^{\frac{1}{\sqrt{t}}})] + \frac{\lambda}{2N^2}} - \\ - P_1(s^{\frac{2}{\sqrt{t}}},0)\frac{\lambda}{2N^2}\frac{e^{-2\nu_1[1-\varphi_1(s^{\frac{1}{\sqrt{t}}})]t}e^{-\frac{\lambda}{2N^2}t}}{\nu_1[1+\varphi_1(s^{\frac{2}{\sqrt{t}}}) - 2\varphi_1(s^{\frac{1}{\sqrt{t}}})] + \frac{\lambda}{2N^2}}\} \tag{33}$$

Dla $t \to \infty$ równanie (33) przyjmuje postać:

$$Z_1(s,\infty) \sim s^{-2\nu_1\varphi_1'(1)\sqrt{t}} e^{-\nu_1[1-\varphi_1(s^{\frac{1}{\sqrt{t}}})]t} \tag{34}$$

Rozwijając funkcję $\varphi_1(s^{\frac{1}{\sqrt{t}}})$ wokół punktu 1 + j0 oraz podstawiając $s = e^{j\omega}$ otrzymujemy:

$$\begin{aligned} -\nu_1[1-\varphi_1(s^{\frac{1}{\sqrt{t}}})]t &= \nu_1 t\left\{\varphi_1'(1)\frac{2j\omega}{\sqrt{t}} - 2\varphi_1'(1)\frac{\omega^2}{t} - 2\varphi_1''(1)\frac{\omega^2}{t}\right\} = \\ & 2j\nu_1\varphi_1'(1)\omega\sqrt{t} - 2\nu_1\varphi_1'(1) - 2\nu_1\varphi_1''(1)\omega^2 \end{aligned} \tag{35}$$

Zatem:

$$\lim_{t\to\infty} Z_1(s,t) = e^{-2\nu_1[\varphi_1'(1)+\varphi_1''(1)]\omega^2} = e^{-2\nu_1\psi_1''(1)\omega^2}\ , \tag{36}$$

co kończy dowód.

5. Momenty zmiennych losowych

Z całej grupy możliwych do obliczenia momentów zdefiniowanych w pracy zmiennych losowych na szczególną uwagę zasługuje moment faktoralny.

$$\begin{aligned} Cov[(X_1 - Y_1)(X_2 - Y_2)] &= Cov[X_1X_2] - Cov[X_1Y_2] - Cov[Y_1X_2] + Cov[Y_1Y_2] = \\ & E[X_1X_2] - E[X_1]E[X_2] - E[X_1Y_2] + E[X_1]E[Y_2] - \\ & -E[Y_1X_2] + E[Y_1]E[X_2] + E[Y_1Y_2] - E[Y_1]E[Y_2] \\ &= 2E[X_1X_2] - 2E[X_1Y_2] \end{aligned} \tag{37}$$

Wyliczając składnik $E[X_1 X_2]$ należy wykorzystać znalezioną wcześniej funkcję tworzącą rozkładu par $P_H(s_1, s_2, t)$. Podwójne jej zróżniczkowanie i podstawienie $s_1 = s_2 = 1$ daje w rezultacie następującą zależność:

$$E[X_1X_2] = E[X_1Y_2] + [C_H(0) - C_D(0)]e^{-(C_1+C_2)t}\ , \tag{38}$$

przy czym

$$\begin{aligned} C_H(0) &= \frac{\partial^2 P_H(s_1,s_2,0)}{\partial s_1 \partial s_2}\Big|_{\substack{s_1=1\\s_2=1}} \\ C_D(0) &= \frac{\partial^2 P_D(s_1,s_2,0)}{\partial s_1 \partial s_2}\Big|_{\substack{s_1=1\\s_2=1}} \end{aligned} \tag{39}$$

oraz

$$C_1 = \lambda r(1 - \frac{1}{2N})$$
$$C_2 = \lambda(1 - r)\frac{1}{2N} \tag{40}$$

Ostatecznie:

$$Cov[X_1 - Y_1, X_2 - Y_2] = 2[C_H(0) - C_D(0)]\, e^{-\lambda r(1-\frac{1}{2N})t} e^{-\lambda(1-r)\frac{1}{2N}t} \tag{41}$$

W przypadku gdy r = 0 (dla bardzo blisko położonych lokusów), zależność (41) przyjmuje postać:

$$Cov[X_1 - Y_1, X_2 - Y_2] = 2[C_H(0) - C_D(0)]\, e^{-\frac{\lambda}{2N}t}\,, \tag{42}$$

natomiast dla $r = \frac{1}{2}$

$$Cov[X_1 - Y_1, X_2 - Y_2] = 2[C_H(0) - C_D(0)]\, e^{-\frac{\lambda}{2}t} \tag{43}$$

Oznacza to, że dla mocno oddalonych lokusów prędkość zanikania kowariancyjnego warunku początkowego będzie znacznie większa niż dla lokusów ulokowanych w relatywnie niewielkiej odległości. Analogiczne wnioski wysunąć można, analizując zachowanie się współczynnika korelacji wyliczanego z próby. Może on być również miarą oceny poziomu rekombinacji w populacji, a tym samym pośrednio wzajemnego ulokowania lokusów. Wiadomo, że:

$$\rho_{X_1X_2} = \frac{|Cov[X_1X_2]|}{\sqrt{Var[X_1]Var[X_2]}}\,, \tag{44}$$

stąd

$$\rho_{(X_1-Y_1)(X_2-Y_2)} = \frac{2\left|[C_H(0) - C_D(0)]\, e^{-\frac{\lambda}{2N}t} e^{-\frac{\lambda}{2}r(1+\frac{1}{2N})t}\right|}{\sqrt{V_{10}e^{-\frac{\lambda}{2N}t} + \frac{4\nu_1\psi_1''(1)N}{\lambda}\left(1 - e^{-\frac{\lambda}{2N}t}\right)}\sqrt{V_{20}e^{-\frac{\lambda}{2N}t} + \frac{4\nu_2\psi_2''(1)N}{\lambda}\left(1 - e^{-\frac{\lambda}{2N}t}\right)}} \tag{45}$$

dla r = 0

$$\rho_{(X_1-Y_1)(X_2-Y_2)} = \frac{2|C_H(0) - C_D(0)|}{\sqrt{V_{10} + \frac{4\nu_1\psi_1''(1)N}{\lambda}\left(e^{\frac{\lambda}{2N}t} - 1\right)}\sqrt{V_{20} + \frac{4\nu_2\psi_2''(1)N}{\lambda}\left(e^{\frac{\lambda}{2N}t} - 1\right)}} \tag{46}$$

zmierza do 0 dla $t \to \infty$ wolniej niż dla $r \to \frac{1}{2}$.

6. Podumowanie

W pracy przedstawiono model matematyczny dynamiki populacji, w której w procesie rowoju występują niezależnie trzy procesy: proces narodzin i śmierci, mutacja i rekombinacja. Uwzględnienie na etapie formułowania zadania wszystkich trzech zjawisk prowadzi wprawdzie do bardzo skomplikowanego modelu, którego jawne rozwikłanie nie jest możliwe, niemniej jednak przeanalizowanie pewnych przypadków szczególnych nakłania do wniosków w istotny sposób ułatwiających tworzenie mapy genetycznej populacji. Uzyskane np. drogą uproszczeń rozkłady graniczne pokrywają się ze znanymi z wcześniejszych publikacji wynikami, co wskazuje na poprawność przeprowadzonego rozumowania. Autorzy planują w najbliższej przyszłości zweryfikowanie uzyskanych rezultatów teoretycznych badaniami symulacyjnymi i na rzeczywistych danych dostępnych w specjalistycznych, internetowych bazach danych.

LITERATURA

1. Carter T., Falconer D.: Stocks for detecting linkage in the mouse and the theory of their design. Journal of Genetics 1951, 50: 307-323.
2. Chakravarti A., Lasher L.K., Reefer J.E.: A maximum likelihood method for estimating genome length using genetic linkage data. Genetics 1991, 128: 175-182.
3. Crow J., Kimura M.: An introduction to population genetics theory. Harper & Row, New York 1970.
4. Goldstein D., Linares A.R., Feldman M., Cavalli-Sforza L.: An evaluation of genetic distances for use with microsatellite loci. Genetics 1995b, 139: 463-471.
5. Haldane J.B.S.: The combination of linkage values, and the calculation of distance between the loci of linked factors" Genetics 1919, 8: 299-309.
6. Kimmel M., Chakraborty R., Stivers D., Deka R.: Dynamics of repeat polymorphism under forward-backward mutation model: Within- and between- population variability at microsatellite loci" Genetics 1996.
7. Kosambi D.D.: „The estimation of map distance from recombination values" Annals of Eugenics 1944, 12: 172-175.
8. „Meiosis and Genetic Recombination", Access Excellence WWW site, Genentech Inc.
9. Moran P.A.: Wandering distribution and the electrophoretic profile" Theor. Pop. Biol. 1975, 8: 318-330.
10. „Mutation of Chromosome During Replication", Access Excellence WWW site, Genentech Inc.

11. Nakamura Y., Leppert M., O'Connell P., Wolf R., Holm T., Culver M., Martin C., Fujimoto E., Hoff M., Kumlin E., White R.: Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 1987, 235: 1616-1622.
12. Ott J.: Estimation of recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. American Journal of Human Genetics 1974, 26: 588-597.
13. Ott J.: Analysis of human genetic linkage Johns Hopkins University Press, Baltimore 1991.
14. Rao D.C., Morton N.E., Lindsten J., Hulten M., Yee S.: A mapping function for man. Human Heredity 1977, 27: 99-104.
15. Tavare S.: Calibrating the clock: Using stochastic processes to measure the rate of evolution w Lander E.S., Waterman M. (eds.) Calculating the secrets of Life National Academy Press, Washington 1995, 114-152.

Recenzent: Prof.dr hab. Ryszard Tadeusiewicz

Wpłynęło do Redakcji 9.12.1997 r.

Abstract

The mathematical model of two-linked microsatellite loci evolution is presented. The described population model is a time-continuous Markov chain. It is transformed into differential equations' system of probability generating function. The obtained model is very complicated because of considering all possible events, and including three genetic processes: mutation, genetic drift, and recombination. Appropriate specification of hypotheses leads to the well-known formulae. The equations could be successfully applied to the stochastic analysis of experimental data.