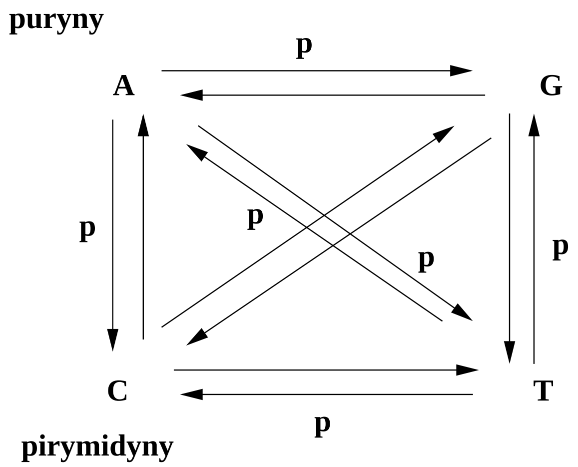
Wykład II: modele ewolucji sekwencji

Anna Gambin
10.10.2002

1 Substytucje nukleotydów w sekwencjach DNA

Podstawowym zdarzeniem w ewolucji sekwencji DNA jest zamiana jednego nukleotydu w inny. Zaproponowano wiele modeli matematycznych opisujących ten proces. Poznamy kilka najprostszych, które pozwalają szacować zmiany zachodzące w sekwencji, a następnie pokażemy jak stosować je do wyznaczania ewolucyjnej odległości pomiędzy dwiema sekwencjami (więcej dla chętnych w [2]).

1.1 Model jednoparametrowy: Jukes & Cantor 1969



Rysunek 1: Jednoparametrowy model Jukes & Cantor

W modelu tym zakłada się, że każdy z nukleotydów może z jednakowym prawdopodobieństwem zmutowac w każdy inny. Dodatkowo zakłada się, że zdarzenia mutacji dla różnych miejsc w sekwencji są od siebie niezależne (oczywiście jest to duże uproszczenie). Oznaczmy prawdopodobieństwo mutacji przez p. Rozważmy pozycję w sekwencji zawierającą A w chwili 0. Policzymy prawdopodobieństwo $P_{A(t)}$ że po t krokach na rozważanej pozycji nadal mamy A. Czyli chcemy policzyć z jakim prawdopodobieństwem prosty łańcuch Markowa (rysunek 1), który modeluje nasz proces, po t krokach znajdzie się w stanie A.

Mamy $P_{A(0)} = 1$, oraz:

$$\begin{aligned} P_{A(1)} &= 1 - 3p \\ P_{A(2)} &= (1-3p)P_{A(1)} + p[1 - P_{A(1)}] \\ P_{A(t+1)} &= (1-3p)P_{A(t)} + p[1 - P_{A(t)}] \quad \forall t \end{aligned}$$

Przekształcając dalej otrzymujemy:

$$P_{A(t+1)} - P_{A(t)} = -3pP_{A(t)} + p[1 - P_{A(t)}] = -4pP_{A(t)} + p$$

Wygodniej jest potraktować rozważany łańcuch jako proces ciągły i interpretować $\Delta P_{A(t)} = P_{A(t+1)} - P_{A(t)}$ jako częstość zmian w odcinku czasu t. Otrzymujemy równanie różniczkowe pierwszego rzędu:

$$\frac{dP_{A(t)}}{dt} = -4pP_{A(t)} + p$$

Pamiętając, że $P_{A(0)} = 1$ otrzymujemy rozwiązanie:

$$P_{A(t)} = \frac{1}{4} + \left(P_{A(0)} - \frac{1}{4}\right)\exp^{-4pt} = \frac{1}{4} + \frac{3}{4}\exp^{-4pt}$$

Jeżeli wystartujemy z nukleotydu innego niż A mamy $P_{A(0)} = 0$ i prawdopodobieństwo znalezienia się w stanie A po t krokach wynosi:

$$P_{A(t)} = \frac{1}{4} - \frac{1}{4}\exp^{-4pt}$$

Niech $P_{AA(t)}$ oznacza prawdopodobieństwo stanu A po t krokach przy założeniu że wystartowaliśmy ze stanu A. Ponieważ w rozważanym modelu wszystkie cztery nukleotydy są równoważne mamy:

$$P_{GA(t)} = P_{CA(t)} = P_{TA(t)}$$

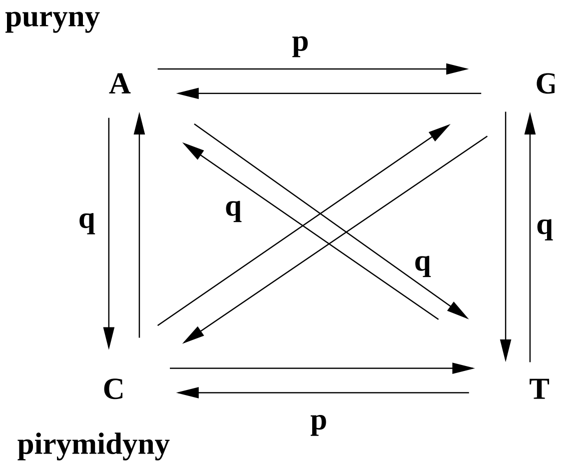
i ogólnie:

$$P_{ii(t)} = \frac{1}{4} + \frac{3}{4}\exp^{-4pt} \tag{1}$$

$$P_{ij(t)} = \frac{1}{4} - \frac{1}{4}\exp^{-4pt} \qquad \text{dla} \quad i \neq j \tag{2}$$

Równania (1) oraz (2) opisują zachowanie się danego miejsca w sekwencji w modelu jednoparametrowym. Patrząc na rysunek 1 zgadujemy natychmiast, że rozkład stacjonarny rozważanego łańcucha Markowa jest jednorodny: prawdopodobieństwo stacjonarne wynosi $\frac{1}{4}$ dla każdego ze stanów (wynika to z faktu że łańcuch jest symetryczny).

1.2 Model dwuparametrowy: Kimura 1980



Rysunek 2: Dwuparametrowy model Kimury

Założenia przyjęte w modelu jednoparametrowym są raczej nierealistyczne. Jak pamiętamy, prawdopodobieństwo tranzycji jest zazwyczaj uznawane za większe niż transwersji. Tę prawidłowość uwzględnia model zaproponowany przez Kimurę (rysunek 2). Niech p oznacza prawdopodobieństwo tranzycji (lub częstość zmian tranzycyjnych w odcinku czasu) oraz q prawdopodobieństwo transwersji. Używając wprowadzonych poprzednio oznaczeń prawdopodobieństwo, że zaobserwujemy nukleotyd A w danym miejscu wynosi:

$$P_{A(t+1)} = (1 - p - 2q)P_{A(t)} + q[P_{T(t)} + P_{C(t)}] + pP_{G(t)}$$

Dla pozostałych stanów łańcucha otrzymujemy następujące równania rekurencyjne:

$$
\begin{aligned}
P_{T(t+1)} =& qP_{A(t)} + (1-p-2q)P_{T(t)} + pP_{C(t)} + qP_{G(t)} \\
P_{C(t+1)} =& qP_{A(t)} + pP_{T(t)} + (1-p-2q)P_{C(t)} + qP_{G(t)} \\
P_{G(t+1)} =& pP_{A(t)} + qP_{T(t)} + qP_{C(t)} + (1-p-2q)P_{G(t)}
\end{aligned}
$$

Ustalając warunki początkowe możemy rozwiązać powyższy układ równań. Podobnie jak poprzednio mamy $P_{AA(t)} = P_{CC(t)} = P_{GG(t)} = P_{TT(t)}$, oznaczmy to prawdopodobieństwo przez $X_{(t)}$. Otrzymujemy:

$$X_{(t)} = \frac{1}{4} + \frac{1}{4}\exp^{-4qt} + \frac{1}{2}\exp^{-2(p+q)t}$$

Policzmy teraz prawdopodobieństwo $Y_{(t)}$, że początkowy nukleotyd i stan w jakim znajduje się łańcuch w chwili t różnią się od siebie tranzycją. Ponieważ prawdopodobieństwo zmian jest symetryczne mamy:

$$Y_{(t)} = P_{AG(t)} = P_{GA(t)} = P_{TC(t)} = P_{CT(t)}$$

Można pokazać, że:

$$Y_{(t)} = \frac{1}{4} + \frac{1}{4}\exp^{-4qt} - \frac{1}{2}\exp^{-2(p+q)t}$$

Podobnie prawdopodobieństwo $Z_{(t)}$, że nukleotyd w chwili t różni się od początkowego nukleotydu transwersją wyraża się wzorem:

$$Z_{(t)} = \frac{1}{4} - \frac{1}{4}\exp^{-4qt}$$

Ponieważ dla każdego stanu mamy dwie możliwe transwersje (a tylko jedną tranzycję) zachodzi:

$$X_{(t)} + Y_{(t)} + 2Z_{(t)} = 1$$

Rozkład stacjonarny łańcucha w modelu Kimury jest jak poprzednio jednorodny. Różnice pomiędzy opisanymi modelami dotyczą szybkości zbieżności i zachowania się funkcji $P_{ii(t)}, P_{ij(t)}, X_{(t)}, Y_{(t)}$ oraz $Z_{(t)}$. Można pokazać, że wartości $P_{ii(t)}, P_{ij(t)}, X_{(t)}$ i $Z_{(t)}$ dążą przy $t \to \infty$ monotonicznie do $\frac{1}{4}$; przy czym prawdopodobieństwa policzone w modelu jednoparametrowym szybciej zbiegają do ekwilibrum (dzieje się tak ponieważ w modelu Kimury większe jest prawdopodobieństwo **odwrotnej mutacji** czyli np. następującego scenariusza A $\rightarrow$ T $\rightarrow$ A). Ciekawsze jest zachowanie funkcji $Y_{(t)}$: mamy $Y_{(O)} = 0$ oraz $\lim_{t\to\infty} Y_{(t)} = \frac{1}{4}$, ale funkcja $Y_{(t)}$ nie jest monotoniczna — osiąga wartość większą od $\frac{1}{4}$ jeśli:

$$\frac{1}{4}\exp^{-4qt} > \frac{1}{2}\exp^{-2(p+q)t}$$

czyli $\frac{1}{2} > \exp^{-2(p-q)t}$. Warunek ten zachodzi o ile $p > q$ oraz $t > \frac{\ln 2}{(p-q)}$.

1.3 Ogólne modele Markowa

Ogólny model jest zdefiniowany przy pomocy macierzy przejść łańcucha Markowa. Dla sekwencji DNA jest to macierz $\boldsymbol{M}$ 4 x 4, w której wartość $\boldsymbol{M}_{ij} = p_{ij}$ dla $i \neq j$ definiuje prawdopodobieństwo zamiany nukleotydu i na j w jednym kroku.

Rozwiązywane przez nas równania mogą być wyrażone jako:

$$\boldsymbol{P}_{(t+1)} = \boldsymbol{P}_{(t)}\boldsymbol{M}$$

gdzie $\boldsymbol{P}_{(t)} = (P_{A(t)}, P_{G(t)}, P_{C(t)}, P_{T(t)})$

Mówimy, że proces substytucji jest **odwracalny w czasie** (inaczej mówiąc definiujący go łańcuch Markowa jest odwracalny) jeżeli prawdopodobieństwo zamiany nukleotydu i na j w pewnym przedziale czasowym jest takie samo jak prawdopodobieństwo mutacji odwrotnej w tym samym przedziale czasowym.

2 Odległość między sekwencjami

Od momentu rozdzielenia się ewolucji dwóch sekwencji każda z nich zaczyna oddzielnie akumulować mutacje. Ażeby dokładnie oszacować stopień podobieństwa/rozbieżności tych sekwencji (czyli ich ewolucyjną odległość) musimy uwzględniać takie zjawiska jak:

- wielokrotne substytucje nukleotydu znajdującego się w danym miejscu w sekwencji;
- odwrotne mutacje;

- substytucje równoległe: w danym miejscu w obydwu sekwencjach następuje mutacja w ten sam nukleotyd, takie zdarzenie zwiększa podobieństwo sekwencji.

2.1 Podobieństwo sekwencji

Często rozważaną miarą podobieństwa dwóch sekwencji jest procent identycznych nukleotydów. Wartość oczekiwana tej wielkości jest równa prawdopodobieństwu, że w danej chwili t w ustalonym miejscu obserwujemy ten sam nukleotyd (oznaczmy te prawdopodobieństwo przez $S_{(t)}$). W modelu Jukes Cantora mamy:

$$S_{(t)} = P^2_{AA(t)} + P^2_{AT(t)} + P^2_{AC(t)} + P^2_{AG(t)}$$

Ostatnie trzy składniki odpowiadają substytucjom równoległym $A \to T$, $A \to C$ oraz $A \to G$ w obydwu sekwencjach. Wykorzystując wzory (1) i (2) otrzymujemy:

$$S_{(t)} = \frac{1}{4} + \frac{3}{4}\exp^{-8pt}$$

Analogicznie w modelu Kimury:

$$S_{(t)} = \frac{1}{4} + \frac{1}{4}\exp^{-8qt} + \frac{1}{2}\exp^{-4(p+q)t}$$

wyraża oczekiwany procent identycznych nukleotydów dla dwóch sekwencji, które rozdzieliły się t jednostek czasu temu.

Zauważmy, że w modelach odwracalnych (np. w modelu Kimury) zachodzi:

$$S_{(t)} = X_{(2t)}$$

czyli oczekiwany procent identycznych nukleotydów dla dwóch sekwencji po czasie t jest równy procentowi nukleotydów, które pozostaną niezmienione w każdej z sekwencji po upływie $2t$ jednostek czasu. Nie jest to prawdą dla modeli, które nie są odwracalne.

2.2 Zróżnicowanie sekwencji

Jako miarę zróżnicowania sekwencji można przyjąć procent nukleotydów je różniących. Tę wielkość można uzyskać z poprzednich rozważań jako $1 - S_{(t)}$.

Inna ciekawszą miarą jest procent tranzycyjnych i transwersyjnych różnic pomiędzy sekwencjami. Na początek policzmy oczekiwany procent roznic tranzycyjnych w modelu jednoparametrowym dla dwóch sekwencji, których ewolucja rozeszła się t jednostek czasu temu (oznaczmy tę wielkość przez $\boldsymbol{P}_{(t)}$). Załóżmy, że w rozważanym miejscu początkowo znajdował się nukleotyd A.

$$\boldsymbol{P}_{(t)} = 2P_{AA(t)}P_{AG(t)} + 2P_{AT(t)}P_{AC(t)} = \frac{1}{4} - \left(\frac{1}{4}\right)\exp^{-8pt}$$

Powyższy wzór możemy też wyprowadzić używając własności odwracalności modelu: mamy $\boldsymbol{P}_{(t)} = P_{ij(2t)}$. Podobnie w modelu dwuparametrowym (również odwracalnym): $\boldsymbol{P}_{(t)} = Y_{(2t)}$. Czyli:

$$\boldsymbol{P}_{(t)} = \frac{1}{4} + \left(\frac{1}{4}\right)\exp^{-8qt} - \left(\frac{1}{2}\right)\exp^{-4(p+q)t} \tag{3}$$

Dla różnic transwersyjnych zachodzi: $\boldsymbol{Q}_{(t)} = 2Z_{(2t)}$; musimy pomnożyć przez 2, gdyż mamy dwa typy transwersji.

$$\boldsymbol{Q}_{(t)} = \frac{1}{2} - \left(\frac{1}{2}\right)\exp^{-8qt} \tag{4}$$

2.3 Odległość ewolucyjna

W omawianych przez nas modelach przyjmowaliśmy (i nadal będziemy tak czynić :) upraszczające założenie, że częstotliwość mutacji jest stała dla wszystkich miejsc w sekwencji, czyli sekwencja jest homogeniczna. Jest to zblizone do prawdy jedynie dla sekwencji, które nie kodują białek (sekwencji niefunkcjonalnych).

Naszym celem w tym rozdziale jest policzenie wielkości K będącej średnią liczbą substytucji nukleotydów przypadającą na jedno miejsce w sekwencji. Ta wielkość jest uznawana za dobre przybliżenie ewolucyjnej odległości pomiędzy sekwencjami i używana w rekonstrukcji drzew filogenetycznych. Niech N oznacza całkowitą liczbę substytucji pomiędzy rozważanymi sekwencjami, jeśli długość jednej z sekwencji (dlugości sekwencji nie muszą być równe) wynosi L mamy: $K = N/L$. Ażeby policzyć N musimy porównać dwie sekwencje. W tym celu rozważa się ich **uliniowienie** czyli wyrównanie długości sekwencji przez dopisanie spacji (reprezentują insercje i delecje).

Mając gotowe uliniowienie możemy policzyć liczbę różnych nukleotydów na odpowiadających sobie miejscach. Teraz wystarczy jeszcze dodać poprawkę na wielokrotne mutacje odbywające się w tym samym miejscu. Bez tej poprawki całkowita liczba mutacji jest niedoszacowana.

Najprostsza metoda wykorzystuje model jednoparametrowy. W tym modelu oczekiwany procent identycznych nukleotydów dla dwóch sekwencji, które rozeszły się przed czasem t, wynosi:

$$S_{(t)} = \frac{1}{4} + \frac{3}{4}\exp^{-8pt}$$

prawdopodobieństwo, że sekwencje różnią się danym miejscem w chwili t jest równe $r = 1 - S_{(t)}$, czyli:

$$\begin{aligned} r =& \frac{3}{4}(1 - \exp^{-8pt}) \\ 8pt =& -\ln\left(1 - \frac{4r}{3}\right) \end{aligned}$$

Zazwyczaj nie znamy czasu t jaki upłynął od rozejścia się sekwencji. Chcemy natomiast oszacować K, które w tym modelu jest równe $2(3pt)$. Czynnik 2 jak poprzednio wnioskujemy z odwracalności, natomiast $3pt$ jest oczekiwaną liczbą substytucji przypadających na jedno miejsce w sekwencji. Reasumując obliczamy K ze wzoru:

$$K = -\left(\frac{3}{4}\right)\ln\left(1 - \frac{4r}{3}\right)$$

gdzie r jest równe obserwowanej proporcji róznych nukleotydów.

Literatura

[1] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis. Probabilistic models of proteins and nucleid acids.* Cambridge University Press, 1998.

[2] W-H. Li. *Molecular Evolution.* Sinauer Associates, 1997.

3 Zadania ,,*domowe*"

1. [1] Rozważamy nieukorzenione drzewa o m wierzchołkach stopnia 4 i n wierzchołkach stopnia 3. Ile liści i ile krawędzi mają takie drzewa ? Niech $N_{m,n}$ oznacza liczbę różnych drzew o etykietowanych liściach; pokaż że:

$$N_{m,n} = (3m + 2n - 1)N_{m,n-1} + (n + 1)N_{m-1,n+1}$$

2. Udowodnij, że

$$N_{m,0} = \prod_{i=1}^{m}(1 + \frac{9i(i-1)}{2})$$