

Tropem ostatniego wspólnego przodka wszystkich żyjących obecnie ludzi

Adrian Musiał

17 kwietnia 2020

1 Wstęp

Wyobraźmy sobie, że cofamy się w czasie pokolenie po pokoleniu, śledząc przy tym jedynie żeńską linię genealogiczną – nasze matki, matki naszych matek itd. Tworzymy w ten sposób żeńskie drzewo genealogiczne ludzkości, którego liśćmi są wszyscy żyjący obecnie ludzie. Badania archeologiczne wskazują, że kobieta znajdująca się u korzenia tego drzewa żyła około 100 lub 200 tys. lat temu, prawdopodobnie w Afryce [2, 3]. Kobietę tę okrzyknięto mianem „mitochondrialnej Ewy”, jako że od niej pochodzi DNA mitochondrialne (tj. DNA znajdujące się w mitochondriach, które dziedziczone jest jedynie po matce) wszystkich żyjących obecnie ludzi. Fascynujące jest nie tylko to, że taki wspólny przodek istnieje, ale również, że żył stosunkowo niedawno. Dla porównania, najstarsze wydobyte skamieliny gatunku *homo sapiens* pochodzą sprzed 300 tys. lat [4].

Nasuwa się jednak pytanie: co by było, gdybyśmy nie ograniczali się jedynie do żeńskiej linii genealogicznej? Ile pokoleń wstecz należałoby prześledzić, aby natrafić na pierwszego człowieka, który jest spokrewniony z każdą żyjącą obecnie osobą? Możliwej odpowiedzi udziela model zaproponowany przez Josepha T. Changa w 1998 roku [1].

2 Założenia i opis modelu

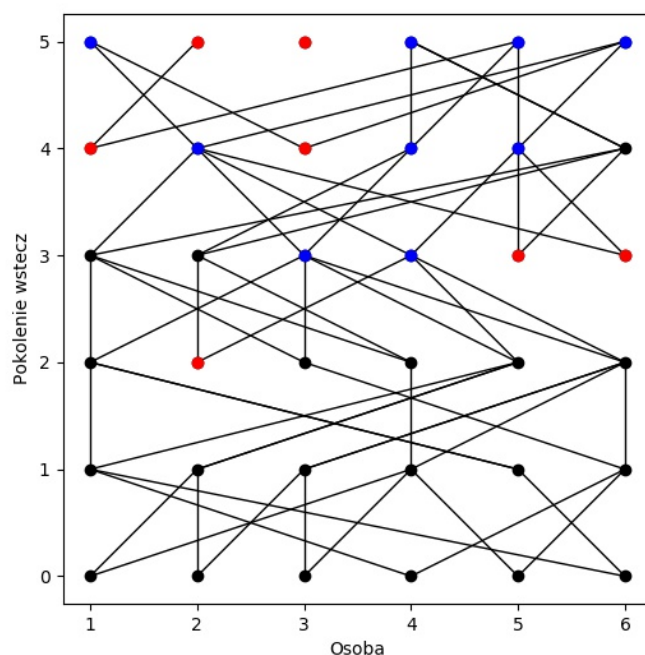
- Liczebność populacji jest równa n i nie zmienia się w czasie.
- Kolejne pokolenia następują skokowo, tzn. wszystkie osoby z danego pokolenia rodzą się w tym samym momencie.
- Pokolenia nie nakładają się, tzn. tylko osoby z tego samego pokolenia mogą mieć potomstwo.
- W momencie narodzin każda osoba wybiera losowo dwóch rodziców z poprzedniego pokolenia. Wszystkie wybory są niezależne i mają

rozkład jednostajny. Może się zdarzyć, że ten sam rodzic zostanie wybrany dwukrotnie – dzieje się tak z prawdopodobieństwem $\frac{1}{n}$.

Model ten stanowi jedynie punkt wyjścia dla dalszych rozważań, więc nie jest oczywiście nazbyt realistyczny. Aby oddawał on rzeczywistość nieco lepiej, możemy w miejsce populacji z n osobnikami rozpatrywać równoważnie populację z n monogamicznymi parami.

3 Przykład

Poniższy przykład ilustruje działanie modelu dla populacji o liczebności $n = 6$. Ograniczamy się do 5 pokoleń wstecz.



Rysunek 1: Zaczynamy od pokolenia 0 (współczesnego). Dla każdego osobnika losujemy dwóch rodziców z pierwszego pokolenia wstecz. Analogicznie postępujemy dla dalszych pokoleń. W drugim pokoleniu wstecz po raz pierwszy pojawia się osobnik, który nie jest spokrewniony z żadną obecnie żyjącą osobą (kolor czerwony), a w trzecim pokoleniu wstecz – osobnik, który jest przodkiem każdej z współcześnie żyjących osób (kolor niebieski). W piątym pokoleniu wstecz każdy osobnik jest spokrewniony albo z wszystkimi osobami z pokolenia 0, albo z żadną z nich.

4 Następstwa modelu

Joseph T. Chang udowodnił w [1] następujące dwa twierdzenia, których dowody, ze względu na ich obszerność i stopień zaawansowania, pomijamy.

Twierdzenie 1. *Niech T_n oznacza liczbę pokoleń licząc wstecz do ostatniego wspólnego przodka wszystkich żyjących obecnie osobników w populacji o liczebności n . Wtedy*

$$\frac{T_n}{\log_2(n)} \xrightarrow{P} 1 \quad (n \rightarrow \infty).$$

Innymi słowy, jeśli populacja liczy dostatecznie dużo osobników, to jest bardzo mało prawdopodobne, że realizacja zmiennej losowej T_n różni się znacznie od $\log_2(n)$. W powyższym przykładzie mamy $T_6 = 3$, podczas gdy $\log_2(6) \approx 2,585$.

Zauważmy, że gdy natrafiłszy już na ostatniego wspólnego przodka (w skrócie ostatniego WP), to wszyscy jego przodkowie również będą WP. Wobec tego im dalej cofamy się w czasie, tym bardziej pospolita jest to własność. Od pewnego pokolenia wstecz każdy osobnik albo jest WP, albo nie ma żyjących potomków, tak jak miało to miejsce w przytoczonym wyżej przykładzie. O tempie osiągania takiego stanu mówi kolejne twierdzenie.

Twierdzenie 2. *Niech U_n oznacza liczbę pokoleń licząc wstecz do pokolenia, w którym każdy osobnik jest albo wspólnym przodkiem wszystkich obecnie żyjących osób, albo nie ma żadnych żyjących potomków. Niech γ oznacza mniejszą spośród dwóch liczb spełniających równanie $\gamma e^{-\gamma} = 2e^{-2}$ oraz niech $\zeta = -\frac{1}{\log_2(\gamma)} \approx 0,7698$. Wtedy*

$$\frac{U_n}{(1 + \zeta) \log_2(n)} \xrightarrow{P} 1 \quad (n \rightarrow \infty).$$

Twierdzenie to możemy rozumieć analogicznie do poprzedniego: jeśli populacja jest odpowiednio liczna, to niewielkie jest prawdopodobieństwo, że realizacja zmiennej losowej U_n odbiega znacząco od $(1 + \zeta) \log_2(n)$. We wcześniejszym przykładzie mamy $U_6 = 5$, natomiast $(1 + \zeta) \log_2(6) \approx 4,573$.

Jasne jest, że zastosowanie modelu do opisu populacji ludzkiej byłoby poważnym nadużyciem. Czynniki takie jak położenie geograficzne czy mobilność społeczna grają zbyt dużą rolę, aby można było je zupełnie zignorować. Nie oznacza to jednak, że nie istnieje populacja (niekoniecznie ludzka), dla której założenia modelu nie są aż tak daleko idącym uproszczeniem. Na myśl przychodzą gatunki zwierząt zamieszkujące niewielkie obszary, bez mocno ustalonej hierarchii społecznej.

Mimo to Czytelnika z pewnością ciekawi, jakie wyniki generuje model w odniesieniu do konkretnych populacji. Jeżeli przyjmiemy za n liczbę żyjących obecnie ludzi (w przybliżeniu 7,8 miliardów), otrzymamy, że ostatni

WP żył około 513 lat temu, co odpowiada $T_n \approx 32,96$ pokoleniom wstecz. Jeśli położymy z kolei $n = 38\,386\,000$ (przybliżona populacja Polski), to U_n będzie równe około 44,57 pokoleń lub 696 lat wstecz. Oznaczałoby to, że każdy żyjący współcześnie Polak jest potomkiem Władysława Łokietka – o ile jego linia genealogiczna w międzyczasie nie wymarła.

5 Symulacja

Przeprowadziłem dodatkowo krótką symulację w Pythonie. Dla wybranych liczebności populacji n wyznaczono 50 realizacji zmiennych losowych T_n oraz U_n . Uśrednione wyniki wraz z ilorazami pojawiającymi się w tezach Twierdzeń 1 i 2 przedstawiono w poniższej tabeli. Ze względu na ograniczone możliwości obliczeniowe i pamięciowe mojego komputera poprzestałem na $n \leq 4000$.

| n | T_n | $\frac{T_n}{\log_2(n)}$ | U_n | $\frac{U_n}{(1+\zeta)\log_2(n)}$ |
|------|-------|-------------------------|-------|----------------------------------|
| 50 | 5,88 | 1,0418 | 12,46 | 1,2481 |
| 100 | 7,00 | 1,0536 | 14,32 | 1,2185 |
| 250 | 8,34 | 1,0470 | 16,94 | 1,2022 |
| 500 | 9,56 | 1,0663 | 18,76 | 1,1829 |
| 1000 | 10,70 | 1,0737 | 21,20 | 1,2026 |
| 2000 | 11,84 | 1,0797 | 22,66 | 1,1682 |
| 4000 | 12,84 | 1,0731 | 24,68 | 1,1660 |

Jak widać, nawet dla niewielkich liczebności populacji n zmienna losowa T_n jest dość dobrze aproksymowana przez $\log_2(n)$. Inaczej sprawy się mają dla U_n – tutaj tempo zbieżności jest znacznie mniejsze.

Literatura

- [1] J. T. Chang *Recent common ancestors of all present-day individuals*, Yale University, 1998.
- [2] R. L. Cann, M. Stoneking, A. C. Wilson *Mitochondrial DNA and human evolution*, Nature 325, 31–36, 1987.
- [3] L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, A. C. Wilson *African populations and the evolution of human mitochondrial DNA*, Science 253, 1503–1507, 1991.
- [4] https://en.wikipedia.org/wiki/Homo_sapiens