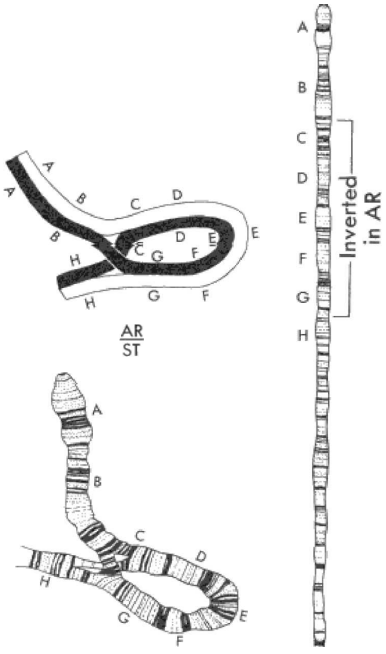
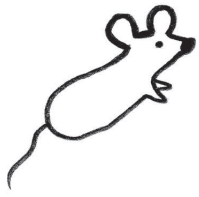



Rys 1. Tak zdefiniowany rewersal modeluje zjawisko molekularne zaobserwowane już w 1938 roku u muszki owocówki. Na rysunku widzimy prawdziwy rewersal, który polega na tym, że fragment chromosomu zawija się w pętelkę, a następnie pętelka ta zostaje powtórnie rozciągnięta w taki sposób, że kolejność genów położonych wewnątrz ulega odwróceniu.



**Naleśniki Billa Gatesa.** Problem sortowania przez rewersale z dodatkowym ograniczeniem, że używamy tylko rewersali typu $\rho(1,i)$, czyli tzw. rewersali prefiksowych, był badany przez studenta Harwardu Billa Gatesa i jego promotora Cristosa Papadimitriou. Podali oni ograniczenia na liczbę $d_{pref}(n) = \max_{\pi \in S_n} d_{pref}(\pi)$. $S_n$ oznacza tutaj zbiór wszystkich permutacji $n$-elementowych. Motywacji do sortowania przez rewersale prefiksowe nie dostarczyła biologia molekularna, lecz następująca zagadka kulinarna: wyobraźmy sobie niestarannego kucharza, który smaży naleśniki różnej wielkości oraz kelnera pedanta, który pragnie podać gościom na stół tacę z naleśnikami ułożonymi w zgrabną piramidę – od największego na dole do najmniejszego na górze. $d_{pref}(\pi)$ jest minimalną liczbą operacji odwrócenia czubka naleśnikowej góry, jaką musi wykonać kelner, żeby posortować naleśniki. Znane oszacowania są następujące: $\frac{17}{16}n \leqslant d_{pref}(n) \leqslant \frac{5}{3}n + \frac{5}{3}$. Problem znalezienia dokładnej wartości $d_{pref}(n)$ nadal czeka na rozwiązanie.

# Bajka o Kocie w Butach, czyli jak zamienić człowieka w mysz

*Anna GAMBIN*[*]

Większość z nas zaakceptowała fakty, do których już od połowy dziewiętnastego stulecia przekonywał nas Karol Darwin i jego następcy, mówiące, że człowieka łączy wyjątkowo bliskie pokrewieństwo z małpami. Dużo trudniej zgodzić się z tym, że gatunek ludzki bardzo wiele wspólnego ma z pospolitymi myszkami – tutaj różnice na pierwszy rzut oka są jeszcze bardziej ewidentne. Okazuje się jednak, że genetycznie jesteśmy bardzo podobni. Jeżeli wyobrazimy sobie genom myszy (20 par chromosomów) jako długi naszyjnik nawleczonych na nitkę genów, to wystarczy pociąć go na około 200 fragmentów i odpowiednio powiązać, aby otrzymać 23 pary chromosomów człowieka. Oznacza to, że stosunkowo niewiele *rearanżacji genomowych* miało miejsce na przestrzeni, bagatela, 80 milionów lat, kiedy to nasza ewolucja podążała odmiennymi torami. Dla ścisłości wyjaśniam, że zaniedbujemy w tym podejściu dużo częstsze zmiany genomu, takie jak mutacje punktowe i uznajemy dwa fragmenty DNA za równoważne, jeśli kodują geny pełniące w rozważanych organizmach taką samą funkcję.

Na pewno pamiętacie bajkę o Kocie w Butach i jego wielkim sukcesie, kiedy namówił złego czarownika do przybrania postaci polnej myszki. Po tym ruchu wystarczyło szybko myszkę połknąć, a rozległe dobra czarownika, w tym przepiękny pałac, stały się własnością Jasia i jego szlachetnie urodzonej małżonki.

Zostaniemy teraz czarownikiem i zamienimy człowieka w mysz za pomocą genetycznych manipulacji, starając się, aby uczynić ten proces jak najbardziej efektywnym. Jak się domyślacie, będzie nam potrzebny matematyczny model zjawiska rearanżacji genomu. Na początek załóżmy, że genom będzie reprezentowany przez permutację $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ zbioru $\{1, 2, \ldots, n\}$. O elementach permutacji możemy myśleć jako o genach albo, co jest bliższe prawdy, większych spójnych fragmentach chromosomu. Do tasowania porządku genów użyjemy operacji o nazwie *rewersal* (oznaczmy ją literką $\rho$). Operacja ta ma dwa parametry, $i$ oraz $j$, a zastosowana do permutacji $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$, odwraca porządek genów od $i$-tego do $j$-tego, czyli:

$$\rho\pi(i,j) = (\pi_1, \ldots, \pi_{i-1}, \pi_j, \pi_{j-1}, \ldots, \pi_{i+1}, \pi_i, \pi_{j+1}, \ldots, \pi_n).$$

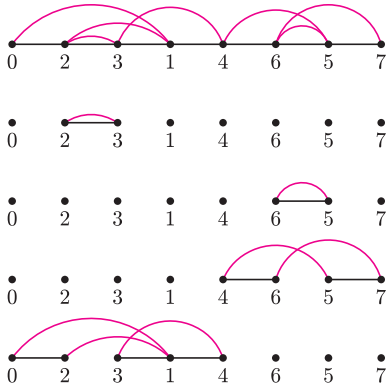
Nasze efektywne czary polegają na przekształceniu permutacji $\pi$ (człowiek) w permutację $\delta$ (myszka) przy użyciu jak najmniejszej liczby rewersali. Tak więc poszukujemy ciągu takich operacji $\rho_1, \rho_2, \ldots, \rho_t$, że $\rho_t \cdots \rho_2 \rho_1 \pi = \delta$ oraz $t$ jest jak najmniejsze. Liczbę $t$ nazwiemy odległością rewersalową pomiędzy permutacjami $\pi$ i $\delta$. Nie jest trudno zauważyć, że możemy bez straty ogólności rozważać problem *sortowania przez rewersale*, czyli znalezienia odległości $d(\pi)$ permutacji $\pi$ od permutacji identycznościowej $(1, 2, \ldots, n)$.

Okazuje się, że posortowanie permutacji przy użyciu rewersali jest zadaniem dość skomplikowanym. Jeśli zajmiemy się jedynie chromosomem X myszy i człowieka (ten chromosom zawiera u większości ssaków bardzo podobne geny i może być reprezentowany jako permutacja długości 7), to stosunkowo nietrudno zweryfikować fakt, że odległość rewersalowa pomiędzy nimi wynosi 6. Sytuacja komplikuje się bardzo, kiedy rozważamy genomy o większej liczbie bloków zachowujących porządek genów, czyli dłuższe permutacje.

Do oszacowania $d(\pi)$ bardzo pomocne okazuje się pojęcie *punktu złamania* permutacji (ang. *breakpoint*). Przyjmijmy oznaczenie $i \sim j$, jeśli $|i - j| = 1$. Dodatkowo na obydwu końcach permutacji $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ dopiszmy $\pi_0 = 0$ oraz $\pi_{n+1} = n + 1$. Parę kolejnych elementów permutacji $(\pi_i, \pi_{i+1})$ dla $0 \leqslant i \leqslant n$ nazwiemy *sąsiedztwem*, jeśli $\pi_i \sim \pi_{i+1}$ czyli są to po prostu kolejne liczby. Przez punkt złamania permutacji będziemy z kolei rozumieć parę $(\pi_i, \pi_{i+1})$, dla której $\pi_i \not\sim \pi_{i+1}$. Zauważmy, że permutacja identycznościowa nie ma żadnego punktu

[*]Instytut Informatyki, Uniwersytet Warszawski

Oczywiście, nie można wykluczyć, że ktoś rozwiąże wartą milion dolarów zagadkę, pokazując, iż klasa problemów wielomianowych jest tożsama z klasą problemów niedeterministycznych wielomianowych, czyli $P = NP$. Więcej o zagadnieniach złożoności obliczeniowej możecie przeczytać w numerze *Delty* 01/2007.



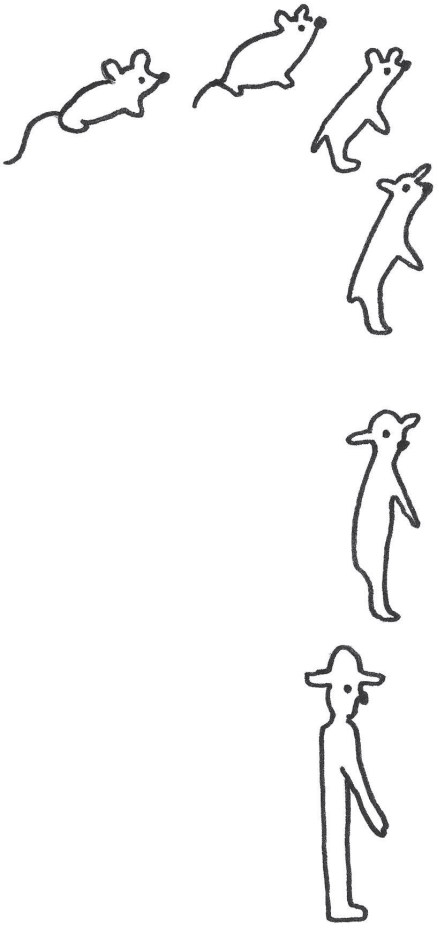

Rys. 2. Graf złamań i jego dekompozycja na cykle dla permutacji $\pi = (2, 3, 1, 4, 6, 5)$.

złamania, czyli proces sortowania polega na eliminacji punktów złamania. Dowolny rewersal potrafi wyeliminować co najwyżej dwa punkty złamania, co pozwala uzyskać oszacowanie: $d(\pi) \geqslant \frac{b(\pi)}{2}$, gdzie $b(\pi)$ oznacza liczbę punktów złamania permutacji $\pi$.

Korzystając z powyższego oszacowania, zaproponowano wielomianowy algorytm aproksymacyjny ze współczynnikiem aproksymacji 2 dla problemu sortowania przez rewersale. Oznacza to, że algorytm wygeneruje ciąg rewersali co najwyżej dwa razy dłuższy niż optymalny. Udowodniono też, że problem sortowania przez rewersale jest $NP$-trudny, czyli najprawdopodobniej nie istnieje algorytm o złożoności wielomianowej sortujący permutację za pomocą minimalnej liczby rewersali.

Ponieważ oszacowanie $d(\pi)$ za pomocą liczby punktów złamania jest najczęściej bardzo niedokładne, zaproponowano kolejne pojęcia mające elegancką, grafową interpretację. Dla permutacji $\pi$ skonstruujemy *graf złamań* $G(\pi)$, który będzie miał $n + 2$ wierzchołki etykietowane liczbami $\{0, 1, 2, \ldots, n, n + 1\}$. Każde dwa wierzchołki $(\pi_i, \pi_{i+1})$ łączymy krawędzią w kolorze czarnym, natomiast parę $(\pi_i, \pi_j)$ łączymy kolorową krawędzią, o ile $\pi_i \sim \pi_j$, czyli są to kolejne liczby w permutacji identycznościowej. Konstrukcję grafu możemy wyobrazić sobie jako nawlekanie na czarną nitkę kolejnych genów jednego organizmu, a następnie nawlekanie na kolorową nić kolejnych genów drugiego organizmu. Cykl w grafie to taka ścieżka, która zaczyna się i kończy w tym samym wierzchołku. Cykl nazwiemy *alternującym*, jeśli kolory kolejnych krawędzi na naszej ścieżce będą zmieniać się naprzemiennie.

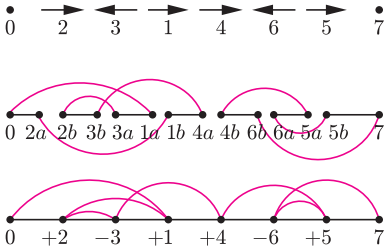
Nasz dwukolorowy graf ma ciekawą własność nazywaną *zbalansowaniem*. Oznacza to, że każdy jego wierzchołek jest zbalansowany, czyli wychodzi z niego taka sama liczba krawędzi kolorowych co czarnych. Są to zawsze dwie czarne krawędzie i dwie kolorowe, z wyjątkiem dwóch skrajnych elementów permutacji, czyli 0 i $n + 1$. Nietrudno pokazać, że grafy zbalansowane mają tzw. *cykl Eulera,* czyli cykl, który zawiera wszystkie krawędzie w grafie i każdą przechodzi dokładnie raz.



Z faktu, że graf złamań dla permutacji ma *alternujący cykl Eulera*, wnioskujemy, iż można przyporządkować wszystkie krawędzie grafu rozłącznym cyklom alternującym w taki sposób, że każda krawędź występuje w dokładnie jednym cyklu. Nazywamy taką operację *dekompozycją na cykle* grafu $G(\pi)$. Graf z rysunku 2 może zostać zdekomponowany na 4 cykle alternujące: $2 \to 3 \to 2$, $6 \to 5 \to 6$, $4 \to 5 \to 7 \to 6 \to 4$ oraz $0 \to 1 \to 3 \to 4 \to 1 \to 2 \to 0$. Okazuje się, że więcej cykli nie uzyskamy w żadnej innej dekompozycji tego grafu, czyli jest to dekompozycja na maksymalną liczbę cykli – oznaczmy tę liczbę przez $c(\pi)$.

Pożyteczna dla nas własność dekompozycji na cykle jest następująca: kiedy zastosujemy dowolny rewersal do permutacji $\pi$, liczba $c(\pi)$ może się zmienić co najwyżej o jeden. Oznacza to po prostu, że wykonując rewersal w celu posortowania permutacji, możemy dodać co najwyżej jeden cykl. Zauważmy, że graf złamań permutacji identycznościowej jest dekomponowalny na maksymalną liczbę cykli alternujących, czyli $n + 1$. Udowodniono, że zachodzi oszacowanie $d(\pi) \geqslant n + 1 - c(\pi)$, które jest o wiele bardziej dokładne, niż rozważane poprzednio $d(\pi) \geqslant b(\pi)/2$ (gdzie, jak pamiętamy, $b(\pi)$ oznacza liczbę punktów złamania permutacji $\pi$).

Niestety, nie zawsze wykonanie rewersalu doda jeden cykl do naszej dekompozycji na maksymalną liczbę cykli alternujących. Dlatego za pomocą liczby cykli potrafimy jedynie oszacować odległość rewersalową. Dobra wiadomość jest taka, że dla permutacji o molekularnym rodowodzie (to znaczy takich, które modelują porządek genów na chromosomie) w powyższym cyklowym oszacowaniu zachodzi najczęściej równość, czyli $d(\pi) = n + 1 - c(\pi)$. Wydawałoby się, że ta obserwacja powinna zagwarantować sukces naszym czarom: wystarczy znaleźć dekompozycję grafu złamań $G(\pi)$ na maksymalną liczbę rozłącznych krawędziowo cykli alternujących $c(\pi)$, policzyć te cykle i wykonać odpowiednie rewersale, a z chromosomu złego czarownika dostaniemy




Rys. 3. Graf złamań i jego *jednoznaczna* dekompozycja na cykle dla permutacji ze znakami $\pi = (+2, -3, +1, +4, -6, +5)$.

chromosom małej myszki. Okazuje się jednak, że nasze czarodziejskie matematyczne zaplecze jest nadal za słabe, bo problem znalezienia wspomnianej dekompozycji jest złożonościowo zbyt trudny.

Na szczęście w sukurs przychodzi nam znów biologia – przypomnijmy sobie, że nić DNA jest obiektem skierowanym (ma swój początek – przez biologów zwany $5'$ i koniec ochrzczony $3'$ – obydwie nazwy pochodzą od położenia atomów węgla w cząsteczce cukru deoksyrybozy). Podobnie geny leżące na nici DNA są skierowane, czyli powinniśmy nasz permutacyjny model organizmu bardziej skomplikować. Komplikacja nie będzie wielka, natomiast zysk z lepszego modelu znaczący. Porządek genów będziemy teraz reprezentować jako *permutację ze znakami*, czyli z każdym elementem związujemy znak '+' lub '−', wskazujący, czy gen jest skierowany w prawo czy też w lewo. Oczywiście, geny leżące na jednej nici DNA są skierowane w tę samą stronę, ale podwójna spirala DNA składa się z dwóch nici DNA biegnących w przeciwnych kierunkach – stąd różne znaki w naszej permutacji odpowiadają temu, że gen pochodzi z jednej bądź z drugiej nici DNA.

Zastanówmy się teraz, jak działa rewersal $\rho(i, j)$ na permutacji ze znakami. Odwraca on na pewno porządek genów, ale też zamienia ich znaki (kierunki). Dla przykładu, stosując rewersal $\rho(2, 5)$ do permutacji $(+1, -5, -4, -3, -2)$, dostaniemy skierowaną permutację identycznościową $(+1, +2, +3, +4, +5)$. Rewersal $\rho(i, i)$ zamienia znak genu $i$. Znowu interesuje nas minimalna liczba rewersali $d(\pi)$, które przekształcą permutację ze znakami $\pi$ w skierowaną permutację identycznościową $(+1, +2, \ldots, +n)$. Dla permutacji ze znakami możemy też zbudować graf złamań, zastępując każdy wierzchołek w grafie $G(\pi)$ przez dwa wierzchołki symbolizujące początek i koniec genu. Oznaczmy te wierzchołki dla genu $i$ jako $ia$ (początek) oraz $ib$ (koniec). Tak więc gen $+2$ będzie reprezentowany przez parę sąsiednich wierzchołków $2a$ i $2b$, natomiast gen $-3$ będzie reprezentowany jako para $3b$ i $3a$. Czarnymi krawędziami łączymy teraz sąsiednie końce genów, czyli dla permutacji z rysunku 3

$$\pi = (+2, -3, +1, +4, -6, +5)$$

łączymy wierzchołki $2b$ i $3b$, następnie $3a$ i $1a$, itd. Kolorowe krawędzie połączą, tak jak poprzednio, geny sąsiednie w permutacji identycznościowej (skierowanej), czyli wierzchołek $1b$ łączymy z $2a$, $2b$ z $3a$, $3b$ z $4a$, itd.

W tej chwili wystarczy, że spojrzycie na obrazek na marginesie, a dokonacie zaskakującego odkrycia: w naszym nowym grafie dekompozycja na alternujące cykle jest jednoznaczna! Ma ona w szczególności maksymalną liczbę cykli. W naszym przykładzie są dwa takie cykle: $0 \to 1a \to 3a \to 2b \to 3b \to 4a \to \to 1b \to 2a \to 0$ oraz $4b \to 5a \to 6a \to 5b \to 7 \to 6b \to 4b$. Wnioskujemy stąd, że aby posortować naszą przykładową permutację, musimy użyć co najmniej $d(\pi) = n + 1 - c(\pi) = 6 + 1 - 2 = 5$ rewersali. Okazuje się, że 5 rewersali wystarczy (spróbujcie!).

Powinniśmy właściwie zakończyć teraz molekularno-matematyczne czary usatysfakcjonowani faktem, że większość złych czarowników potrafimy efektywnie zamieniać w wiele niegroźnych zwierzaków. Warto jeszcze dodać, że uparci matematycy nie dali za wygraną i znaleźli dla permutacji ze znakiem dokładną liczbę rewersali niezbędną do posortowania. Wynosi ona $d(\pi) = n + 1 - c(\pi) + h(\pi) + f(\pi)$, gdzie $c(\pi)$ jest znaną nam liczbą cykli alternujących, $h(\pi)$ jest liczbą *płotków* w permutacji, natomiast $f(\pi)$ wynosi 1, kiedy permutacja $\pi$ jest *fortecą* (jeśli $\pi$ nie jest fortecą, to $f(\pi) = 0$). Nazwy *płotek* i *forteca* sugerują przeszkody, które trzeba pokonać, sortując permutację. Ich zdefiniowanie, niestety, wymagałoby o wiele dłuższych rozważań.

Na zakończenie wspomnijmy, że matematyczna teoria sortowania przez rewersale pozwoliła dokonać ciekawego biologicznie odkrycia: rozważając najbardziej efektywny scenariusz transformacji jednego organizmu w drugi, spotykamy rewersale, które bardzo często są zaczepione w ustalonym punkcie w chromosomie. Własności takich „wrażliwych" miejsc, odkrytych przez matematyków, są teraz badane przez biologów.