O pewnym modelu zawijania białek

Marcin WIERZBIŃSKI[*,†], *Karolina L. TKACZUK*[†], *Alessandro CRIMI*[†]

[*] Doktorant, Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski
[†] Sano Centrum Medycyny Obliczeniowej w Krakowie

Aminokwasy to małe cząsteczki stanowiące główny budulec białka. Można o nich myśleć jak o koralikach nawleczonych na sznurek. Sposób ich ułożenia w przestrzeni decyduje o funkcji białka w komórce oraz o tym, jak oddziałuje ono z innymi elementami komórki. Z tego względu struktura przestrzenna danego białka jest dla badaczy bardzo pożądaną informacją, przydatną m.in. w przypadku projektowania leków.

Istnieje wiele możliwości odtworzenia kształtu białka występującego w naturze. Jedną ze standardowych metod jest „hodowla" cząsteczki w warunkach laboratoryjnych, imitujących te naturalne. Jednak to metoda bardzo czasochłonna i nie zawsze skuteczna, gdyż uchwycenie kształtu białka niekiedy bywa niemożliwe – białko może okazać się niestabilne i rozpaść się przed odtworzeniem jego kształtu. Alternatywą dla tej metody jest tworzenie trójwymiarowych struktur „w krzemie" (tj. przy użyciu komputera) w oparciu o modele teoretyczne. Podstawą tych ostatnich są istniejące już struktury białek (wcześniej odtworzone eksperymentalnie i zebrane w bazie białek *Protein Data Bank*). W ostatnich latach temat ten zyskał duże zainteresowanie mediów, ze względu na sukcesy projektu AlphaFold (`alphafold.ebi.ac.uk`), który do przewidywania struktur białek wykorzystuje sztuczną inteligencję.

Jednym z najprostszych modeli struktury przestrzennej białek jest model hydrofobowo-polarny (HP), w którym aminokwasy podzielone są na dwa typy: H (aminokwas hydrofobowy) i P (aminokwas hydrofilowy). Białko jest w nim reprezentowane przez zagięcie danej skończonej sekwencji $s \in \{H,P\}^k$ w kracie $L = \mathbb{Z}^3$. *Zagięcie* (lub *zwinięcie*) można formalnie zdefiniować jako przekształcenie różnowartościowe $\omega : \{1,\dots,k\} \to L$, takie że sąsiednie liczby odpowiadają sąsiednim punktom z kraty, tzn.

$$\omega(i) \neq \omega(j) \quad \text{oraz} \quad |\omega(i) - \omega(i+1)| = 1 \quad \text{dla } 1 \leqslant i < j \leqslant k,$$

przy czym $|p-q|$ to zwyczajna (euklidesowa) odległość między punktami p i q. Poszukiwane jest zagięcie o największej liczbie E par sąsiadujących w kracie aminokwasów typu H. Wartość $-E$ określa się w tym kontekście mianem *energii*; innymi słowy, poszukiwane jest zagięcie o najmniejszej energii.

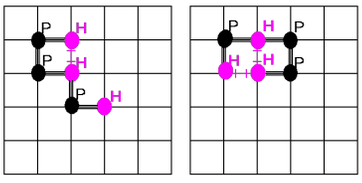
Dla krótkich sekwencji aminokwasów (tzn. niewielkich wartości k), optymalne zagięcie można odnaleźć, obliczając energię każdego możliwego zagięcia. W praktyce jednak sekwencje mogą liczyć od 50 do ponad 1500 aminokwasów, a liczba możliwych zagięć rośnie bardzo szybko wraz ze wzrostem k! W umieszczonej na marginesie tabeli przedstawiono liczbę zagięć w kracie $\mathbb{Z}^3$ dla wybranych wartości k. Jasno wynika z niej, że przeszukiwanie wszystkich konfiguracji nie wchodzi w grę.

To, że przestrzeń możliwości jest ogromna, nie oznacza jeszcze, że z informatycznego punktu widzenia problem jest nierozwiązywalny. Przyjrzyjmy się bliżej trudności omawianego zadania. Sformułowany przez nas problem ma charakter obliczeniowy – szukamy zagięcia ω minimalizującego energię zadaną pewnym wzorem. To zdanie może zostać przeformułowane w następujący problem decyzyjny (nazwijmy go ZB od zwijania białek).

> **Wejście**: Ustalona sekwencja $s \in \{H,P\}^k$, liczba naturalna $m \in \mathbb{N}$.
>
> **Pytanie:** Czy istnieje zawinięcie sekwencji $s \in \{H,P\}^k$ w kracie $\mathbb{Z}^3$ o energii co najwyżej $-m$?

Gdybyśmy znali szybki algorytm rozwiązujący ZB, moglibyśmy uruchamiać go dla coraz większych wartości m, by uzyskać informacje o minimalnej energii zagięcia. Okazuje się jednak, że nasze decyzyjne zadanie jest problemem NP–zupełnym, czyli z punktu widzenia informatyki bardzo trudnym (o takich problemach można przeczytać np. w Δ_{17}^{01} oraz Δ_{17}^{11}).

Rozwiązanie zadania M 1733.
Każda z liczb $2, 3, \dots, p_n$ dzieli albo a, albo b, stąd nie dzieli $a+b$. Wobec tego każdy dzielnik pierwszy $a+b$ musi być równy co najmniej p_{n+1}. Stąd też, o ile liczba $a+b$ sama nie jest pierwsza, to jest iloczynem co najmniej dwóch liczb pierwszych nie mniejszych od p_{n+1} (niekoniecznie różnych), a w konsekwencji $a+b \geqslant p_{n+1}^2$ – co przeczy warunkom zadania.



Przykłady zagięcia sekwencji $s = HPPHPH$ w kracie $\mathbb{Z}^2$. Po lewej stronie energia całkowita jest równa -1, a po prawej stronie -2. Co więcej, zagięcie przedstawione po prawej stronie realizuje globalne minimum energii całkowitej

k	liczba zagięć w $\mathbb{Z}^3$
1	6
2	30
3	150
4	726
5	3534
10	$\sim 8{,}8 \cdot 10^6$
15	$\sim 21{,}2 \cdot 10^9$
20	$\sim 49{,}9 \cdot 10^{12}$

O przybliżaniu liczby możliwych zagięć metodami Monte Carlo można przeczytać w artykule Wojciecha Niemiro *Monte Carlo, spacery i polimery* w Δ_{15}^5 .

Rozwiązanie zadania F 1064.
Ciepło reakcji Q oznacza energię uwalnianą ($Q > 0$) lub pochłanianą ($Q < 0$) podczas reakcji i w naszym przypadku jest równe różnicy sumy energii spoczynkowych substratów i sumy energii spoczynkowych produktów. $Q < 0$ oznacza, że warunkiem zajścia reakcji jest dostarczenie energii $|Q|$. Energia ta zostanie „pobrana" z energii kinetycznej substratów (jąder ^{2}H i ^{14}N). Tylko w układzie środka masy możliwa jest sytuacja, że tuż po reakcji jej produkty spoczywają (względem tego układu) – odpowiadająca tej sytuacji suma energii kinetycznych substratów obliczona w układzie środka masy równa się $|Q|$. Podana w zadaniu wartość $|Q|$ jest ponad 100 razy mniejsza od energii spoczynkowej najlżejszego z występujących jąder atomowych, można więc wykonać obliczenia, stosując wzory mechaniki klasycznej. Niech M oznacza masę ^{14}N (tarczy), m – masę deuteronu ^{2}H (pocisku), a v jego prędkość. Prędkość układu środka masy, V_{CM}, wynosi:

$$V_{CM} = \frac{mv}{m+M}.$$

W układzie środka masy warunkiem zajścia reakcji jest, by suma energii kinetycznych substratów była nie mniejsza niż $|Q|$:

$$\frac{m(v-V_{CM})^2}{2} + \frac{MV_{CM}^2}{2} \geqslant |Q|.$$

Po kilku prostych przekształceniach otrzymujemy warunek na energię kinetyczną pocisku w układzie laboratorium:

$$E_k = \frac{mv^2}{2} \geqslant \frac{m+M}{M}|Q|.$$

Liczbowo: $E_k \geqslant 11{,}54$ MeV.
Dla dociekliwych komentarz do rozwiązania na str. 12.

NP–zupełność problemu zagięć białek została wykazana w artykule *Protein Folding in the Hydrophobic–Hydrophilic (HP) Model is NP–Complete* autorstwa Bonniego Bergera i Toma Leightona. Schemat tego (miejscami mocno skomplikowanego) dowodu jest następujący: zamiast problemu ZB autorzy rozważają problem WS (*Wypełnianie Sześcianu*):

> **Wejście**: Liczba naturalna n i ustalona sekwencja $s \in \{H,P\}^k$, w której znajduje się n^3 liter H.
> **Pytanie:** Czy istnieje zawinięcie sekwencji $s \in \{H,P\}^k$ w kracie $\mathbb{Z}^3$, w którym litery H wypełniają sześcian $n \times n \times n$?

Jest dość intuicyjne, że umiejętność rozwiązania problemu ZB pociąga za sobą zdolność do rozwikłania WS; wystarczy wziąć $m = 3n^2(n-1)$, czyli tyle, ile jest krawędzi w sześciennej siatce $n \times n \times n$. Nietrudno bowiem uwierzyć, że sześcienna siatka jest optymalnym rozwiązaniem z punktu widzenia liczby sąsiadujących liter H. Następnie dowodzi się, że problem WS jest tak samo trudny jak „problem pakowania" (*bin packing*), czyli uogólnienie problemu plecakowego, o którym wiadomo, że jest NP–trudny:

> **Wejście**: Liczby naturalne B i K, zbiór U oraz funkcja $s : U \to \mathbb{N}$ przyjmująca wartości parzyste, taka że $\sum_{u\in U} s(u) = BK$.
> **Pytanie:** Czy można podzielić zbiór U na K podzbiorów U_i $(i = 1,\ldots,K)$ w taki sposób, że dla każdego i mamy $\sum_{u\in U_i} s(u) = B$?

Wiemy już, że zadanie znajdowania optymalnego zwinięcia jest problemem trudnym. Czy to oznacza, że jesteśmy skazani na przypadkowe wybieranie losowych zwinięć i wskazywanie spośród nich tego o najmniejszej energii, choć wiemy, że w ten sposób przeglądamy tylko malutką część całej przestrzeni? Na szczęście nie; można czynić coś istotnie mądrzejszego. Jednym z podejść jest oznaczanie tras, którymi się już przechodziło i które nie dały interesującego zagięcia. Taki pomysł stosowany jest w algorytmie *Monte Carlo Tree Search*. Algorytm został wykorzystany m.in. w słynnym programie Alpha Go, w którym w listopadzie 2015 roku jako pierwszy automat pokonał zawodowego gracza w grę Go, Fan Hui, a w marcu 2016 pokonał jednego z najlepszych zawodowych graczy, Lee Sedola.

W naszym przypadku „gra" (jednoosobowa) może być określona jako wydłużanie zawinięć tak, by uzyskać jak najmniejszą energię. Monte Carlo Tree Search dla modelu hydrofobowo–polarnego na każdym etapie działania algorytmu operuje na drzewie dotychczas sprawdzonych ruchów, czyli zawinięć częściowych. Każde z tych zawinięć ma obliczone dotychczasowe empiryczne oszacowanie jakości. Krok algorytmu składa się z wymienionych niżej działań:

- Startując od korzenia (który tworzy pierwszy aminokwas), schodzimy w dół aż do pewnego liścia $\mathcal{L}$ aktualnego drzewa, w każdym kroku wybierając takie dziecko wierzchołka, które maksymalizuje pewną funkcję (zależną od aktualnego oszacowania jakości dziecka i tego, jak często było odwiedzane na dotychczasowych etapach).
- Rozszerzamy zwinięcie $\mathcal{L}$ o kolejny aminokwas z sekwencji w losowym kierunku i dołączamy tak powstałe zwinięcie $\mathcal{C}$ do aktualnego drzewa.
- Rozszerzamy $\mathcal{C}$ o kolejne aminokwasy z sekwencji w losowych kierunkach aż do otrzymania zwinięcia długości k. Powtarzamy tę operację wielokrotnie i na tej podstawie wyznaczamy jakość $\mathcal{C}$ jako średnią energię końcowych zwinięć uzyskanych w symulacji.
- Aktualizujemy jakość wierzchołków drzewa na drodze od $\mathcal{C}$ do korzenia drzewa (na podstawie wcześniejszych symulacji z $\mathcal{C}$).

Mam nadzieję, że tym przykładem przekonaliśmy Czytelników, że algorytmy losowe mają ciekawe zastosowanie praktyczne i przydają się w poszukiwaniu przybliżonych rozwiązań dla określonych modeli. Sam model hydrofobowo–polarny daje bardzo ciekawe teoretyczne wyniki i wiąże się z istotnymi problemami informatyki, statystyki i matematyki.