STRUKTURY PRZESTRZENNE

Piotr Formanowicz

Cząsteczki biologiczne, takie jak DNA, RNA i białka, są obiektami trójwymiarowymi.

Dotąd jednak zajmowaliśmy się badaniem ich liniowej struktury, tj. sekwencji nukleotydów w przypadku DNA i RNA oraz sekwencji aminokwasów w przypadku białek.

Sekwencje te determinują strukturę trójwymiarową cząsteczek, jednak dotąd nie wiadomo w jaki dokładnie sposób struktura trójwymiarowa zależy od sekwencji (poza pewnymi szczególnymi przypadkami).

Problem określenia struktury trzeciorzędowej na podstawie struktury pierwszorzędowej jest jednym z najważniejszych problemów biologii molekularnej i obliczeniowej.

Cząsteczka RNA zbudowana jest z pojedynczego łańcucha nukleotydów A, C, G, U i to istotnie ją różni od cząsteczki DNA, która składa się z dwóch łańcuchów nukleotydów (A, C, G, T).

Jednoniciowa budowa cząsteczki RNA powoduje, że nukleotyd znajdujący się w pewnej części tej nici może hybrydyzować z nukleotydem znajdującym się w innej jej części, co powoduje, że cząsteczka RNA zwija się.

Ponadto, sekwencja nukleotydów jednoznacznie określa, w jaki sposób cząsteczka się zwija i dlatego można próbować określić kształt cząsteczki na podstawie analizy sekwencji.

Ze względu na fakt, że określenie struktury trzeciorzędowej jest bardzo trudne, skupimy się na przewidywaniu struktury drugorzędowej (która może być podstawą do przewidywania struktury trzeciorzędowej), czyli na określeniu, które pary nukleotydów ze sobą hybrydyzują.

Przedstawiona tu metoda opisana jest m. in. w [1].

Cząsteczkę RNA można postrzegać jako ciąg n znaków $R = r_1 r_2 \dots r_n$, przy czym $r_i \in \{A, C, G, U\}$.

Struktura drugorzędowa cząsteczki RNA jest zbiorem S par (r_i, r_j) nukleotydów, takich że $1 \leq i \leq j \leq n$.

Jeżeli $(r_i, r_j) \in S$, to r_i jest komplementarny do r_j oraz $j - i > t$, gdzie t jest pewnym progiem.

Metody przewidywania struktur przestrzennych są oparte na obliczaniu struktur o minimalnej energii swobodnej.

Zakładamy, iż nie uwzględniamy węzłów, tj. struktur takich, że $(r_i, r_j) \in S$ i $(r_k, r_l) \in S$, gdzie $i < k < j < l$. Jeżeli węzły nie występują, struktura S może być opisana za pomocą grafu planarnego.

W rzeczywistości węzły występują w cząsteczkach RNA, ale nieuwzględnianie ich upraszcza problem. Ponadto, struktury drugorzędowe są wykorzystywane do wyznaczania struktur trzeciorzędowych, a węzły można przewidywać na etapie przewidywania tych struktur.

Przewidywanie struktur przestrzennych można oprzeć na poszukiwaniu struktur o minimalnej energii swobodnej.

W związku z tym potrzebna jest pewna metoda przypisywania takiej energii do struktur.

Można przyjąć, iż istnieją emirycznie określone funkcje dokonujące tego rodzaju przypisania.

Najprostszy algorytm mógłby działać w ten sposób, że wyznaczałby wszystkie możliwe struktury drugorzędowe, obliczałby ich energie swobodne, a następnie wybierałby strukturę z najmniejszą energią.

Niestety, struktur drugorzędowych jest wykładniczo wiele w stosunku do liczby nukleotydów. Stąd, potrzebne są bardziej wyrafinowane algorytmy.

NIEZALEŻNE PARY NUKLEOTYDÓW

Podejście to oparte jest na założeniu, że energia danej pary nukleotydów jest niezależna od energii innych par. Jest to pewne przybliżenie.

Całkowita energia E struktury drugorzędowej S określona jest zależnością

$$E(S) = \sum_{(r_i, r_j) \in S} \alpha(r_i, r_j),$$

gdzie $\alpha(r_i, r_j)$ jest energią swobodną pary (r_i, r_j).

Przyjmujemy, że $\alpha(r_i, r_j) < 0$ dla $i \neq j$ oraz $\alpha(r_i, r_j) = 0$ dla $i = j$.

Założenie o niezależności energii pozwala wykorzystywać rozwiązania dla krótszych ciągów do obliczania rozwiązań dla dłuższych ciągów, co jest podstawą konstrukcji efektywnego algorytmu.

Załóżmy, że chcemy określić strukturę drugorzędową $S_{i,j}$ o minimalnej energii swobodnej dla ciągu $R_{i,j} = r_i r_{i+1} \dots r_j$.

Sprawdźmy, co może nastąpić w przypadku nukleotydu r_j.

Może on utworzyć parę z jakimś nukleotydem pomiędzy r_i oraz r_j lub może takiej pary nie utworzyć.

Jeżeli takiej pary nie utworzy, to

$$E(S_{i,j}) = E(S_{i,j-1}).$$

Jeżeli r_j tworzy parę z r_i, to

$$E(S_{i,j}) = \alpha(r_i, r_j) + E(S_{i+1,j-1}).$$

Jeżeli natomiast r_j tworzy parę z r_k, gdzie $i < k < j$, to możemy podzielić ciąg $R_{i,j}$ na dwa ciągi $R_{i,k-1}$ i $R_{k,j}$. W takim przypadku mamy

$$E(S_{i,j}) = E(S_{i,k-1}) + E(S_{k,j})$$

dla pewnego k (można tak zrobić, ponieważ założyliśmy, że nie ma węzłów).

Problem polega na tym, że nie wiadomo, które k daje najmniejszą całkowitą energię. Zatem można napisać

$$E(S_{i,j}) = \min\{E(S_{i,k-1}) + E(S_{k,j})\},$$

gdzie $i < k < j$.

Oznacza to, że można obliczyć energię $E(S_{i,j})$ na podstawie energii krótszych ciągów. Można to zrobić korzystając z programowania dynamicznego w następujący sposób:

$$E(S_{i,j}) = \min \begin{cases} E(S_{i+1,j-1}) + \alpha(r_i, r_j) \\ \min\{E(S_{i,k-1}) + E(S_{k,j})\} \text{ dla } i < k \leq j \end{cases}$$

(Przypadek, w którym r_j nie tworzy pary jest uwzględniony w drugiej z powyższych zależności, ponieważ, gdy $k = j$, zachodzi $E(S_{j,j}) = 0$.)

Algorytm obliczania wartości $E(S_{1,n})$ jest bardzo podobny do algorytmu programowania dynamicznego dla problemu dopasowania dwóch sekwencji.

Złożoność zaprezentowanego algorytmu wynosi $O(n^3)$.

Literatura

[1] J. Setubal, J. Meidanis. *Introduction to Computational Molecular Biology.* PWS Publishing Company, Boston 1997.

v. 1.0.0