Algorytm symulujący proces nieenzymatycznej degradacji RNA w oparciu o przewidywanie struktury drugorzędowej

1. Wstęp

Badania prowadzone nad procesem degradacji RNA w warunkach charakteryzujących się nieobecnością czynników komórkowych (czyli tzw. hydrolizą nieenzymatyczną), pokazały, że jest to proces silnie zależny od struktury przestrzennej RNA (zarówno pierwszo-, drugo- jak i trzeciorzędowej). W pierwszych eksperymentach biochemicznych, przeprowadzonych na krótkich oligorybonukleotydach zauważono, że cięcia występowały jedynie w rejonach jednoniciowych, natomiast rejony dwuniciowe pozostawały stabilne. Dodatkowo zaobserwowano, że poszczególne wiązania internukleotydowe (pomiędzy nukleotydami w łańcuchu RNA np. dla łańcucha AGG są to wiązania pomiędzy A i G oraz G i G) w łańcuchu RNA wykazują różną podatność na cięcie. W związku z tym, pewne wiązania (np. YA lub YC, gdzie Y oznacza U lub C) są bardziej preferowane niż pozostałe. Powoduje to, że każda cząsteczka RNA ma trochę inny wzór degradacji (listę fragmentów na jakie ulega rozpadowi).

Analiza wzoru degradacji cząsteczki RNA dostarcza wielu informacji, m.in.: pozwala wnioskować o jej stabilności, przewidywać czy powstałe z niej degradanty mają szansę przetrwać (nie ulec dalszej degradacji) i pełnić określone funkcje, czy też projektować sztuczne cząsteczki RNA, tak, aby rozpadały się generując zadane, stabilne degradanty.

Laboratoryjne zbadanie wzoru degradacji dowolnej cząsteczki RNA jest żmudne i czasochłonne, dlatego też zaproponowano algorytm podziału i odcięć (ang. *branch and cut*) symulujący przebieg procesu nieenzymatycznej degradacji RNA. Uwzględnia on sekwencję oraz strukturę przestrzenną zadanej cząsteczki RNA, a także reguły stabilności opracowane przez Kierzka i współpracowników, tak aby na bazie tych informacji przewidzieć jej wzór degradacji.

2. Reguły stabilności RNA

Nieenzymatyczna degradacja RNA jest procesem zależnym od struktury cząsteczki RNA, a cięcia zachodzą jedynie w jej rejonach jednoniciowych. Dodatkowo zaobserwowano też różną podatność na hydrolizę poszczególnych wiązań fosfodiestrowych łączących nukleotydy w łańcuchu RNA, i tak, najbardziej niestabilnym okazało się być wiązanie między UA, które było hydrolizowane 1.5-2 razy szybciej niż między CA. Wiązanie YC (UC, CC) było 3-5 razy bardziej stabilne niż YA (UA, CA), a wiązanie pomiędzy YG (UG, CG) i YU (UU, CU) rozpadało się 20-50 razy wolniej niż pomiędzy YC (UC, CC) i YA (UA, CA). Pozostałe wiązania RR (AA, AG, GG, GA) i RY (AU, AC, GU, GC) wykazywały stabilność w badanych warunkach cięcia (w tych miejscach cięcia nie następowały). Analiza omawianych wiązań pozwoliła na uszeregowanie ich począwszy od najmniej do najbardziej stabilnych w następujący sposób: UA>CA>YC>YG>YU, gdzie Y oznacza C lub U, a R oznacza A lub G.

W ramach przeprowadzonych eksperymentów, chemikom nie udało się przypisać powyższym wiązaniom konkretnych wartości, zdołali jednak zauważyć pewne zależności pomiędzy ich stabilnością (np. UA jest 1.5 razy szybciej hydrolizowany niż CA). Matematyczne wyrażenie tych zależności wymagało więc zaproponowania takiego podejścia, które ze względu na wystąpienie istotnych błędów pomiarowych, brałoby pod uwagę "zaszumienie" tych wartości. Zdecydowano się przyjąć podejście stosowane w teorii zbiorów rozmytych (ang. *fuzzy sets*). Przypadek z którym mamy do czynienia w kontekście reguł stabilności Kierzka i współpracowników bliski jest przypadkowi "im większe tym lepsze" (ang. *the larger the better*)

opisanemu przez Taguchi (np. gdy celowe jest zmaksymalizowanie pewnych pożądanych cech produktu) oraz mierze przynależności z nim związanej:

$$\mu_i(y) = \frac{y_i^q}{a + y_i^q} \quad \text{lub} \quad \mu_i(y) = \frac{\Delta y_i^q}{a + \Delta y_i^q} \quad \text{dla} \quad 0 \leq \mu_i(y) \leq 1$$

gdzie a = 1, q=1/2 (w rozważanym przypadku), a Δy_i odzwierciedla przedział zmienności y_i wraz z przypisaną miarą przynależności, w ramach którego zlokalizowane jest rozwiązanie rozważanego problemu.

Niech P={UA, CA, UC, CC, UG, CG, UU, CU} będzie zbiorem wszystkich dopuszczalnych miejsc cięcia w cząsteczce RNA zgodnie z danymi pochodzącymi z eksperymentów wykonanych przez Kierzka i współpracowników.

Na podstawie zależności zaobserwowanych pomiędzy szybkością rozpadu poszczególnych wiązań internukleotydowych w cząsteczce RNA, opracowano następujące przedziały wartości [c, d] dla każdego dopuszczalnego miejsca cięcia w zbiorze P:
Dla wiązań pomiędzy YG oraz YU:

$$c_{(YG,YU)} = 0$$

$$d_{(YG,YU)} = 0.1$$

Dla wiązań pomiędzy YC:

$$c_{(YC)} = 20$$

$$d_{(YC)} = 50$$

Dla wiązania pomiędzy CA:

$$c_{(CA)} = 3 * c_{(YC)}$$

$$d_{(CA)} = 5 * d_{(YC)}$$

Dla wiązania pomiędzy UA:

$$c_{(UA)} = 1.5 * c_{(CA)}$$

$$d_{(UA)} = 2 * d_{(CA)}$$

Dla każdego miejsca cięcia ze zbioru P, uwzględniając błędy pomiarowe, zdefiniowane zostały następujące miary degradacji (przynależności):

$$\mu(UA) = \frac{(d_{(UA)} - c_{(UA)})^{\frac{1}{2}}}{1 + (d_{(UA)} - c_{(UA)})^{\frac{1}{2}}} = 0.953$$

gdzie $\mu_{(UA)}$ oznacza miarę, biorącą pod uwagę niedokładność pomiarów pochodzących z eksperymentu biochemicznego. Miara ta określa jak duża jest szansa na to, że badana cząsteczka RNA pęka w konkretnym miejscu (w tym przypadku pomiędzy nukleotydami UA). Wartości miary w pozostałych miejscach cięcia zostały zdefiniowane analogicznie, jak przedstawiono poniżej:

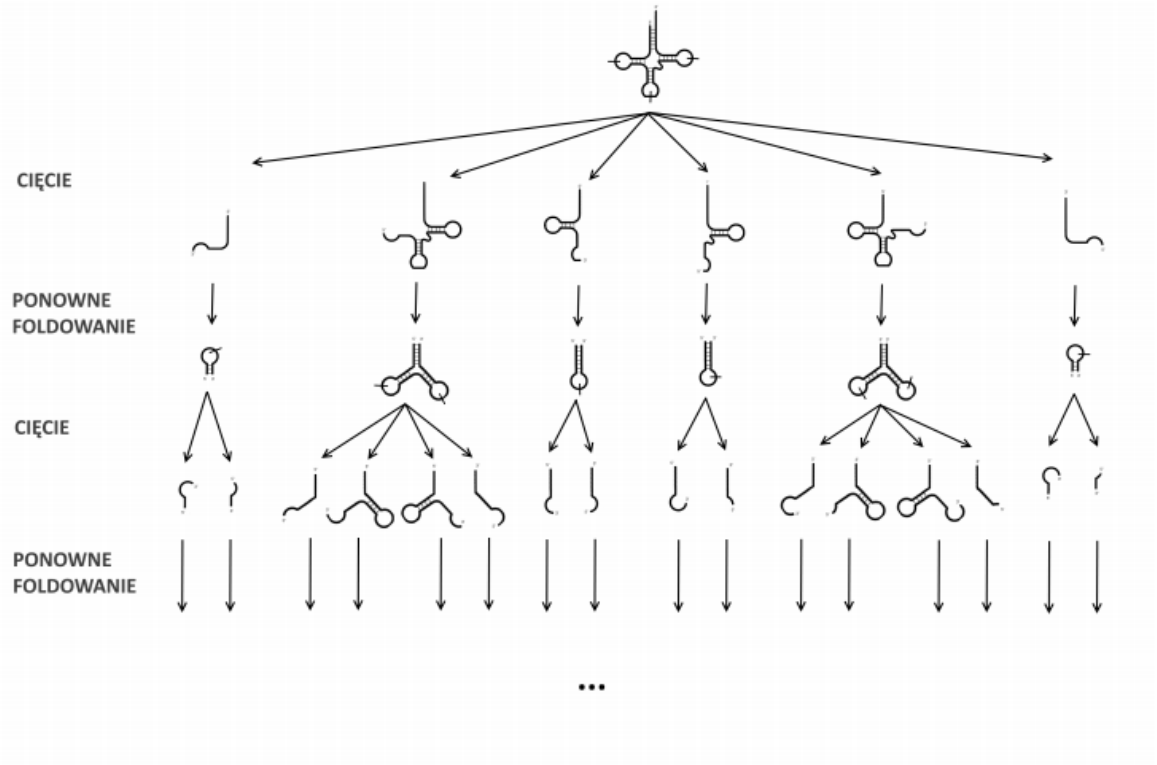
$$\mu(CA) = \frac{(d_{(CA)} - c_{(CA)})^{\frac{1}{2}}}{1 + (d_{(CA)} - c_{(CA)})^{\frac{1}{2}}} = 0.932$$

$$\mu(YC) = \frac{(d_{(YC)} - c_{(YC)})^{\frac{1}{2}}}{1 + (d_{(YC)} - c_{(YC)})^{\frac{1}{2}}} = 0.846$$

$$\mu(YG, YU) = 0.1$$

3. Algorytm podziału i odcięć

W celu przewidzenia wzoru degradacji dowolnej cząsteczki RNA, proponowany jest algorytm podziału i odcięć (ang. *branch-and-cut*), który zaimplementowany będzie w języku Java. Algorytm symuluje proces nieenzymatycznej degradacji RNA podanej na jego wejściu cząsteczki RNA, na bazie jej sekwencji oraz reguł stabilności zaproponowanych przez Kierzka i współpracowników. Zastosowanie tych reguł, wymaga, aby została wykonana predykcja struktury drugorzędowej, do czego wykorzystywane jest zewnętrzne narzędzie. W uzyskanej strukturze 2D wyszukiwane są odcinki jednoniciowe, po czym rozpoznawane są w nich miejsca cięcia obecne w zbiorze P, a następnie w wybranym miejscu wykonywane jest cięcie, którego wynikiem jest zawsze para fragmentów RNA. Ważnym założeniem jest, że zaraz po hydrolizie, oba wynikowe fragmenty oddysocjowują (oddzielają się od siebie i stają się od siebie niezależne) i każdy z nich przybiera nową strukturę drugorzędową. W związku z tym, w celu przewidzenia dalszych produktów spontanicznej degradacji RNA, konieczne jest ponowne wykorzystanie zewnętrznego programu służącego do predykcji struktury 2D RNA. Ogólny schemat działania algorytmu został przedstawiony na poniższym Rysunku:



RYSUNEK Ogólna idea działania algorytmu pokazana na przykładzie cząsteczki tRNA. Załóżmy, że w pierwszym kroku tRNA cięty jest w trzech miejscach, w wyniku czego otrzymujemy sześć fragmentów. Następnie każdy z nich ulega ponownemu foldowaniu (przyjmuje nową strukturę drugorzędową) i ponownie podlega degradacji. Cała procedura powtarzana jest tak długo, dopóki powstawać będą fragmenty, które posiadają dozwolone miejsca cięcia

Proponowana metoda poszukiwania rozwiązania została częściowo oparta na wykorzystaniu zbiorów rozmytych oraz planowaniu eksperymentu wg Taguchi. Bazując na tym, przyjęto, że poprawny wynik eksperymentu może być osiągnięty w kilku krokach (metoda

podziału i odcięć), gdzie do węzłów drzewa przypisane są miary przynależności (miary będące wynikiem intersekcji (mnożenia):

$$\mu(x, \Theta) = \mu_1(x, \Theta) * \mu_2(x, \Theta) * \ldots * \mu_k(x, \Theta)$$

co jest typowym rozumowaniem związanym z poszukiwaniem rozwiązania dokładnego, stosowanym w systemach wnioskowania rozmytego.

Przyjęta miara przynależności $(0 \leq \mu(x, \Theta) \leq 1)$, może być interpretowana jako wzięcie pod uwagę faktu, że prawdopodobieństwo znalezienia rozwiązania rośnie w sytuacji, gdy analizowany przedział jest szerszy. Oznacza to, że dla większego przedziału, wartość miary przynależności będzie wyższa.

W kolejnych krokach algorytmu, w wyniku zastosowania operacji mnożenia (intersekcji), wartość miary przynależności zmniejsza się, co należy rozumieć jako zmniejszenie obszaru, w ramach którego dostępne są wszystkie rozwiązania (przestrzeni wszystkich rozwiązań) i jednocześnie wzrost prawdopodobieństwa znalezienia konkretnego, poszukiwanego rozwiązania. Warto zauważyć, że według tej teorii, projektant algorytmu ma możliwość swobodnego wyboru miary przynależności. Jedynym sposobem zweryfikowania przyjętych założeń jest wykonanie eksperymentu biochemicznego.

Danymi wejściowymi (parametrami) algorytmu są: sekwencja RNA, próg odcięcia dla iloczynu miar degradacji RNA (domyślnie ε = 0.01), minimalna długość fragmentu dla którego może zostać jeszcze przewidziana struktura 2D RNA (domyślnie minLen = 20nt) oraz zewnętrzny program wykorzystywany do predykcji struktury 2D (RNAfold, CentroidFold lub ContextFold).

Każdy węzeł drzewa zawiera wartość miary degradacji ε_{node}, która obliczana jest jako iloczyn miary przynależności węzła rodzicielskiego oraz bieżącej wartości miary, przydzielonej do wybranego miejsca cięcia. Wynikiem działania algorytmu jest zbiór wszystkich fragmentów będących wynikiem degradacji wejściowej cząsteczki RNA, zapisanych w postaci par (x, y) określających ich dokładne położenie w ramach cząsteczki wejściowej, uporządkowanych zgodnie z przydzieloną im punktacją (ranking). Gdy cząsteczka RNA ulega degradacji to fragmenty (degradanty), które powstaną w wyniku cięć w miejscach najbardziej niestabilnych będą powstawały w największej ilości, a przez to zwiększa się szansa na ich uzyskanie i potwierdzenie w ramach eksperymentu biochemicznego. To powinno w punktacji być uwzględnione.

Aby rozwiązać problem, algorytm buduje oraz przeszukuje drzewo rozwiązań. Liście drzewa reprezentują częściowe wyniki degradacji i odpowiadają elementom kompletnego rozwiązania. W każdym węźle wykonywana jest predykcja struktury drugorzędowej z wykorzystaniem wybranego zewnętrznego narzędzia. W kolejnym kroku, w uzyskanej strukturze 2D poszukiwane są odcinki jednoniciowe (zewnętrzny skrypt) i dla każdego z nich poszukiwane są dopuszczalne miejsca cięcia na bazie zbioru P i przypisywane są do nich odpowiadające im miary degradacji. Zakłada się, że aby mogło dojść do cięcia, przynajmniej 1 nukleotyd tworzący wiązanie fosfodiestrowe musi być niesparowany (stanowić część odcinka jednoniciowego). Wszystkie miejsca cięcia, dla których miara degradacji przemnożona przez miarę degradacji przechowywaną w bieżącym węźle jest większa od progu odcięcia dla iloczynu miar degradacji ε (który jest parametrem podanym na wejściu algorytmu przez użytkownika) podlegają dalszej analizie. Fragmenty o długości mniejszej niż minLen dodawane są do zbioru wynikowego (z odpowiednią punktacją).

W sytuacji, gdy dla rozważanego fragmentu RNA nie udało się znaleźć dopuszczalnego miejsca cięcia (zgodnie ze zbiorem P) lub gdy iloczyn miar degradacji miał wartość mniejszą niż

ε, taki fragment uznawany jest za końcowy produkt degradacji i dodawany do zbioru wynikowego (z odpowiednią punktacją).

Algorytm kończy pracę, gdy nie istnieje już żaden fragment, który zawierałyby dopuszczalne miejsca cięcia. Wyniki działania algorytmu (wzór degradacji cząsteczki RNA w postaci listy fragmentów będących wynikiem nieenzymatycznej degradacji wraz z ich punktacją) znajduje się w zbiorze wynikowym. W zbiorze wynikowym fragmenty powinny być uporządkowane zgodnie z punktacją, dla każdego z nich powinna zostać podana jego sekwencja oraz lokalizacja w ramach (względem) cząsteczki wejściowej (czyli dla degradantów z kolejnych etapów trzeba przeliczyć ich lokalizację na lokalizację w ramach cząsteczki wejściowej).

Ostatnim etapem jest zaproponowanie punktu odcięcia – czyli dla posortowanych zgodnie z ich punktacją degradantów, należy wskazać wartość punktacji, poniżej której pojawienie się tych fragmentów jest już statystycznie bardzo mało prawdopodobne. Celem jest też ograniczenie liczby uzyskanych fragmentów, do tych występujących w największych ilościach w wyniku degradacji, ponieważ to one mogą pełnić dodatkowe funkcje w komórce i mają szansę na wykrycie metodami biologii molekularnej.