STRESZCZENIE ROZPRAWY DOKTORSKIEJ

Analiza grup i sygnałów używanych do budowy struktury białek z lokalnych deskryptorów

Michał Drabikowski

1 Wstęp

Białka to cząsteczki o fundamentalnym znaczeniu dla organizmów żywych: są podstawowym składnikiem suchej masy komórek, realizują wszystkie ich ważniejsze funkcje. To, jaką funkcję pełni dane białko, determinowane jest bezpośrednio przez przyjmowaną przez nie strukturę przestrzenną, czyli konformację. Konformacja określona jest zaś jednoznacznie przez sekwencję aminokwasów: aby zminimalizować swoją energię swobodną, białko spontanicznie zwija się w środowisku wodnym w określony sposób — zawsze tak samo! — dzięki czemu może pełnić określoną funkcję.

W dobie sekwencjonowania genomów organizmów na wielką skalę w olbrzymim tempie przyrasta wiedza o sekwencjach białkowych kodowanych przez nukleotydy na nici DNA. Jednak wiedza ta ma niewielkie znaczenie praktyczne — dopiero znajomość struktury przestrzennej, a co za tym idzie w dalszej kolejności funkcji białek, pozwala rozwijać się najróżniejszym działom biotechnologii: od produkcji kosmetyków czy pestycydów, aż po zastosowania w wielu sektorach przemysłu farmakologicznego (na przykład przy projektowaniu leków białkowych i eliminowaniu efektów ubocznych ich działania).

Stosowane obecnie metody eksperymentalne służące poznaniu struktury przestrzennej białek (krystalografia rentgenowska, jądrowy rezonans magnetyczny) są niezwykle kosztowne i czasochłonne. Dlatego właśnie problem przewidywania konformacji białek w oparciu o ich sekwencję (na którą można patrzeć jak na słowo zbudowane nad dwudziestoelementowym alfabetem) jest jednym z najważ-

niejszych, choć ciągle otwartych problemów stojących przed bioinformatyką.

Rozwinięte białko przyjmuje spontanicznie swoją strukturę przestrzenną w czasie od kilku milisekund do kilku minut. Symulacja tego procesu w pełnym atomowym modelu jest praktycznie niewykonalna obliczeniowo. Dlatego, aby przewidywać strukturę przestrzenną białek, konieczne są prostsze modele, uwzględniające pewne specyficzne cechy białek. Obecnie przewiduje się ich strukturę, stosując jedno z trzech podstawowych podejść:

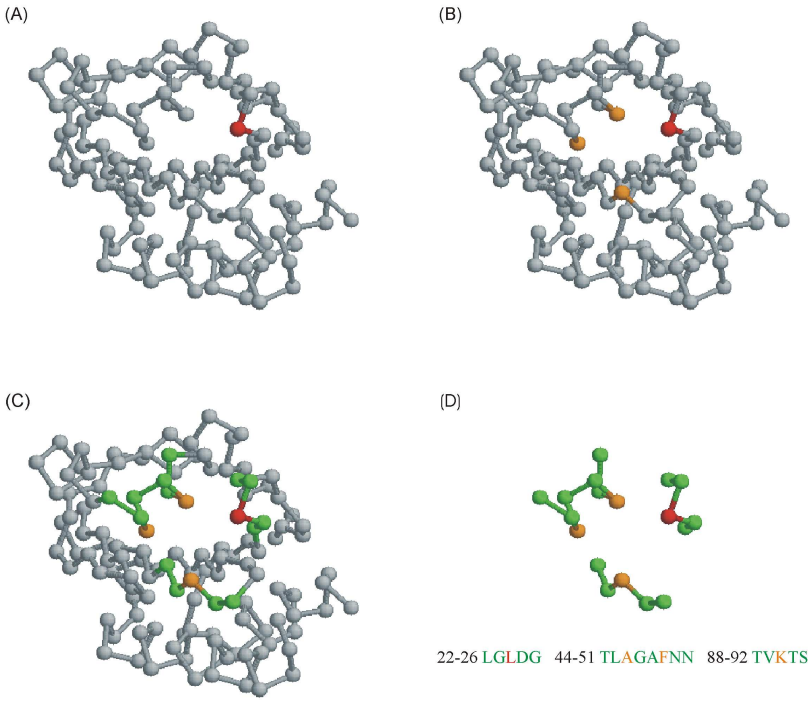
- *modelowanie komparatywne* — mające zastosowanie wtedy, gdy dla rozważanego białka o nieznanej strukturze można znaleźć białko podobne sekwencyjnie (a zatem i strukturalnie) o znanej strukturze: służy ono za wzorzec przy budowie szukanej struktury [1, 2],

- wypierana ostatnio przez modelowanie komparatywne metoda *nawlekania*, polegająca na „nawlekaniu" rozważanej sekwencji białkowej na białka o znanej strukturze i badaniu, czy otrzymane struktury są stabilne w rozumieniu pewnej zdefiniowanej wcześniej funkcji potencjału [3],

- predykcja *ab initio* — stosowana wtedy, gdy zawodzą wyżej wspomniane techniki: standardowe rozwiązanie polega na stworzeniu modelu reprezentującego strukturę przestrzenną białek w pewnej przestrzeni poszukiwań oraz definiującego funkcję potencjału oceniającą energię układu, a następnie na poszukiwaniu struktury przestrzennej analizowanego białka minimalizującej energię w tym modelu [4].

Prawie wszystkie liczące się obecnie metody przewidywania struktury białek nie wykazujących podobieństwa sekwencyjnego do jakiegokolwiek białka o znanej strukturze łączą technikę *ab initio* z próbą składania białek z pewnych — najczęściej kilku- lub kilkunastoaminokwasowych — fragmentów struktur białkowych zbudowanych w oparciu o wszystkie białka o znanej strukturze. Jednym z prekursorów takiego podejcia jest David Baker [5], który od wielu lat odnosi wraz ze swoim zespołem sukcesy w eksperymencie CASP (http://predictioncenter.org), będącym najważniejszym wydarzeniem i głównym źródłem wiedzy o postępie w badaniach nad predykcją struktury białek.

2 Metoda lokalnych deskryptorów



Rysunek 1: Schemat tworzenia deskryptorów. Dla dowolnego białka o znanej strukturze i dowolnego aminokwasu w tym białku (A) znajdowane są aminokwasy bliskie strukturalnie (B), które następnie „rozszerza się" poprzez dodanie z obu stron dwóch najbliższych sekwencyjnie „sąsiadów" (C): w efekcie powstaje deskryptor niosący informację zarówno strukturalną, jak i sekwencyjną (D).

Punktem wyjścia niniejszej pracy jest zaproponowane przez Krzysztofa Fidelisa [6] nowatorskie podejście do definiowania motywów białkowych, pozwalające na identyfikację i charakterystykę regularności w sekwencji i strukturze białek. Główną ideą tego podejścia jest analizowanie nie tylko sąsiadujących w białku aminokwasów (jak robi to między innymi David Baker), ale również kontaktów dalekiego zasięgu między aminokwasami. Owocuje to konstrukcją *lokalnych deskryptorów*, czyli zbiorów ciągłych fragmentów białkowych zwanych *segmentami*. Schemat tworzenia deskryptorów opisujących lokalne otoczenie przestrzenne dowolnego aminokwasu w dowolnym białku przedstawiono na rysunku 1.

Utworzone na podstawie wszystkich białek o znanej strukturze lokalne deskryptory mogą być wykorzystane do przewidywania struktury przestrzennej białek (stosowaną tu metodę można traktować jako swego rodzaju uogólnienie modelowania komparatywnego, polegające na korzystaniu z wielu wzorców). Po połączeniu podobnych strukturalnie deskryptorów w grupy, wewnątrz każdej z nich wykrywana jest zależność sekwencyjno-strukturalna. Zależności tej — czyli *sygnału sekwencyjnego* — używa się do znajdowania przypisań grup do sekwencji danego białka (o nieznanej strukturze), które to przypisania ze znacznym prawdopodobieństwem opisują poprawnie lokalną strukturę białka. Następny krok to przewidywanie jego globalnej struktury poprzez zidentyfikowanie możliwie dużego zbioru zgodnych strukturalnie przypisań.

3 Znajdowanie i uzgadnianie przypisań

Pierwszym krokiem było wprowadzenie matematycznego formalizmu opisującego deskryptory i grupy deskryptorowe. Wprowadzono też kluczowe dla dalszych rozważań pojęcie *przypisania*: przypisanie danej grupy deskryptorowej g do dowolnego białka s to przyporządkowanie segmentom należącym do g rozłącznych spójnych fragmentów białka s tej samej długości co przypisywane segmenty. Każde takie przypisanie wyznacza pewien hipotetyczny lokalny opis sekwencji i struktury białka s indukowany przez grupę g.

Z każdym przypisaniem związać można jego jakość: gdy znana jest struktura białka, wyrażamy tę jakość przy pomocy kryteriów strukturalnych (funkcja $RMSD$ określona na ciągach punktów przestrzeni trójwymiarowej); gdy znana jest natomiast jedynie sekwencja białka, jakość tę wyrażamy przy pomocy kryteriów sekwencyjnych (funkcja sygnału sekwencyjnego pozwala określić w terminach probabilistycznych podobieństwo macierzy aminokwasów do sekwencji aminokwasów). Dzięki temu możemy sformułować w każdym przypadku problem znajdowania najlepszych przypisań analizowanej grupy do dowolnej sekwencji. Niestety oba rozważane problemy okazały się złożone obliczeniowo: udowodniliśmy NP-zupełność decyzyjnej wersji zarówno problemu znajdowania najlepszych przypisań strukturalnych (poprzez redukcję silnie NP-zupełnego problemu pakowania koszyków), jak i problemu najlepszych przypisań sekwencyjnych (poprzez

redukcję problemu spełnialności pewnych formuł logicznych), mimo przyjęcia w tym ostatnim przypadku upraszczającego założenia o *addytywności* funkcji sygnału sekwencyjnego (założenie to mówi, że jakość przypisania grupy g może być obliczona jako suma jakości przypisań wszystkich segmentów należących do g).

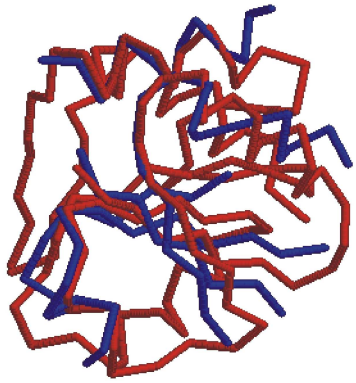
Wykazana NP-zupełność oznacza, że nie można spodziewać się wielomianowego — ze względu na liczbę segmentów w rozważanej grupie — algorytmu rozwiązującego postawione problemy. Okazuje się jednak, że do ich rozwiązania nie jest konieczne analizowanie przestrzeni wszystkich przypisań grupy deskryptorowej do analizowanego białka (przypisań takich jest $\Theta(n^k)$, gdzie n to długość białka: od kilkudziesięciu do kilku tysięcy aminokwasów, a k — liczba segmentów w analizowanej grupie, przy czym $k \leqslant 10$ dla wszystkich analizowanych grup deskryptorowych). Zaproponowaliśmy dokładne algorytmy pozwalające efektywnie rozwiązywać oba problemy w odniesieniu do rozważanych danych: w przypadku przypisań sekwencyjnych zastosowano programowanie dynamiczne (co pozwala znajdować najlepsze addytywne przypisanie sekwencyjne w czasie $O(k2^k)O(n)$), natomiast w przypadku przypisań strukturalnych znacznie ograniczono przestrzeń poszukiwań, wykorzystując pewne własności funkcji $RMSD$.

Przypisania grup deskryptorowych do ustalonego białka opisują lokalnie jego strukturę. Przedstawiliśmy efektywny algorytm pozwalający uzgadniać takie opisy, tak by mógł powstać opis globalny. Sformułowaliśmy wreszcie problem znajdowania optymalnego globalnego opisu (czyli predykcji — gdy nie jest znana struktura białka, albo też rekonstrukcji struktury) na podstawie danego zbioru przypisań grup deskryptorowych. Również w tym przypadku udowodniliśmy jego NP-zupełność (redukcja problemu kliki), proponując do rozwiązania tego problemu algorytm heurystyczny oparty o programowanie zachłanne.

4 Zastosowania

Sformułowanie i rozwiązanie przedstawionych problemów teoretycznych pozwoliło przystąpić do praktycznych zastosowań biologicznych zaproponowanej przez nas metody. Na początek pokazano, jak spośród wszystkich grup deskryptorowych wybrać niewielki podzbiór opisujący przestrzeń wszystkich białek — otrzymano w ten sposób siedem baz grup deskryptorowych. Aby wykazać ich przydatność

do opisu występujących w przyrodzie białek, użyto każdej z baz do rekonstrukcji struktury białek z pewnego zbioru testowego w oparciu o przypisania strukturalne. Pokazano, jak przypisania te mogą być użyte do przewidywania sekwencji białek na podstawie ich struktury (uzyskane wyniki pozwalają myśleć o prezentowanej metodzie przewidywania sekwencji białek jako o pierwszym kroku w kierunku projektowania leków białkowych). Następnie przeanalizowano, jak rozważane bazy grup deskryptorowych i różne funkcje sygnału pozwalają przewidywać lokalną strukturę białek ze zbioru testowego. Na koniec przedstawiono główny wynik eksperymentalny niniejszej pracy: przewidywanie struktury przestrzennej białek na podstawie ich sekwencji — wykorzystano do tego najlepszą (w świetle przeprowadzonych testów) z rozważanych baz grup deskryptorowych i najlepszą funkcję sygnału. Przykładową predykcję uzyskaną naszą metodą przedstawia rysunek 2.



Rysunek 2: Przykładowa predykcja domeny o symbolu $T0202\_1$ (z eksperymentu CASP 6): kolorem czerwonym zaznaczono prawdziwą strukturę, a kolorem granatowym — strukturę przewidzianą przy użyciu metody lokalnych deskryptorów.

Otrzymane w całkowicie zautomatyzowany sposób predykcje struktury białek porównano z predykcjami otrzymanymi przez inne zespoły badawcze. Okazało się, że w przypadku „trudnych" białek (dla których nie jest możliwe wskazanie wzorca strukturalnego przy pomocy technik sekwencyjnych) uzyskane przez nas wyniki porównywalne są z uzyskiwanymi przez najlepsze serwery (dla 8 z 23 białek testowych nasze predykcje okazały się lepsze niż predykcje serwera Robetta

[7], uchodzącego za zdecydowanie najlepszy serwer przewidujący strukturę „trudnych" białek). Choć przedstawione wyniki przewidywania struktury przestrzennej białek uznać należy za bardzo wartościowe, to jednak nie należy ich traktować jako wyników ostatecznych prezentowanej metody. Rozważać można bowiem kilka modyfikacji stosowanego podejścia, które choć nie zostały uwzględnione w niniejszej pracy, powinny owocować jeszcze lepszym przewidywaniem konformacji białek.

Literatura

[1] R. Sanchez, A. Sali (1997). Advances in comparative protein-structure modeling. *Structural Biology* 7, 206–214.

[2] C. Venclovas (2001). Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. *Proteins Suppl.* 5, 47–54.

[3] R. Thiele, R. Zimmer, T. Lengauer (1999). Protein Threading by Recursive Dynamic Programming. *J. Mol. Biol.* 290, 757–779.

[4] A. Koliński, M. R. Betancourt, D. Kihara, P. Rotkiewicz, J. Skolnick (2001). Generalized Comparative Modeling (GENECOMP): A Combination of Sequence Comparison, Threading, and Lattice Modeling for Protein Structure Prediction and Refinement. *PROTEINS: Structure, Function, and Genetics* 44, 133–149.

[5] C. Bystroff, D. Baker (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281, 565–577.

[6] T. R. Hvidsten, A. Kryshtafovych, J. Komorowski, K. Fidelis (2003). A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics* 19, II81–II91.

[7] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. Strauss, D. Baker (2001). ROBETTA in CASP4: Progress in ab initio protein structure prediction. *Proteins Suppl.* 5, 119–126.