Metody wspomagania modelowania struktury trzeciorzędowej białek za pomocą predyktorów miejsc kontaktowych

1. Wstęp

Aby zrozumieć sposób funkcjonowania danego białka wymagana jest wiedza o jego strukturze. Znajomość struktury jest m.in. konieczna w procesie projektowania leków. Niestety spośród wszystkich poznanych sekwencji aminokwasowych jedynie około 0.15% posiada również znaną strukturę. Wynika to przede wszystkim z tego, że wykorzystywane w tym celu metody eksperymentalne takie jak krystalografia rentgenowska, spektroskopia NMR, czy mikroskopia elektronowa są czasochłonne i kosztowne. Rozwiązania poszukuje się więc w wykorzystaniu komputerowych metod odtwarzania struktury białek.

Jednymi z najpopularniejszych metod używanych w komputerowej predykcji struktur białkowych są metody przewidujące tzw. miejsca kontaktowe, które następnie wykorzystywane są podczas odtwarzania struktury. Miejsce kontaktowe jest zazwyczaj zdefiniowane jako dwa aminokwasy, których atomy C_β (C_α w przypadku glicyny) są od siebie oddalone o odległość niewiększą niż 8 Å i oddzielone w sekwencji przynajmniej czterema aminokwasami.

Istnieje wiele metod predykcji miejsc kontaktowych. Najskuteczniejsze obecnie są metody wykorzystujące tzw. analizę sprzężeń bezpośrednich (z ang. direct coupling analysis, DCA) mutacji skorelowanych występujących w dopasowaniu wielosekwencyjnym (z ang. Mutliple Sequence Alignment, MSA) analizowanej sekwencji białka. Główne założenie takich metod mówi, że jeżeli dwie pozycje w sekwencji białka są istotne dla funkcji lub struktury tego białka, to gdy aminokwas na jednej z tych pozycji mutuje, musi dojść do mutacji również na drugiej pozycji tak, aby zachować istotną interakcję pomiędzy tymi pozycjami. W związku z tym, że miejsca kontaktowe reprezentują punkty w strukturze białka istotne dla jej zachowania, to ich występowania pomiędzy konkretnymi pozycjami w sekwencji także może być wskazane poprzez wystąpienie mutacji skorelowanych.

Niestety, nawet najlepsze obecnie metody wykorzystujące DCA osiągają skuteczność na poziomie 40% dla stu przewidzianych miejsc kontaktowych. Nie jest to wystarczające dla wiarygodnego, automatycznego odtwarzania struktury białka z przewidzianych kontaktów. Poza tym, raportowana skuteczność predykcji takich metod jest wartością średnią dla analizowanego zbioru białek. Dla konkretnego białka może ona osiągać wartość bardzo odmienną od 40%. Jest to szczególnie problematyczne gdy przewidywane miejsca kontaktowe są głównym źródłem danych w badaniach nad strukturą danego białka. Celem pracy doktorskiej było zatem zwiększenie wiarygodności i możliwości wykorzystania przewidzianych miejsc kontaktowych.

2. Wyniki

Pierwszym etapem badań była analiza charakterystyk miejsc kontaktowych w białkach w odniesieniu do definicji miejsca kontaktowego oraz grupy strukturalnej białka. Zaproponowano parametry, które opisywały powinowactwo różnych typów aminokwasów do tworzenia miejsc kontaktowych w dwóch grupach strukturalnych: klasach i topologiach. Jednym z wyników było wskazanie takich par aminokwasów, które najbardziej różnicują klasy strukturalne Alfa i Beta pod względem częstości tworzenia miejsc kontaktowych. Ponadto, charakterystyki miejsc kontaktowych w tych klasach zostały porównane z charakterystykami ich pojedynczych topologii. W efekcie wskazano i omówiono topologie, które najbardziej odbiegają charakterystykami miejsc kontaktowych od swoich klas.

Druga część badań poświęcona była zastosowaniu charakterystyk miejsc kontaktowych w ich predykcji. Zaproponowano algorytm mający zwiększyć skuteczność działania metody mfDCA. Algorytm miał analizować charakterystyki miejsc kontaktowych w 200 przewidywanych kontaktach dla danego białka i dopasowywać je do tych charakterystyk dla klasy strukturalnej do jakiej to białko należy. Wyniki pokazały, że zastosowanie opisanego algorytmu do wyników mfDCA może zwiększyć skuteczność predykcji tej metody nawet o średnio 5% dla 10 kontaktów przewidzianych z największym prawdopodobieństwem. Poza tym największa poprawa skuteczności została zanotowana dla białek pochodzących z klasy Alfa. Wyniki tych badań, opublikowane w pracy (**Wozniak i Kotulska, 2014**), udowodniły tym samym **pierwszą tezę** pracy doktorskiej mówiącą o tym, że *wykorzystanie charakterystyk miejsc kontaktowych może przyczynić się do poprawy skuteczności ich predykcji*.

W kolejnych badaniach pokazano, że DCA może wspomóc wybór prawidłowego modelu strukturalnego białka spośród struktury prawidłowej i fałszywej. Miejsca kontaktowe zostały przewidziane za pomocą jednej z najlepszych obecnie metod opartych na DCA - gplmDCA. Krokiem decyzyjnym w wyborze prawidłowej struktury było porównanie ilości poprawnie przewidzianych kontaktów dla każdej z analizowanych struktur - struktury natywnej i nieprawidłowego modelu. Upraszczając, jako prawidłowa była wskazywana ta struktura, która miała więcej kontaktów przewidzianych poprawnie. Wyniki pokazały, że zaproponowana metoda bardzo skutecznie wskazuje prawidłową strukturę białka spośród dwóch modeli. W szczególności, gdy struktura nieprawidłowa została zaprojektowana specjalnie na potrzeby eksperymentu, opublikowana metoda wykazywała 100% skuteczności. Działanie metody zostało również ocenione na zbiorze białek z bazy PDB, które po kilku latach po opublikowaniu zostały zastąpione przez ich dokładniejsze, bardziej prawidłowe struktury. Ponadto, w publikacji przeanalizowano te pary białek z bazy PDB, dla których metoda zawiodła. Analiza struktur tych białek pokazała, że aktualna struktura dostępna w bazie PDB wcale nie musi być tą poprawniejszą. Metoda opublikowana w (**Wozniak i in., 2017A**) może być szczególnie przydatna gdy w pracy eksperymentalnej należy wybrać poprawniejszą wersję struktury białka spośród dostępnych modeli. Wyniki udowodniły **drugą tezę** pracy doktorskiej mówiącą o tym, że *przewidziane miejsca kontaktowe mogą rozróżnić strukturę prawidłową od fałszywego modelu*.

Ostatnim etapem pracy doktorskiej było zaprojektowanie modeli regresyjnych do przewidywania skuteczności predykcji miejsc kontaktowych dla konkretnego białka. Przewidywana była skuteczność dwóch metod opartych na DCA: gplmDCA i PSICOV. Zbadano różne modele oraz ich kombinacje. Modele zbudowano na parametrach opisujących cechy danego białka takie jak MSA, przewidziana struktura drugorzędowa, przewidziana dostępność rozpuszczalnika oraz wartość korelacji zwrócona przez DCA. Najwyższą, bardzo wysoką skuteczność predykcji zaprezentował model będący średnią z Lasso, SVM, oraz lasów losowych. Dla predykcji 200 miejsc kontaktowych model ten przewidział skuteczność gplmDCA z dokładnością około 7 punktów procentowych. Zaprojektowane modele okazały się również skuteczne dla tzw. meta-metod jak np. RaptorX. W badaniach określono także istotność poszczególnych parametrów na skuteczność modeli. Badania opublikowane w pracy (**Wozniak i in., 2017B**) mogą istotnie wpłynąć na wiarygodność wykorzystywania przewidywanych miejsc kontaktowych w odtwarzaniu struktur białkowych. Ponadto wyniki potwierdzają **trzecią tezę** pracy doktorskiej mówiącą o tym, że *skuteczność predykcji miejsc kontaktowych dla konkretnego białka może zostać wcześniej przewidziana*.

3. Podsumowanie

Wszystkie tezy pracy doktorskiej zostały potwierdzone. Po pierwsze zaproponowano procedurę poprawy skuteczności predykcji miejsc kontaktowych z wykorzystaniem ich charakterystyk. Po drugie, udowodniono, że przewidywane miejsca kontaktowe mogą być wykorzystane w procesie wyboru struktury prawidłowej spośród dostępnych modeli. Po trzecie, zaprojektowano metodę umożliwiającą określenie skuteczności predykcji miejsc kontaktowych przez metody oparte na DCA dla konkretnego białka.

Wszystkie przeprowadzone w ramach pracy doktorskiej badania przyczyniły się bardzo istotnie do zwiększenia użyteczności miejsc kontaktowych w modelowaniu struktur białkowych. Aby jednak modelowanie to było możliwe w sposób zautomatyzowany, badania muszą być kontynuowane.