

mgr Aleksandra Elżbieta Badaczewska-Dawid
Pracownia Teorii Biopolimerów
Wydział Chemii
Uniwersytet Warszawski

Autoreferat pracy doktorskiej

***Modelowanie struktury i dynamiki białek z użyciem modeli gruboziarnistych
o różnej skali rozdzielczości***

promotor: dr hab. Dominik Gront

Opisanie struktury przestrzennej i dynamiki białek ma fundamentalne znaczenie dla wielu dziedzin współczesnej biologii, biomedycyny, a w szczególności dla racjonalnego projektowania nowych leków i wyjaśniania złożonych mechanizmów procesów komórkowych. Liczba danych eksperymentalnych, jaką dotychczas zgromadzono o różnych systemach białkowych jest ogromna. Rozwój metod eksperymentalnych oraz postęp technologiczny powoduje, że dane te są wciąż gromadzone w coraz szybszym tempie także dla układów, których badanie do tej pory było kłopotliwe lub nieosiągalne (białka błonowe, białka inherentnie nieuporządkowane, duże kompleksy). Struktura przestrzenna została poznana dla zaledwie 0.1% białek o znanej sekwencji, a ponad 5000 rodzin białkowych w bazie Pfam wciąż nie posiada swojego strukturalnego reprezentanta. Dla pozostałych rodzin poznano przynajmniej jedną strukturę przestrzenną.

Eksperymentalne oznaczanie struktury wielu białek jest kłopotliwe i kosztowne, co skłoniło do poszukiwania metod teoretycznych, pozwalających przewidzieć strukturę z wykorzystaniem strukturalnego szablonu (modelowanie porównawcze) lub wyłącznie na podstawie sekwencji aminokwasowej (modelowanie bezpośrednie). Pierwsze podejście wykorzystuje obserwację, że ewolucyjnie powiązane białka zachowują wysokie podobieństwo strukturalne, zatem pojedyncze znane struktury mogą być użyte jako szablony do modelowania innych białek z rodziny. W modelowaniu bezpośrednim nie wykorzystuje się szablonów struktury przestrzennej, ale za pomocą różnych potencjałów statystycznych zależnych od sekwencji aminokwasowej często uwzględniane są ogólne prawidłowości geometryczne, obserwowane w znanych strukturach białek.

Stałym wyzwaniem każdej strategii komputerowego modelowania jest wydajne próbkowanie przestrzeni konformacyjnej, realizowane zwykle w różnych wariantach dynamiki molekularnej lub dynamiki Monte Carlo, często w schemacie wymiany replik zwiększającej efektywność algorytmu. Jednak tylko dla podzbioru stosunkowo małych białek możliwe jest obecnie obliczeniowe przewidywanie struktury przestrzennej bezpośrednio w rozdzielczości pełnoatomowej. Jest to powód rosnącej roli metod modelowania gruboziarnistego (ang. coarse-grained, CG). Niektóre z najbardziej udanych modeli CG są ukierunkowane na przewidywanie struktury, a w połączeniu z odpowiednimi narzędziami bioinformatycznymi mogą być bardzo skuteczne (ROSETTA, I-TASSER). Inne modele są bardziej uniwersalne i umożliwiają nie tylko przewidywanie struktury, ale także badanie dynamiki białek. Typowymi przykładami takich modeli CG o średniej rozdzielczości są UNRES lub CABS, zastępujące pojedynczą resztę aminokwasową kilkoma (2 - 4, zależnie od typu reszty) zjednoczonymi atomami. Taki poziom gruboziarnistego uproszczenia zapewnia przyspieszenie obliczeniowe o 3 - 4 rzędy wielkości w porównaniu

z symulacjami pełnoatomowymi, pozwalając jednocześnie na realistyczną rekonstrukcję szczegółów atomowych. Wyniki konkursu CASP dowodzą, że modelowanie molekularne struktury białka z wykorzystaniem uproszczonej reprezentacji oraz wiarygodnej rekonstrukcji jest skuteczne.

Wyznaczanie trójwymiarowych struktur pojedynczych cząsteczek białka, zarówno eksperymentalne i obliczeniowe, stanowi tylko niewielką część biologii strukturalnej. Potrzeba również precyzyjnej wiedzy na temat mechanizmów zwijania się białek, dynamiki cząsteczek i ich wzajemnego oddziaływania. Struktury i właściwości dynamiczne takich kompleksów są znacznie większym wyzwaniem dla badań eksperymentalnych i obecnie stają się ważnym kierunkiem rozwoju metod modelowania obliczeniowego. Sprawdzenia skuteczności narzędzi w tym zakresie dokonuje się w ogólnosiwiatowym eksperymencie CAPRI, oceniającym przewidziane struktury kompleksów białko-białko. Ze względu na skalę czasową i rozmiar układów biomakrocząsteczkowych konieczne staje się projektowanie efektywnych modeli o niskiej rozdzielczości, takich jak prezentowany w tej rozprawie SURPASS, które umożliwiają symulacje dużych układów oraz mogą być włączone w zintegrowany protokół modelowania wieloskalowego. Takie podejście jest nakierowane na badanie kompleksów białkowych oraz oddziaływań biomakrocząsteczkowych zmierzające do pełnego opisu procesów zachodzących na poziomie całej komórki.

Przedmiotem pracy było modelowanie struktury i dynamiki białek z użyciem modeli gruboziarnistych o różnej skali rozdzielczości. Wykorzystano istniejące modele gruboziarniste średniej rozdzielczości, takie jak CABS i oparty na strukturze model elastycznej sieci (DynOmics) oraz stworzono nowy model o niskiej rozdzielczości – SURPASS. Jako dane referencyjne posłużyły także trajektorie z symulacji dynamiką molekularną w reprezentacji łańcucha zbudowanego z węgli alfa oraz liczne znane struktury pełnoatomowe białek i inne dane eksperymentalne (np. więzy NOE). Głównym celem pracy było stworzenie nowego, wydajnego modelu gruboziarnistego o niskiej rozdzielczości, pokonującego pewne ograniczenia istniejących narzędzi (skala czasu, rozmiar układu). W ramach realizacji tego celu stworzono gruboziarnistą reprezentację modelu SURPASS o niskiej rozdzielczości, wyprowadzono zestaw potencjałów statystycznych opisujących podstawowy schemat oddziaływań pomiędzy pseudo atomami w układzie, zoptymalizowano wagi poszczególnych członów pola siłowego opartego na wiedzy oraz zaproponowano algorytm rekonstrukcji położenia węgli alfa z silnie uproszczonej reprezentacji SURPASS. Drugim ważnym celem była ocena skuteczności i wydajności modelu SURPASS w modelowaniu struktury i dynamiki białka. Ewaluację przeprowadzono na podstawie analizy wyników długich symulacji modelowania de novo w schemacie próbkowania dynamiką Monte Carlo z wymianą replik dla zestawu kilkudziesięciu jednodomenowych białek globularnych. W tym eksperymencie szczegółowo oceniono globalną dynamikę układu w procesie zwijania od łańcucha całkowicie rozwiniętego do struktury około natywnej oraz zdolność tworzenia konformacji białkopodobnych (w tym frakcji struktur około natywnych). Dodatkowo zbadano wpływ dokładności założonej struktury drugorzędowej na wynik modelowania. W innym eksperymencie obliczeniowym porównano elastyczność konformacyjną białka w pobliżu stanu natywnego przewidzianą przez różne modele gruboziarniste oraz pełnoatomową dynamiką molekularną dla zestawu ponad stu różnorodnych białek.

W pracy doktorskiej przedstawiono gruboziarnisty model SURPASS oraz jego zastosowania do modelowania molekularnego struktury i dynamiki białek. W szczególności zaproponowano unikalną, silnie uproszczoną reprezentację struktury białek i wyprowadzono zestaw potencjałów statystycznych tworzących oparte na wiedzy pole siłowe SURPASS. Koncepcja modelu zakłada uśrednianie krótkich 4-resztowych fragmentów struktury drugorzędowej. Specyficzny schemat oddziaływań odróżnia białkopodobny łańcuch SURPASS od losowego polimeru. Potencjały statystyczne opisują lokalne

prawidłowości strukturalne charakterystyczne dla większości białek globularnych. Wagi składowych funkcji energii oraz parametry programu symulacyjnego dla różnych strategii próbkowania przestrzeni stanów zostały zoptymalizowane podczas długich symulacji testowych. W celu zwiększenia użyteczności modelu opracowano nietrywialny algorytm rekonstrukcji położeń węgli alfa z reprezentacji SURPASS. Jego nieodłączną część stanowi biblioteka par fragmentów pochodzących ze znanych struktur białek (SURELib). Zrekonstruowany ślad C α odtwarza lokalną geometrię łańcucha polipeptydowego i poprawną orientację przestrzenną elementów struktury drugorzędowej.

W symulacjach z użyciem modelu SURPASS oraz próbkowaniem w schemacie dynamiki Monte Carlo z wymianą replik zbadano proces zwijania jednodomenowych białek globularnych. Analiza wyników pozwoliła ocenić skuteczność narzędzia w tworzeniu białkopodobnych konformacji, w tym szczególnie identyfikacji struktur około natywnych. W badaniach tych jedyną informacją zależną od sekwencji aminokwasowej była preferowana struktura drugorzędowa. Analizie poddano rolę struktury drugorzędowej w procesie zwijania się białka oraz wpływ krytycznych błędów w przyporządkowanej strukturze drugorzędowej na zdolność odtworzenia prawidłowej topologii. Uzyskane wyniki dowodzą, że struktura drugorzędowa w dużej mierze definiuje geometrię zwoju. Nawet fragmentaryczne informacje o preferowanej strukturze drugorzędowej są wystarczające do skutecznej identyfikacji topologii natywnej. Model SURPASS odtwarza podstawowe właściwości strukturalne białek, a dokładność otrzymanych modeli jest zaskakująco dobra zarówno w przypadku małych, jak i znacznie większych układów. Dodatkową poprawę dokładności modelu można osiągnąć dzięki wykorzystaniu cząstkowych danych doświadczalnych, co wykazano w symulacjach z wiązami NOE.

Przeprowadzono także eksperymenty symulacyjne, mające na celu porównanie obrazu dynamiki struktury białka w pobliżu stanu natywnego odtworzonego przez modele gruboziarniste średniej (CABS) i niskiej (SURPASS) rozdzielczości, model elastycznej sieci (DynOmics) i pełnoatomową dynamikę molekularną. Zarejestrowane profile fluktuacji zostały dodatkowo odniesione do danych eksperymentalnych NMR, z którymi najwyższą zgodność wykazały wyniki metod ENM i CABS. Model SURPASS okazał się mniej dokładny od pozostałych metod, choć uzyskane profile fluktuacji dość realistycznie odtworzyły lokalną dynamikę struktury. Przeprowadzone badania dowiodły, że mimo silnie uśrednionego charakteru, model SURPASS bardzo dobrze odtwarza podstawowe właściwości strukturalne białek.

Model SURPASS umożliwia bardzo efektywne próbkowanie całej przestrzeni konformacyjnej białka. Pozwala też przewyciężyć ograniczenia gruboziarnistych modeli o średniej rozdzielczości, które są nadal zbyt kosztowne obliczeniowo, aby efektywnie modelować duże układy biocząsteczkowe. Zaproponowany model może zostać łatwo włączony do różnych strategii modelowania wieloskalowego. Otworzy to możliwość efektywnego modelowania integracyjnego dużych układów, w tym modelowania de novo i odległego modelowania porównawczego białek i ich kompleksów. Takie hierarchiczne podejście powinno być odpowiednio dostosowane do konkretnych problemów i wykorzystywać fragmentaryczne dane strukturalne dostarczane przez krystalografię rentgenowską, NMR, cryo-EM, SAXS, HDX i inne metody eksperymentalne.

Praca doktorska była realizowana w ramach projektu: TEAM: "*Development and integration of new multiscale modeling tools for molecular biology: structure, dynamics and thermodynamics*", finansowanego przez FNP oraz MAESTRO: "*New and efficient methods for the flexible molecular docking of proteins*", finansowanego przez NCN. Wyniki zawarte w rozprawie doktorskiej zostały opublikowane w pracach oryginalnych oraz przeglądowych w czasopismach *J. Chem. Theory Comput.*, *Chem. Rev.* oraz *Int J Mol Sci.*