

Streszczenie rozprawy doktorskiej

Maciej Dziubiński

“Development of selected mesoscopic physical models with the aid of machine learning methods and their applications in studies of molecular systems”,

w języku polskim,

“Rozwijanie wybranych mezoskopowych modeli fizycznych z wykorzystaniem metod uczenia maszynowego i ich zastosowania w badaniach układów molekularnych”

Wstęp i cel rozprawy

Przedmiotem niniejszej rozprawy było opracowywanie i wykorzystanie metod uczenia maszynowego bez nadzoru do wybranych problemów w dziedzinach teoretycznej biofizyki molekularnej i bioinformatyki. Uczenie maszynowe dostarcza teoretycznych i informatycznych narzędzi pozwalających na wyodrębnianie i filtrowanie danych w celu uzyskania istotnych z merytorycznego punktu widzenia informacji pochodzących z dużych, wielowymiarowych zbiorów danych, generowanych w szczególności podczas symulacji i eksperymentalnych analiz złożonych układów. Jednakże metody uczenia maszynowego wymagają dostosowania do złożonych i częściowo chaotycznych danych spotykanych w dziedzinach teoretycznej biofizyce obliczeniowej i w biologii strukturalnej. Znaczna część tej pracy poświęcona jest adaptacji standardowych algorytmów uczenia bez nadzoru do wybranych problemów z w/w obszarów.

W niniejszej pracy wykorzystano metodologie uczenia bez nadzoru do dwóch problemów. Pierwszy z nich polegał na identyfikacji względnie sztywnych fragmentów strukturalnych podczas wewnątrzcząsteczkowych ruchów atomów w białkach, zaś drugi dotyczył identyfikacji i analizy kooperatywnych mechanizmów wewnątrz cząsteczki, napędzającej przemiany konformacyjne – w obydwu przypadkach symulowanych metodami dynamiki molekularnej. Obydwa te problemy odnosiły się do dynamicznych właściwości układów (bio)molekularnych a analizy przedstawione w tej rozprawie pozwoliły na stworzenie niżej wymiarowych i bardziej poglądowych modeli tych zjawisk.

W szczególności, przedstawiane wykorzystanie uczenia maszynowego bez nadzoru pozwoliło na lepsze wytłumaczenie zawłości ukrytych w pozornie chaotycznych danych z symulacji metodami dynamiki molekularnej. Choć przedstawione metody objaśniają złożone zjawiska, zwrócona została również uwaga na potencjalne trudności związane z ich użyciem do problemów biofizycznych.

Współczesne metody badań układów (bio)molekularnych (zarówno symulacyjne jak i eksperymentalne) dostarczają ogromnych ilości danych. Dwa istotne powody, dla których obserwujemy ten trend to:

- układy będące przedmiotem badań stają się coraz większe,
- transformacje strukturalne tych układów stanowią zawile ścieżki w wielowymiarowych przestrzeniach konfiguracyjnych.

Obydwa przyczyny stawiają coraz wyższe wymagania narzędziom analitycznym. Potrzebne są metody które

mogłyby ujawniać ukryte w danych wartościowe z merytorycznego punktu widzenia informacje, pozwalające lepiej zrozumieć obserwowane zjawiska i formułować dalsze hipotezy badawcze.

Celem rozprawy było opracowanie optymalnej metodologii i narzędzi pozwalających stworzyć uproszczone opisy złożonych, dynamicznych układów molekularnych. Wykorzystano metody uczenia maszynowego bez nadzoru do dwóch problemów.

Pierwszy z nich to identyfikacja dynamicznych domen gdzie danymi na wejściu były konfiguracje pozyskane z symulacji metodami dynamiki molekularnej, bądź w eksperymentach NMR. Choć struktura większości białek podlega ciągłym zmianom, wiele z nich dysponuje fragmentami strukturalnymi, których kształt pozostaje względnie sztywny. Jednym ze sposobów uproszczenia opisu ruchu takich białek jest identyfikacja takich fragmentów, często określanych mianem dynamicznych domen. Wówczas opis transformacji strukturalnych zachodzących w białkach, lub w innych (bio)polimerach, możemy wyrazić poprzez wzajemne transformacje owych dynamicznych domen, co znacząco ułatwia zrozumienie zachodzących w tych układach przemian.

Drugim problemem rozważanym w tej pracy jest identyfikacja podukładów molekularnych, odgrywających istotną rolę w przemianach strukturalnych całego układu. Tego typu podukłady nazwano w tej pracy molekularnymi trybami (ang. *molecular cogs*). Rozwiązanie tego problemu było bardziej wymagające, przez co układami podlegającymi przedstawionej analizie były stosunkowo niewielkie cząsteczki. Projekt ten nastawiony był głównie na pokonywanie koncepcyjnych i metodologicznych trudności, które mogłyby stanowić blokadę w zastosowaniu do większych układów takich jak białka.

Wprowadzające uwagi merytoryczne

Rozwiązania obydwu problemów bazuje na połączeniu metod uczenia maszynowego bez nadzoru i wiedzy na temat fizycznych właściwości układów (bio)molekularnych. Własności fizyczne zostały wykorzystane do sformułowania odpowiednich miar zależności między obiektami, które następnie łączone były w grupy, nazywane też *klastrami*. Chociaż metody uczenia maszynowego stają się coraz efektywniejsze, bez wykorzystania informacji o fizycznych właściwościach badanych układów (bio)molekularnych nawet najbardziej zaawansowane algorytmy klastrujące okazują się być niewystarczające.

Dynamiczne domen

ResiCon – stworzony do identyfikacji dynamicznych domen – bazował na kilku pomysłach, z których dwa okazały się kluczowe. Pierwszy polegał na wykorzystaniu koncepcji kontaktów pomiędzy resztami aminokwasowymi, po to by wyrazić miarę wewnętrznej zmienności strukturalnej białek w kontekście ciągłych przemian strukturalnych symulowanych metodami MD. Drugim istotnym pomysłem było zastosowanie odpowiedniego algorytmu klastrującego, wykorzystującego spektralne własności macierzy akumulującej niezidentyfikowane informacje o zależnościach między obiektami (w tym przypadku między aminokwasami). Wynikiem klasteryzacji były grupy aminokwasów, które w trakcie transformacji

strukturalnych białka pozostawały względem siebie w stosunkowo stałych relacjach przestrzennych.

Zaproponowana miara zależności wykorzystana w ResiConie miała na celu uchwycić nie tylko relacji przestrzenne, ale również „siłę oddziaływań” między resztami aminokwasowymi. W Rozdziale 2 pokazujemy, że klasteryzacja wykorzystująca tę miarę prowadzi nie tylko do spójnych dynamicznych domen, ale pozwala również na zaobserwowanie ciekawych dynamicznych motywów w transformacjach strukturalnych. Zauważono m.in. że w wynikach MD dla proteazy HIV-1 pojedyncze aminokwasy były przydzielane do innych klastrow niż ich bezpośredni sekwencyjni sąsiedzi. Okazało się, że nie był to artefakt algorytmu klasteryzującego, ale że aminokwasy te tworzą pewnego rodzaju „zawias”, względem którego występują znaczne zmiany konformacyjne.

Warto również wspomnieć, że algorytm klasteryzujący wykorzystany w ResiConie umożliwiał ocenienie, jaka liczba dynamicznych domen jest optymalna dla danego białka. Było to istotne, gdyż w dwóch dotychczas stosowanych metodach – GeoStaS i PiSQRD – nawet najmniejsze zmiany w parametrach tych metod prowadzą do wyników o znacznie różnych liczbach klastrow. W wielu przykładach prezentowanych w Rozdziale 2 (i w towarzyszących mu materiałach uzupełniających) konkurencyjne metody prowadziły do absurdalnych dynamicznych domen (składających się np. z pojedynczych aminokwasów), podczas gdy ResiCon w sposób konsystentny identyfikował spójne, sztywne dynamiczne domeny, bądź wskazywał na brak takowych. Serwis ResiCon jest dostępny pod adresem: <http://dworkowa.imdik.pan.pl/EP/ResiCon>.

Tryby molekularne

W Rozdziale 3 przedstawiona została nowa metodologia pozyskiwania informacji o wewnętrznych mechanizmach małych cząsteczek prowadzących do określonych przemian strukturalnych. Badania skoncentrowane zostały na pokazaniu, że jest to obiecujący front badań i starano się odkryć wszelkie trudności koncepcyjne i techniczne które mogłyby stanowić przeszkodę w wykorzystaniu tego typu podejścia do analizy większych układów (bio)molekularnych. Zaproponowana miara zależności tłumaczyła obserwowane w mikroskopowych symulacjach MD przemiany, a dodatkowo w kilku przypadkach podkreśliła bardziej subtelne właściwości badanych cząsteczek. Jednak, co warto podkreślić, miara zależności wykorzystana do detekcji trybów molekularnych posiada jednoznaczną interpretację w kontekście zmiany energii swobodnej związanej z przemianą strukturalną.

Uzyskano zatem dobrą miarę zależności, która połączona została z odpowiednim algorytmem klastrującym. W celu znalezienia trybów molekularnych, potrzebna była odpowiednia funkcja celu, do minimalizacji której wykorzystano genetyczny algorytm klastrujący. Efektem końcowym był podział cząsteczki na dwa podukłady, których wewnętrzne oddziaływania podnosiły, bądź obniżały energię swobodną związaną z transformacją całego układu.

Podsumowanie

W niniejszej pracy rozwinięto i zastosowano metody uczenia maszynowego w wybranych zagadnieniach teoretycznej biofizyki molekularnej i bioinformatyki (identyfikacja dynamicznych domen w wybranym układzie enzymatycznym oraz molekularnych trybów w niewielkim, modelowym układzie molekularnym) jak również zaprezentowano korzyści związane z wykorzystaniem tych metod do analizy danych pochodzących z symulacji metodami dynamiki molekularnej. Przedstawiano jednocześnie pewne metodologiczne trudności, które pojawiły się w trakcie realizowanych badań. Rozdziały 2 i 3 omawiają trzy główne wyzwania związane z badaniami tego typu:

1. wybór/definicja odpowiedniej miary zależności,
2. wybór odpowiedniego algorytmu klasteryzującego oraz wybór i/lub optymalizacja liczby klastrów,
3. metodologia weryfikacji wyników.

Miary zależności jakie wykorzystano w obydwu rozdziałach bazowały na właściwościach fizycznych badanych układów. Był to kluczowy element nie tylko dlatego, że prowadził do wysokiej jakości klastrów, ale również dlatego że pozwalał na bardziej wiarygodną interpretację uzyskanych wyników. Drugim istotnym elementem był wybór właściwego algorytmu klasteryzującego dostosowanego zarówno do miary zależności jak i do zagadnień, do których był używany. Jednakże bez trzeciego elementu – zewnętrznej metody weryfikacji uzyskanych klastrów – wiarygodna ocena końcowych wyników nie byłaby w pełni możliwa. Te trzy współzależne elementy są nieodzowne w rozwiązywaniu problemów w w/w dziedzinach przy użyciu metod uczenia maszynowego bez nadzoru.

Maciej Dziubiński