Struktura pierwszorzędowa białek i kwasów nukleinowych to ciąg reszt (odpowiednio - aminokwasowych lub nukleotydowych), które tworzą daną sekwencję biologiczną. To najprostszy sposób reprezentacji sekwencji, pomijający wyższe poziomy organizacji, takie jak ułożenie przestrzenne łańcuchów reszt względem siebie. Struktura pierwszorzędowa pozostaje jedyną znaną informacją na temat większości białek zdeponowanych w bazach danych i z tego powodu jest często wykorzystywana do przewidywania właściwości tych sekwencji bez konieczności przeprowadzania kosztownych eksperymentów. Algorytmy te działają tym lepiej, im efektywniej uzyskują informacje z sekwencji biologicznych.

W mojej pracy doktorskiej zajmuję się problemem efektywnego dekodowania informacji zawartych w sekwencjach białek i kwasów nukleinowych. Wykorzystam do tego dwie metody: macierzy n-gramowych i uproszczonych alfabetów aminokwasowych. Opracowaną metodologię planuję zastosować w analizie dwóch problemów biologicznych: przewidywaniu lokalizacji subkomórkowej i amyloidogenności białek.

Macierze n-gramowe, to sposob reprezentacji sekwencji biologicznej za pomocą zliczeń występujących w nich krótkich sekwencji o długości n (stąd nazwa n-gramy). Przykładowo, dla alfabetu nukleotydowego (A, C, G, T) w sekwencji ATATA występują następujące 3-gramy: dwukrotnie ATA i raz TAT. Pozostałe możliwe 3-gramy (AAA, AAC itd.) nie występują w tej sekwencji. Mankamentem takiej reprezentacji sekwencji jest duża liczba mało informatywnych n-gramów, które muszą zostać usunięte za pomocą czasochłonnych algorytmów z reguły opartych na testach permutacyjnych. W mojej pracy doktorskiej wprowadzam nowy test nazwany QuiPT pozwalający na precyzyjniejsze i szybsze wyszukiwanie informatywnych n-gramów.

Z kolei uproszczone alfabety aminokwasowe polegają na grupowaniu aminokwasów pod względem ich podobieństwa fizykochemicznego lub biochemicznego. Aby wykorzystać je w reprezentacji sekwencji, oryginalne aminokwasy zastępuje się indeksami grup, do których je przypisano. W ten sposób wprowadza się do modelu nową informację biologiczną o podobieństwie aminokwasów, co pozwala na tworzenie prostszych i skuteczniejszych modeli rozpoznających różne grupy sekwencji o specyficznych funkcjach czy strukturze. Tego typu podejście było już z sukcesem wykorzystywane np. przy przewidywaniu struktury białek. Jednakże wciąż nie ma gotowych do użycia algorytmów pozwalających na generowanie optymalnie uproszczone alfabety. W trakcie moich badań rozwinę dotychczasowe metody redukcji alfabetu i zaproponuję kolejne oparte na algorytmach genetycznych.

Lokalizacja subkomórkowa białka to jego ostateczne miejsce przeznaczenia w komórce, które można przewidzieć na podstawie obecności w sekwencji charakterystycznych sygnałów kierujących. Sygnały te są dość zróżnicowane, a ich prawidłowe funkcjonowanie nie zależy od obecności konkretnych aminokwasów, ale od ich właściwości fizykochemicznych. Stosując zaproponowaną przeze mnie metodologię stworzyłem zmodyfikowaną wersję najpopularniejszego programu do wykrywania peptydów sygnałowych - SignalP 4.1. Po zastosowaniu uproszczonego alfabetu, program jest w stanie precyzyjnie rozpoznawać petydy sygnałowe groźnego pasożyta zarodźca malarii, co dotychczas nie było możliwe. W trakcie prac nad moją rozprawą doktorską przygotuję bardziej uniwersalny program do określania lokalizacji subkomórkowej białek.

Drugim moim tematem badawczym będa amyloidy, które są zróżnicowaną grupą białek związanych m.in. z neurodegeneracyjnymi chorobami układu nerwowego, jak choroby Alzheimera, Parkinsona i Creutzfeldta-Jakoba. Białka te charakteryzują się zdolnością do tworzenia agregatów, co jest inicjowane w charakterystycznych regionach sekwencji. Macierze n-gramowe stworzone z użyciem skróconego alfabetu pozwoliły na opracowanie nowego predyktora amyloidogenności, który jest czulszy niż istniejące w tej chwili programy. Dodatkowo, dzięki jasnemu opisowi sekwencji przez macierze n-gramowe mogłem wyznaczyć motywy aminokwasowe występujące w regionach odpowiedzialnych za inicjację agregacji.

Wszystkie metody, które powstaną w trakcie realizacji mojej pracy doktorskiej, zostaną opublikowane w postaci otwartego oprogramowania naukowego. Nie tylko zwiększy to powtarzalność prowadzonych przeze mnie badań, ale również pozwoli innym badaczom na wykorzystanie opracowanych przeze mnie rozwiązań. Rozwijane przeze mnie narzędzia pozwolą na zbudowanie bardziej uniwersalnych i specyficznych modeli statystycznych przewidujących właściwości sekwencji biologicznych.