Język stanowi sposób w jaki nieskończona ilość znaczeń może być wyrażona z wykorzystaniem skończonej liczbę symboli oraz reguł. Już kilkadziesiąt lat temu zauważono, że sposób ten jest wspólny dla języka ludzkiego oraz biopolimerów, takich jak kwasy nukleinowe i białka. Z zaledwie dwudziestu aminokwasów (liter lub wyrazów) powstały miliony sekwencji (zdań) zwiniętych w tysiące rodzajów konformacji (syntaktyka) spełniających najróżniejsze funkcje w organizmach żywych (semantyka). Podobieństwo pomiędzy sekwencjami białkowymi a zdaniami języka naturalnego znacząco wykracza poza zbieżnośc formy zapisu jako ciągu znaków. Nić białka stanowi łańcuch aminokwasów fizycznie połączonych wiązaniami peptydowymi. Podobnie do zdań języka naturalnego, sekwencja białka może być wieloznaczna (ten sam łańcuch aminokwasów zwija się w różne struktury w zależności od otoczenia), często zawiera zależności dalekozasięgowe i struktury rekurencyjne. W szczególnym przypadku zaobserwowano zjawisko podwójnej artykulacji, specyficzne dla przeważającej większości języków ludzkich: bardzo duży zbiór możliwych powierzchni molekularnych (zdań) otrzymany z ograniczonego zestawu powtórzonych fragmentów sekwencji (słów), składających się z kilku aminokwasów poddanych pozytywnej selekcji (głosek).

Dzięki temu podobieństwu metody bazujące na teorii języków formalnych nadają się wyjątkowo dobrze do analizy sekwencji białkowych. Jednakże metody lingwistyczne używane są do badania białek przeważnie w charakterze tzw. czarnej skrzynki. Ponadto popularne modele nie potrafią uwzględnić oddziaływań pomiędzy oddalonymi od siebie aminokwasami, które są bardzo ważne dla struktury i funkcji białek. W ramach projektu będziemy pracować nad zastosowaniem do analizy białek bardziej zaawansowanego modelu: probabilistycznych gramatyk bezkontekstowych, który umożliwia na bezpośrednie ujęcie oddziaływań pomiędzy oddalonymi od siebie aminokwasami. Naszym celem jest opracowanie skutecznych metod tworzenia modelu białka na poziomie gramatyki bezkontekstowej. Trzy podstawowe zastosowania takiego modelu to: (1) przeszukiwanie bazy danych sekwencji w poszukiwaniu podobnych białek lub ich fragmentów; (2) analiza struktury automatycznie utworzonej gramatyki oraz gramatycznego rozbioru sekwencji (podobny do rozbioru zdania języka polskiego) w celu wychwycenia charakterystycznych cech i zależności - niejako poznawania "języka białek"; (3) zapisywanie różnych hipotez naukowych w formie gramatyki i porównywanie która gramatyk bardziej pasuje do przebadanych białek.

Szczególnie dwa ostatnie zastosowania były jak dotąd zaniedbywane przez bioinformatyków. Tymczasem nasze wcześniejsze prace wskazują, że ze struktury automatycznie otrzymanej (nauczonej) gramatyki bezkontekstowej lub z jej drzewa rozkładu można odczytać informacje dotyczące biologicznie istotnych cech białka. Dlatego zamierzamy systematycznie porównywać gramatykami i drzewami rozkładu z cechami opisywanego przez nie białka. Każdy model jest redukcją rzeczywistości do jakiegoś jej aspektu. Aby określić która z hipotez reprezentowanych przez probabilistyczne modele gramatyczne lepiej opisuje badany wycinek świata może wystarczyć porównanie prawdopodobieństw z jakimi modele te utworzyłyby występujące w rzeczywistości obiekty. Czasami jednak warto zastosować także inną metodę sprawdzenia, która polega na wytworzeniu bardzo dużej liczby obiektów zgodnych z modelem, a następnie porównaniu czy ich cechy charakterystyczne rozkładają się podobnie jak cechy istniejących w rzeczywistości obiektów. W takim przypadku niezbędne jest określenie mierzalnych cech charakterystycznych. Ze względu na podobieństwo pomiędzy sekwencjami białkowymi a zdaniami języka naturalnego pomocne mogą okazać się cechy służące do analizy tekstów literackich, ale także cechy używane przy opisie systemów dynamicznych.

Również w przypadku zastosowania modelu gramatycznego do przeszukiwania bazy danych możliwe jest osiągnięcie znaczącego postępu. W ostatnich latach, dzięki lawinowo rosnącej liczbie dostępnych sekwencji białkowych z różnych genomów oraz dzięki nowym metodom matematycznym, pojawiła się możliwość określenia z dużym prawdopodobieństwem, które aminokwasy w sekwencji są od siebie zależne. Jak dotąd doprowadziło to do poprawy w przewidywaniu struktury białka. W naszym przekonaniu włączenie informacji o tzw. mutacjach zależnych do procesu tworzenia gramatyki spowoduje znaczącą poprawę szybkości uczenia oraz jakości uzyskanych modeli. Byłoby to znaczące osiągnięcie w zakresie bioinformatyki.

Praktycznym efektem projektu będzie udostępnienie społeczności naukowej opracowanych w jego trakcie metod w formie pakietu programów oraz serwisu on-line. W szerszej perspektywie rozwój lingwistyki białek wydaje się nieodzowny do systematycznego projektowania peptydów i białek, czy wręcz programowania całych podsystemów molekularnych (np. immunologicznych) - w zgodzie z ich naturą.