Wnioskowanie gramatyczne jest intensywnie rozwijanym obszarem badań łączącym w sobie zagadnienia gramatyk formalnych, przetwarzania języków, uczenia maszynowego i teorii wyuczalności. Problem ten, obejmujący głównie indukcję gramatyk i automatów, jest ważny pod względem teoretycznym. Istotnymi pytaniami otwartymi, które zamierzamy badać są zagadnienia dotyczące gramatyk probabilistycznych, w szczególności roli kontrprzykładów w uczeniu tych gramatyk, a także efektywnych metod oraz algorytmów indukowania gramatyk bezkontekstowych oraz automatów niedeterministycznych. Rozważane problemy mają znaczenie praktyczne z uwagi na zastosowania w takich dziedzinach jak genomika obliczeniowa, w tym klasyfikacja danych bioinformatycznych, którą będziemy się zajmować w projekcie.

Nieformalnie, wnioskowanie gramatyczne (ang. grammatical inference) jest problemem wnioskowania indukcyjnego, w którym dziedziną jest klasa gramatyk. Uczenie rozumiane jest jako identyfikacja gramatyki nieznanego języka L z użyciem próbki uczącej $S = (S_+, S_-)$, gdzie S_+ jest zbiorem słów zwanych przykładami (ang. examples) należących do języka L, a S_- zbiorem słów zwanych kontrprzykładami (ang. counterexamples) nie należących do języka. Mówi się również, że zbiory S_+ i S_- zawierają, odpowiednio, przykłady pozytywne i negatywne.

Podstawowym celem naukowym projektu jest opracowanie metod indukcji gramatyk probabilistycznych i bezkontekstowych, grafów słownych, a także deterministycznych i niedeterministycznych automatów skończonych opisujących język L pasujący do zadanych białek amyloidowych (czyli pewnych słów nad alfabetem złożonym z 20 symboli). W projekcie opracujemy oryginalny system indukcji gramatyk probabilistycznych, weryfikując hipotezę o istotnej roli kontrprzykładów w indukcji tego typu gramatyk. Zaproponujemy metody indukowania acyklicznych gramatyk bezkontekstowych oraz automatów niedeterministycznych. Skuteczność zaproponowanych metod zweryfikujemy na drodze praktycznej przez skonstruowanie algorytmów oraz ich implementacji komputerowych (programów) realizujących procesy indukcji gramatyk i automatów. Uzyskane w ten sposób klasyfikatory posłużą także do generowania nowych sekwencji białkowych, które mogą okazać się nieznanymi jeszcze amyloidami. Taką teoretyczną weryfikację można przeprowadzić opierając się na przewidywanych fizykochemicznych właściwościach tych nowych białek o potencjalnie nieprawidłowej (z punktu widzenia funkcjonowania żywego organizmu) strukturze.

Badania projektu będą dotyczyły następujących zadań szczegółowych.

- Indukcja gramatyk probabilistycznych: (1) system indukcji gramatyk probabilistycznych, (2) rola kontrprzykładów w uczeniu się gramatyk probabilistycznych, (3) zastosowanie systemu indukcji gramatyk w klasyfikacji białek amyloidowych.

- Indukcja acyklicznych gramatyk bezkontekstowych oraz grafów słownych: (4) indukcja acyklicznych gramatyk bezkontekstowych, (5) indukcja grafów słownych, (6) zastosowanie acyklicznych gramatyk bezkontekstowych oraz grafów słownych w klasyfikacji białek amyloidowych.

- Indukcja niedeterministycznych automatów oraz dekompozycja języków skończonych: (7) indukcja niedeterministycznych automatów skończonych, (8) dekompozycja języków skończonych, (9) zastosowanie dekompozycji języków w klasyfikacji białek amyloidowych.