Grammatical inference is an intensively studied area of research that sits at the intersection of several fields including finite state machines, formal grammars, machine learning, language processing, learnability theory. Grammar and automata induction is important in theory. There are essential theoretical issues that should be investigated. For example, the role of counterexamples in learning probabilistic grammars, or complexity classes of algorithms to induce context-free grammars and nondeterministic automata. Induction problems are also of practical importance. Their solutions are applied in such fields as linguistic engineering, natural language processing, computational genomics, and in particular, in classification of amyloidogenic proteins that is going to be investigated in this project.

Informally, *grammatical inference* is a problem of inductive reasoning, whose domain is a class of grammars. Grammar learning is understood as identification of a grammar of an unknown language $L$ by using a learning sample $S = (S_+, S_-)$, where $S_+$ is a set of words, called examples, belonging to language $L$, and $S_-$ is a set of words, called counterexamples, that do not belong to language $L$. It is also said that sets $S_+$ and $S_-$ contain positive and negative examples, respectively.

The scientific objective of the project is to develop the methods for induction of context-free and probabilistic grammars, word graphs and deterministic and nondeterministic finite automata describing language $L$ matching given amyloidogenic sequences (i.e. words over an alphabet consisting of 20 symbols). We are going to develop a system of induction of probabilistic grammars and to verify a scientific hypothesis of positive impact of presence of counterexamples in a learning sample. We intend to develop methods of induction of noncircular context-free grammars, word graphs and nondeterministic automata. The proposed methods will be validated in practice by designing algorithms, including parallel ones, and their computer implementations for carrying out processes of induction. The resulting classifiers will be also applied to the process of new protein sequences generation, possibly identifying some unknown amylome. Such a theoretical verification can be conducted based on physico-chemical properties of the new proteins having potentially incorrect (from the point of view of living organism functioning) structure.

Here are the groups of tasks that will be the subject of project research:

- Induction of probabilistic grammars: (1) system of induction of probabilistic grammars, (2) role of counterexamples in learning of probabilistic grammars, (3) application of the system of grammar induction in classification of amyloidogenic proteins.

- Induction of noncircular context-free grammars and word graphs: (4) induction of noncircular context-free grammars, (5) induction of word graphs, (6) application of noncircular context-free grammars and word graphs in classification of amyloidogenic proteins.

- Induction of nondeterministic automata and decomposition of finite languages: (7) induction of nondeterministic automata, (8) decomposition of finite languages, (9) application of decomposition of languages in classification of amyloidogenic proteins.