Language is a way in which an infinite number of meanings can be expressed by a finite number of symbols using a finite number of rules. Since several decades, scientists have been aware that this creativity is a feature shared by human languages and biopolymers, such as nucleic acids and proteins. Indeed, merely 20 aminoacids (letters or words) build millions of sequences (sentences) folded in thousands of different folds (syntactics) playing various functions in living organisms (semantics). Similarity of nucleic acids and proteins to sentences in natural languages is more than just a matter of notation as a string. Indeed, the nucleic acid or protein sequence actually forms a string of tokens (aminoacids) physically linked by peptide bonds. Similarily to the natural language sentences, protein sequences may be ambiguous (the same amino acid sequence folds into different structures depending on the environment), and often includes long-range dependencies and recursive structures. In a particular case we observed in proteins the phenomenon of double articulation, specific for the vast majority of human language: a very large set of possible molecular surfaces (sentences) obtained from a limited set of repeated sequence fragments (words), consisting of several amino acids subjected to positive selection (phonems).

Due to this similarity, methods based on the formal languages theory are specifically well suited to analysis of protein sequences. However linguistic methods are typically used in proteomics as a "black box". Moreover, popular methods cannot take into account dependencies between aminoacids distant in the sequence, which are very important for protein structure and function. In this project, we will work on applicating to the analysis of proteins a more advanced grammatical model: the probabilistic context-free grammar, which allows for direct representation of interactions between aminoacids distant in the sequence. This type of grammar is a very powerful tool: for example it is often utilized to define programming languagues. Our goal is to develop methods for deducing grammatical models of proteins. Three major applications of the model are: (1) searching databases to find similar proteins sequences; (2) analysis of the structure of the automatically-generated grammar to capture specific features and relationships - perhaps biologically relevant; (3) representing various scientific hypotheses in the form of grammars and compare which grammar best fits proteins of interest.

Last two application have been so far neglected by bioinformaticians. Meanwhile, our earlier studies indicated that the structures of automatically generated context-free grammars may contain information regarding biologically relevant features of proteins. Therefore, we plan to study systematically the correspondence between grammatical models and features of proteins. Each model is a reduction of reality to some of its aspect. To determine which of the hypotheses represented by probabilistic grammatical models better describe the problem, it may be sufficient to calculate and compare probabilities that the models generate objects observed in the reality. Sometimes, however, it is beneficial to use another method of hypothesis testing which consists in producing a very large number of objects that match the model, and then comparing if their characteristic features are similarily distributed to characteristic features of actually exisiting objects. In this case, it is necessary to define measurable characteristics. Due to the similarity between protein sequences and natural language sentences, it is possible to apply characteristics used in analysis of literary texts, but also the characteristics used to describe dynamic systems.

A significant progress seems to be achieveble also in the application of grammatical models for searching databases of sequences. Since recently, thanks to rapidly growing number of available protein sequences and to newly developed mathematical methods, it has become possible to determine with high accuracy aminoacids in the sequence which are correlated. So far the approach has led to improvements in the protein structure prediction. We are convinced that including information about correlated mutations to the process of grammar generation will result in a significant improvement in the speed and quality of the model learning, and therefore it will constitute a significant achievement in the field of bioinformatics.

As a practical effect of the project, we will make methods developed in this project available to the community in the form of a software package and a web service. In a longer horizon, development of protein linguistics seems to be indispensable for systematic design of peptides and proteins, or - even - for programming of entire molecular subsystems (e.g. immunological); in accordance with their nature.