Usually thought of as the canonical Watson-Crick double helix, DNA can in reality fold into many different structures, some of which have profound effects on DNA's biological role. This project focuses on one of these non-canonical DNA forms, termed G-quadruplexes (G4), whose structure is based on the formation of the so-called G-tetrad, a planar array of four guanine bases kept together by a cyclic arrangement of hydrogen bonds. By stacking on top of each other, G-tetrads make up the core of the G-quadruplex, with four guanine tracts being connected by three intervening loops of variable sequence and conformation. Sequences capable of forming G-quadruplexes are widespread across genomes (e.g., over 700,000 seq. in the human genome) and have been found to be significantly enriched in regulatory regions, including telomeres (up to 25 % of all formed G4), and gene promoters. Accordingly, there is accumulating evidence that G-quadruplexes are involved in regulation of gene expression and maintaining chromosomal stability. At the same time, due to their versatility and plasticity, engineered G-quadruplex segments have also attracted attention as convenient and potentially programmable building blocks in chemistry, material sciences and nanotechnology. Structural polymorphism of G-quadruplexes with many possible folded states (topologies) adopted depending on the sequence and environmental conditions (in particular, type and concentration of alkali metal cations) presents both opportunities and challenges to rational design. On the one hand, it allows for multiple design choices, but on the other, it requires a reliable way of predicting and controlling the structure of G-quadruplexes. Unfortunately, despite the extensive research efforts, a practical and general framework allowing for such predictions has yet to be established.

Accordingly, the goal of the current project is to **thoroughly understand the relation between the sequence of guanine-rich DNA strands and the structure of G-quadruplexes so as to allow for a design of DNA G-quadruplexes with a desired folded topology**. This will involve a systematic search through all possible folded G4 structures of a wide range of DNA sequences and analysis of structural and energetic principles underlying the relative stability of these structures. The project focuses on three primary research objectives (detailed below) that will be addressed using an integrative approach combining multiscale computer simulations and complementary experimental methods. To enable such a multidisciplinary approach the PI, whose research focuses on the application of computational methods to study biomolecular systems, has established collaboration with a group having the expertise in experimental studies of G-quadruplex structures. **1) Determination and validation of relative stabilities for an extensive set of guanine-rich DNA sequences capable of forming G-quadruplexes with two G-tetrads.** A structural simplicity of two-tetrad G-quadruplexes makes them an excellent model system for stability prediction and explanation by means of computer simulations. Therefore, we will first use atomistic and coarse-grained simulations to evaluate relative stabilities of all theoretically possible two-tetrad G-quadruplexes for a systematic set of G4-forming sequences. Next, these predictions will be validated using an array of experimental methods, including NMR, CD and UV spectroscopy. **2) Detailed analysis of structural and energetic factors underlying the relative stability of possible G-quadruplex folded structures.** To identify the molecular origin of the observed differences in stability, we will analyze large datasets obtained in 1) by means of statistical, machine learning and information-theoretic methods. The computed stabilities of the folded structures will be then explained in terms of energetic and entropic contributions, allowing for examination of the balance of forces underlying the stability of G4 DNA. **3) Extending our analysis to three-tetrad G-quadruplexes and testing the developed model as a tool for prediction and design of G-quadruplex structure.** To evaluate and, if necessary, extend the applicability of our approach to the three-tetrad G-quadruplexes, we will test two major hypotheses about the energetic relation between the three- and two-tetrad G-quadruplexes. Finally, we will assess the predictive power of our model by designing and experimentally validating a set of three-tetrad G-quadruplexes.

We believe that successful completion of these tasks will result in a novel framework for predicting dominant folded structures of G-quadruplexes based on their sequence. By providing the first fully structure-based mapping of G-quadruplex stability, this framework will also be capable of explaining the observed stabilities in terms of interactions within the DNA chain and between this chain and the solvent. Both predictive and explanatory insights offered by the developed model should greatly facilitate structural studies of DNA sequences containing G4-forming motifs. We expect that our results will be especially important for DNA nanotechnology, as sequence-based structure prediction would largely eliminate resource-intensive trial-and-error protocols usually employed when designing G4 structures with desired structural properties for use in chemistry, material science and biomedical engineering.