Bridging Molecular Modeling and Machine Learning: New Perspectives on Protein-Glycosaminoglycan Recognition

Krzysztof K. Bojarski

Department of Physical Chemistry, Gdańsk University of Technology, 80-233 Gdańsk, Poland

In recent years, glycosaminoglycans (GAGs) have gained increasing attention due to their structural diversity and critical biological functions. These negatively charged, sulfated, and linear carbohydrates are composed of repeating disaccharide units, typically consisting of an amino sugar and either glucuronic or iduronic acid (except for keratan sulfate). GAGs are primarily found in the extracellular matrix and lysosomes, where, with the exception of hyaluronic acid, they are covalently attached to proteoglycans. Through predominantly electrostatic interactions, they modulate the biological activity of various protein targets, including cathepsins, growth factors, chemokines, antithrombin, and bone morphogenetic proteins. GAGs play essential roles in cell signaling, adhesion, angiogenesis, anticoagulation, and apoptosis. Their dysregulation has been implicated in numerous pathological conditions, including cancer, autoimmune disorders, Alzheimer's and Parkinson's diseases, arthritis, and mucopolysaccharidoses. A deeper understanding of protein-GAG interactions is thus crucial for developing novel therapeutic strategies targeting these diseases.

In molecular studies of protein-GAG interactions, *in silico* methods often complement experimental approaches. However, designing a computational pipeline for molecular modeling of these interactions presents several challenges due to the unique properties of GAGs. These molecules are highly flexible, with their conformational variability arising from glycosidic linkages and ring puckering, making it difficult to define a preferred conformation. Additionally, GAGs can bind to multiple regions of a protein with similar free energy values, further complicating the identification of the most favorable binding mode. Despite their relatively simple carbohydrate backbone, GAGs exhibit significant structural complexity due to their sulfation patterns, which play a crucial role in protein recognition, structural properties, and biological activity. Moreover, protein-GAG complexes can display comparable stability even when GAGs bind within the same region but adopt different conformations. These factors collectively make it challenging to determine the native binding pose and conformation. Machine learning (ML) algorithms, which are increasingly employed in molecular modeling, also face limitations in this context due to the scarcity of high-quality experimental structural and thermodynamic data. As a result, developing accurate scoring functions remains a considerable challenge. Nevertheless, despite these obstacles, *in silico* methods have been successfully applied numerous times to characterize the complex nature of protein-GAG interactions.

This study provides an overview of recent advancements in molecular modeling of protein-GAG systems. First, the ff14SB/GLYCAM06 (AMBER) and CHARMM36m (CHARMM) force fields are compared in terms of their ability to describe the conformational and energetic properties of protein-GAG interactions. While both force fields yield comparable energy profiles for GAG unbinding, they differ in their description of conformational properties, such as protein root mean square fluctuation (RMSF) and glycosidic linkage flexibility. Second, the impact of five explicit solvent models on the structural properties of heparin is assessed. Results suggest that in CHARMM, the choice of solvent model may have a lesser impact on heparin structure compared to AMBER. Finally, the Repulsive Scaling Replica Exchange Molecular Dynamics (RS-REMD) method is implemented in CHARMM as an advanced molecular docking approach and tested on seven protein-carbohydrate systems. The predictive power of MM-GBSA-based scoring is evaluated, and a deep learning-based scoring function is introduced, significantly improving the selection of the most accurate binding structures.