# Amino Acid Composition and Folding of the Monosaccharide Binding Sites of Enzymic Proteins

Khuri S., Nassif H., Al-Ali H., Khalaf K., Khachfe H., and Keyrouz W.

## Abstract

Hexoses are single-unit sugars which play an essential role in many chemical pathways within the cell. The sugar molecules bind to protein molecules at specific binding sites to realize these roles. Identifying these binding sites remains an open research problem.

This work uses Support Vector Machines (SVM), a well-known pattern recognition technique, to find the binding sites of *hexoses*. The technique is applied in two phases. In the *learning* or *training* phase, SVM processes descriptions of currently known binding sites of hexoses in terms of their geometric and chemical properties. It extracts from these descriptions a pattern that the binding sites match. In the *discovery* phase, SVM explores existing descriptions of other proteins to locate sites that match the pattern identified in the learning phase.

## Background & Motivation

*Problem motivation*—Many proteins whose functions are unknown bind to hexose sugars. Identifying the sugar binding sites of these proteins contributes to their functional description. However, the identification task is further complicated as these proteins belong to different families that do not share significant sequence and structural similarities.

*Approach motivation*—It is not feasible to use numerical simulations to identify binding sites as these simulations are computationally too expensive. As such, one should resort to simulations only on promising binding sites.

We want to identify candidate sites by examining the chemical and geometric properties at a binding site as these properties determine the site's functionality. Furthermore, we want to use a pattern recognition technique that act on descriptions of the properties to provide a first-level filter of potential sites. These binding sites will then be confirmed by more detailed analytical and numeric techniques and, later on, by experimental techniques.

## Methodology

This work focuses on the Galactose, Glucose, and Mannose binding sites. It treats a binding site as a chemical and geometric environment where hexose docking takes place. This environment consists of the sphere centered at the hexose's pyranose ring and the enclosing protein binding residues.

The system we are developing represents the spherical binding environment as a feature vector. These features include residue type, polarity, secondary structure, dihedral angle and solvent accessibility (using STRIDE), and evolutionary data (sequence conservation, HMMSTR's amino acid profile and context descriptor, *etc.*).

The system extracts the feature vectors from Protein Data Bank files and feeds them to an SVM-based pattern recognition component. During SVM's learning phase, the system will process *positive* examples—feature vectors that characterize known hexose binding sites. It will also process *negative* examples—non-sites and sites that do not bind hexoses. The outcome of this phase is a *pattern* that the system then uses to match against candidate feature vectors in the discovery phase.

A sensitivity analysis step will later on explore the radius of the binding sphere and the relevance of certain features.

## Progress Report

We have so far extracted a non-redundant set of hexose-binding proteins by filtering the outcome of queries to the PDB and through an extensive literature search.

We have also identified an extensive set of features to describe these binding environments through a literature review.