AMERICAN UNIVERSITY OF BEIRUT

A PATTERN RECOGNITION BASED MODEL FOR CHARACTERIZING AND PREDICTING GLUCOSE-BINDING SITES

HOUSSAM GEORGES NASSIF

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Computer Science
of the Faculty of Arts and Sciences
at the American University of Beirut.

Beirut, Lebanon.
June 2006

AMERICAN UNIVERSITY OF BEIRUT

A PATTERN RECOGNITION BASED MODEL FOR CHARACTERIZING AND PREDICTING GLUCOSE-BINDING SITES

HOUSSAM GEORGES NASSIF

Approved by:

Dr. Walid Keyrouz, Assistant Professor — Advisor
Computer Science

Dr. Chiraz Benabdelkader, Assistant Professor — Member of Committee
Computer Science

Dr. Rabih Talhouk, Professor — Member of Committee
Biology

Date of thesis presentation: June 7, 2006

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

I, Houssam Georges Nassif,

☐ authorize the American University of Beirut to supply copies of my thesis to libraries or individuals upon request.

☐ do not authorize the American University of Beirut to supply copies of my thesis to libraries or individuals for a period of two years starting with the date of the thesis defense.

_____

Signature

_____

Date

ACKNOWLEDGMENTS

I want to express my gratitude to Dr. Sawsan Khuri and Mr. Hassan Al-Ali. Dr. Khuri introduced me to Bioinformatics and acted as a co-supervisor of this thesis despite her being physically in Miami, FL. Mr. Hassan Al-Ali provided me with the binding sites data and gave me valuable biochemical feedback. Without them this work would never have come into existence.

I would like to thank my supervisor, Dr. Walid Keyrouz, for his dedicated guidance, help and endorsement. I am grateful to Dr. Khachfe for his many discussions. I am also thankful to my examiners, Dr. Chiraz Benabdelkader and Dr. Rabih Talhouk for managing to read and comment the whole document. I should also mention Dr. Jihad Boulos for his guidance through the early months of chaos and confusion.

I warmly thank Dr. Giri Narasimhan and Mr. Leonardo Bobadilla of Florida International University. Dr. Narasimhan took the time to review and comment on this document. Mr. Bobadilla shared with me his SVM feature selection technique.

I am appreciative to all my colleagues and friends who constantly motivated and supported me throughout this thesis. I especially mention Deema for her caring incentives, Dania for her concern, Mark for his presence, and Shant and Rami for their help.

I am forever indebted to Wael and Rawane for their understanding and encouragement, to K.Roll for her constant care, joyful company and tender affection, and to my parents, Georges and Najat, for my upbringing and their endless patience and love. To them I dedicate this thesis.

Beirut, Lebanon — Houssam Nassif
June 7, 2006

AN ABSTRACT OF THE THESIS OF

Houssam Georges Nassif for Master of Science
Major: Computer Science

Title: A Pattern Recognition Based Model for Characterizing and Predicting Glucose-Binding Sites

Glucoses are single-unit sugars that play essential roles in many biochemical pathways within the cell. To realize these roles, glucose binds to protein molecules at specific binding sites. The biochemical process by which the glucose or any other ligand binds to the binding site is termed *docking*. Binding sites are specific to their respective ligands. Identifying the ligand docking at a certain binding site remains an open research problem.

This thesis attempts to predict and model glucose-binding sites. It uses Support Vector Machines (SVM) and k-Nearest-Neighbors (kNN), two well-known classification techniques, to find glucose-binding sites. The thesis represents the binding site as a vector of geometric and biochemical features. The prediction techniques operate in two phases. In the learning phase, the classifier processes descriptions of known binding and non-binding glucose sites to build a predictive model. The classifier then uses this model during the testing phase to predict if unknown sites are glucose binding sites.

The thesis refines the predictive model by first applying dimensionality reduction techniques to identify the characteristic features of the glucose-binding sites. It then retrains the classifiers using the reduced feature vectors.

CONTENTS

ACKNOWLEDGMENTS v

ABSTRACT vi

1 INTRODUCTION 1
1.1 Problem Overview 1
1.1.1 Problem Description 1
1.1.2 Motivation 1
1.1.3 Description of the Approach 2
1.1.3.1 Binding Site Representation 2
1.1.3.2 Feature Extraction 3
1.1.3.3 Learning Phase 3
1.1.3.4 Sensitivity Analysis 4
1.1.3.5 Testing Phase 4
1.2 Organization 4

2 BIOCHEMICAL BACKGROUND 5
2.1 Carbohydrates 5
2.1.1 Monosaccharides 5
2.1.1.1 Chemical Classification 5
2.1.1.2 Stereochemistry 7
2.1.2 Aldohexoses 8
2.1.2.1 Configuration 8
2.1.2.2 Cyclization 9
2.2 Proteins 10
2.2.1 Amino Acids 10
2.2.2 Protein Structure 12
2.3 Biochemical Interactions 13
2.3.1 Chemical Bonds 13
2.3.1.1 Covalent Bonds and Polarity 13
2.3.1.2 Hydrogen Bonds 14
2.3.2 Intermolecular and Intramolecular Forces and Interactions . . 15
2.3.2.1 Van der Waals Forces 15
2.3.2.2 Hydrophobic Interactions 16
2.3.3 Protein-Hexose Interaction 16
2.4 Previous Work 17
2.4.1 Support Vector Machines Applied to Biological Problems . . . 17
2.4.2 Identification of Hexose Binding Sites 17

3 PATTERN RECOGNITION . . . 22
3.1 Statistical Pattern Recognition . . . 22
3.1.1 Basic Classification Concepts . . . 22
3.1.1.1 Pattern Classification Approaches . . . 22
3.1.1.2 Bayesian Learning . . . 24
3.1.1.3 Density Estimation . . . 25
3.1.2 k-Nearest-Neighbor . . . 26
3.1.2.1 Simple kNN . . . 26
3.1.2.2 Distance-Weighted kNN . . . 28
3.1.3 Support Vector Machines . . . 29
3.1.3.1 Linear Discriminants . . . 30
3.1.3.2 Kernel Mapping . . . 33
3.1.3.3 Support Vector Machines Implementation . . . 34
3.2 Feature Representation . . . 37
3.2.1 Standardizing Data . . . 37
3.2.2 Curse of Dimensionality . . . 38
3.2.3 Random Forests . . . 38
3.2.3.1 RF Implementation Overview . . . 39
3.2.3.2 Feature Importance Computation . . . 39

4 PROBLEM REPRESENTATION AND IMPLEMENTATION . . . 41
4.1 Data . . . 41
4.1.1 Database Availability . . . 41
4.1.2 Data mining step . . . 42
4.1.3 PDB Filtering . . . 43
4.1.4 Negative Entries . . . 46
4.2 Solution Approach . . . 48
4.2.1 Feature Extraction . . . 48
4.2.1.1 Atomic Features . . . 49
4.2.1.2 Residue Features . . . 51
4.2.2 Binding Site Representation . . . 53
4.2.2.1 Concentric Layers Sphere . . . 53
4.2.2.2 Sphere Center . . . 54
4.2.3 SVM and kNN Classifiers . . . 55
4.2.4 Error Estimation . . . 56
4.2.5 Modeling Glucose-Binding Sites . . . 57

5 EXPERIMENTAL RESULTS 58
5.1 Introduction 58
5.2 Learning Phase 59
5.2.1 Primary Findings 59
5.2.1.1 Atomic Properties Comparison 60
5.2.1.2 Comparison of Classifiers 61
5.2.2 Water and Ions Inclusion 61
5.2.3 Residue Schemes Analysis 63
5.2.3.1 Residue Features Comparison 63
5.2.3.2 Atomic and Residue Properties Combination 64
5.2.4 Atomic Properties Analysis 67
5.2.4.1 The Hydrogen Bond Atomic Property 67
5.2.4.2 The Hydrophobicity Atomic Property 68
5.2.4.3 The Charge Atomic Property 70
5.3 Feature Selection 70
5.3.1 Atomic Properties 70
5.3.1.1 Charge Feature Selection 71
5.3.1.2 Hydrogen Bond Feature Selection 72
5.3.1.3 Hydrophobicity Feature Selection 73
5.3.2 Residue Properties 75
5.3.2.1 Simplified Schemes Feature Selection 75
5.3.2.2 Detailed Schemes Feature Selection 76
5.3.3 Combining Atomic and Residue Properties 77
5.3.4 Feature Selection Findings 80
5.4 Testing Phase 81

6 CONCLUSION 83
6.1 Summary 83
6.2 Future Work 84

REFERENCES 85

LIST OF FIGURES

2.1 Structures of carbonyl and hydroxyl groups 6
2.2 Structures of aldehyde and ketone groups 6
2.3 Structures of glyceraldehyde aldose and dihydroxyacetone ketose . . . 7
2.4 Fisher projections of glyceraldehyde 8
2.5 Structures of glucose, mannose and galactose 9
2.6 Glucose conformations 10
2.7 General structural formula of an amino acid. 10
2.8 A peptide bond linking two amino acids together. 11
2.9 Polar molecules 14
2.10 Hydrogen bond 14

3.1 A separating hyperplane segregating two classes in a $\mathbb{R}^2$ space 31
3.2 Projection of input $\boldsymbol{x}$ onto hyperplane R 32
3.3 A separating hyperplane with its support vectors highlighted. 37

4.1 The classifier algorithm outline. 49
4.2 Glucose bound to a hydrolase, PDB entry 1I8A 53

5.1 Simple *vs.* weighted kNN plot. 62
5.2 Comparison of detailed1 and detailed2 schemes. 65
5.3 Comparison of simplified2 and simplified3 schemes. 66
5.4 Analysis of the hydrophobicity property. 69
5.5 Importance of charge features according to RF 72
5.6 Importance of hydrogen bond features according to RF 73
5.7 Importance of hydrophobicity features according to RF 74
5.8 Importance of "simplified3" features according to RF 76
5.9 The 5 highest "detailed1" features importance measure as returned by RF 77
5.10 The 7 highest "charge + hbond + hydro + detailed" features importance measure as returned by RF 79

LIST OF TABLES

2.1 Standard amino acids names, three- and one-letter abbreviations . . . 11
2.2 Differences between Group I and Group II carbohydrate-binding proteins. 18
2.3 Natural logarithm of the sugar interface propensity values for the standard amino acids 20

3.1 Typical kNN kernel functions 29

4.1 The various sections of a PDB file 42
4.2 Inventory of the D-Glucose-binding proteins. 44
4.3 Inventory of the positive training set binding sites. 45
4.4 Inventory of the positive testing set binding sites. 45
4.5 Inventory of the set 1 negative training sites 46
4.6 Inventory of the set 2 negative training sites. 47
4.7 Inventory of the set 3 negative training sites 47
4.8 Inventory of the set 4 negative testing sites. 48
4.9 Atomic features. 50
4.10 The two detailed residue subgrouping schemes, "detailed1" and "detailed2". 51
4.11 The different simplified residue subgrouping schemes. 52

5.1 Misclassification rates using the "detailed1" residue scheme 60
5.2 Testing the importance of water and ions to glucose specificity. 62
5.3 Comparison of the different residue schemes. 63
5.4 Comparison of atom and residue properties. 64
5.5 Classifier training using an exclusively non-binding sites negative set. 68
5.6 Comparison of classifiers' performance on atomic data with and without RF feature selection. 71
5.7 Comparison of classifiers' performance on residue data with and without RF feature selection. 75
5.8 Comparison of classifiers' performance on combined atomic and residue data with and without RF feature selection. 78
5.9 The feature combination achieving the minimal error for both kNN and SVM classifiers. 79
5.10 Testing phase results. 81

LIST OF ALGORITHMS

4.1 Compute layer feature vector 54

4.2 Compute sphere feature vector 54

CHAPTER 1

INTRODUCTION

1.1 Problem Overview

1.1.1 Problem Description

Hexoses are 6-carbon sugar molecules that play a key role in many different biochemical pathways, including cellular energy release, signaling, carbohydrate genesis and gene expression regulation. Different types of proteins bind the hexoses, resulting in various hexose cellular functions. These proteins belong to different protein families that have little, if any, overall sequence or structural similarity. Hexose-binding proteins are characterized by a binding site, where protein-hexose interaction occurs. The hexose binding process is termed *docking*. The hexose, or any other molecule docking to a binding site, is called a *ligand*.

The 3-D structure and functionality of a protein directly result from its amino acid sequence and its folding geometry. The few amino acids present at the binding site, rather than the overall protein fold, determine the binding site's topology and biochemical properties and hence the ligand type and the protein's functionality [39]. As a result, the prediction of the hexose ligand type based on sequence alignment and structural homology is inaccurate. Therefore, characterizing these binding site features and incorporating them in a prediction/modeling algorithm should lead to tangible results.

1.1.2 Motivation

Many proteins with unknown or not fully characterized functions are thought to bind to hexose sugars. Identifying the sugar binding to these proteins contributes greatly to their functional description. However, the identification task is further

complicated as these proteins belong to different families that do not share significant sequence or structural similarities.

It is not feasible to use numerical simulation to identify binding hexoses at a certain site as these simulations are computationally very expensive: one hexose docking simulation using Insight-II [2] may take more than one hour. As such, one should resort to simulations only on known ligands.

We want to identify candidate sites by examining the chemical and geometric properties at a binding site as these properties determine the functionality of such a site. Furthermore, we want to use a machine learning technique that acts 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 by experimental methods in extensions to this work.

1.1.3 Description of the Approach

This work focuses on glucose binding sites since glucose is one of the most biochemically active sugars in animal biology. It aims to build a statistical classifier that classifies glucose-binding sites. The approach consists of the following five steps: we begin by the problem representation including binding site representation and feature extraction. We then conduct the learning phase experiments and a sensitivity analysis step for feature selection. Finally we validate our results during the testing phase.

1.1.3.1 Binding Site Representation

The approach 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 (the hexose core ring formed from five carbons and one oxygen atoms) and includes the protein binding amino acids.

1.1.3.2 Feature Extraction

This work represents the spherical binding environment as a feature vector. These features include residue type and properties, polarity, hydrophobicity and hydrogen bonding.

The system extracts the feature vectors from Protein Data Bank files and feeds them to the classifier. During the learning phase, the system will process positive examples—feature vectors that characterize known glucose binding sites. It will also process negative examples—non-binding sites and binding-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.

1.1.3.3 Learning Phase

Statistical classification techniques are easy-to-use methods capable of discovering the hidden rules behind a classification outcome. The thesis uses two classification techniques, Support Vector Machines and k-Nearest-Neighbors.

Support Vector Machines are classifier algorithms used for predictive modeling. The structure of these algorithms is very flexible and can be applied to virtually any classification problem that uses predefined classes. The SVM algorithm has been found to outperform, in many cases, other classification methods, such as neural networks and decision trees, in terms of both computational efficiency and accuracy [37].

We compare SVM results to k-Nearest-Neighbors, since the latter remains accurate to small training samples. kNN is a non-parametric approach to classification and is particularly suitable for numeric feature vectors [48]. It is a simple algorithm whose resulting classifying rules are easy to analyze.

1.1.3.4 Sensitivity Analysis

The binding site representation and the feature extraction steps can be conceived in numerous ways. It is unknown, a priori, which representation or features correctly model the glucose-binding sites. A sensitivity analysis step explores the representation of the binding sphere and the relevance of certain features.

1.1.3.5 Testing Phase

The learning phase, coupled with the sensitivity analysis step, builds a glucose-binding site classifier. To evaluate the classifier's accuracy and predictive powers, we run it on a previously-unused set of data.

1.2 Organization

The rest of this document is organized as follows: Chapter 2 describes the biochemical properties of proteins, glucose and their interactions. It includes a review of previous work. Chapter 3 details the classifying techniques used. It covers SVM, kNN and feature representation. Chapter 4 specifies data gathering and preprocessing, together with the solution approach. The results are shown, analyzed and discussed in Chapter 5. Finally, Chapter 6 draws conclusions and sets perspectives for future work.

CHAPTER 2

BIOCHEMICAL BACKGROUND

Hexose-binding sites involve the interaction of two different biochemical families: hexoses, which are carbohydrates, and proteins. This chapter introduces the relevant biochemical properties of these classes and their interaction. The final section lists some biological problems solved using SVM, together with previous computational attempts to identify hexose binding sites.

2.1 Carbohydrates

Carbohydrates are one of the four major classes of organic molecules, together with lipids, proteins and nucleic acids. Carbohydrates, commonly referred to as *sugars*, are the most abundant biological molecules on earth [21]. This major class includes, among others, the walls of plant cells, starch, sugars, cellulose, fibers and wood. Carbohydrates provide most of the energy for both the human body and animal life.

2.1.1 Monosaccharides

2.1.1.1 Chemical Classification

Large carbohydrates are *polymers*, large molecules formed by the repetitive combination of smaller molecules, termed *monomers*. The *hydrolysis* chemical process decomposes a polymer into its constituent monomers. A *monosaccharide* is the smallest repetitive unit, or monomer, of a carbohydrate polymer. It is the simplest carbohydrate, which does not yield further carbohydrates upon hydrolysis. The monosaccharide formula is:

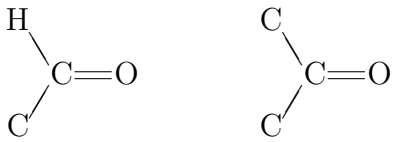
$$(CH_2O)_n \qquad \text{for } 3 \leq n \leq 9 \tag{2.1}$$

where C refers to carbon, H to hydrogen and O to oxygen. Few monosaccharide monomers linearly linked together form an *oligomer*. A polymer is a large oligomer, formed by the combination of many monosaccharide monomers. It is worth to note that the name carbohydrate is derived from the 2.1 chemical formula, or "hydrate (water) of carbon", H_2O being the chemical formula for water.

Chemically, a monosaccharide contains one *carbonyl group* (CO) and each one of the remaining $n-1$ carbon atoms is bonded to a *hydroxyl group* (OH) (see Figure 2.1). The carbon in the carbonyl group is linked to the oxygen by a double bond. An *aldehyde* is a compound containing a carbonyl group linked to a carbon and a hydrogen atom; a *ketone* is a compound containing a carbonyl group linked to two carbon atoms (see Figure 2.2).

(a) (b)

Figure 2.1: (a) Carbonyl group (b) Hydroxyl group.



(a) (b)

Figure 2.2: (a) Aldehyde; (b) Ketone.

It follows that monosaccharides are either *polyhydroxy aldehydes* or *polyhydroxy ketones*, according to their carbonyl group type. They are named *aldoses* and *ketoses* respectively [9] (see Figure 2.3). Aldoses and ketoses are further classified

according to the number of carbon atoms they contain. Monosaccharides containing 3 carbons are named *trioses*, 4 carbons compounds are *tetroses*, and so on. An *aldotriose* is an aldehydic triose where an aldehyde is the core of the triose; a *ketotriose* is a ketonic triose where a ketose is the core of the triose. The same goes for aldotetrose, aldopentose, ketotetrose, ketopentose.... Knowing this nomenclature, Figure 2.3(a) is an aldotriose, while Figure 2.3(b) is a ketotriose.

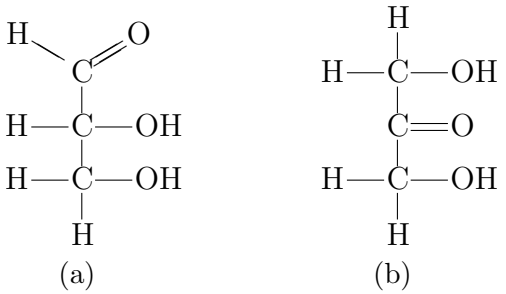


Figure 2.3: (a) Glyceraldehyde Aldose; (b) Dihydroxyacetone Ketose.

It is worth noting that both molecules in Figure 2.3 have the same empirical formula $(CH_2O)_n$ (see Equation 2.1), but have different constitutions. Such similar-formula molecules are called *isomers*. All *n*-oses (triose, tetrose...) are isomers.

2.1.1.2 Stereochemistry

Stereochemistry (or chemistry in three dimensions) adds an extra complexity level to monosaccharide classification. Molecules may have the same constitution, but differ in the spatial arrangement of their atoms [9]. This is the case when two objects are mirror images of each others. A left hand and a right hand are mirror images, they have the same constitution, but differ in their spacial configuration. The same principle applies to molecules. Taking glyceraldehyde (Figure 2.3(a)) as an example, it has two mirror-image forms, L and D forms, as shown in Figures 2.4(a) and 2.4(b). Figures 2.4(c) and 2.4(d) illustrate the different glyceraldehyde conformations in 3-D structure: the ring is the central carbon atom, the horizontal

bonds protrude out of the document and the vertical bonds penetrate through the document.

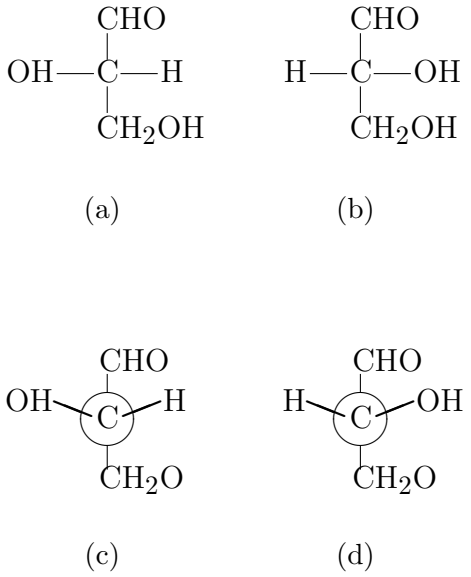


Figure 2.4: Planar (Fischer) projection of L-glyceraldehyde (a) and D-glyceraldehyde (b); together with their respective 3-D conformation (c) and (d).

L-Glyceraldehyde and D-Glyceraldehyde are called *stereoisomers*, isomers having the same constitution but with different 3-D atom arrangements. Although stereoisomers might appear similar, they are different compounds with different biochemical properties. Compounds exhibiting this stereo characteristic are said to be *chiral* [21]. Most monosaccharides are chiral compounds, and thus exist in both L- and D-forms.

2.1.2 Aldohexoses

2.1.2.1 Configuration

Based on the previous discussion, aldohexoses are 6-carbon monosaccharide aldehydes. The most biochemically active sugars in animal biology are three aldohexoses, namely D-Galactose, D-Glucose and D-Mannose, shown in Figure 2.5.

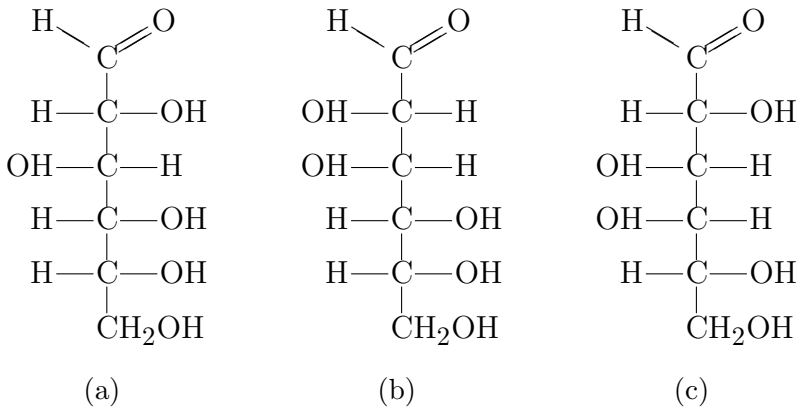


Figure 2.5: (a) D-Glucose; (b) D-Mannose; (c) D-Galactose

2.1.2.2 Cyclization

Carbonyl and hydroxyl groups (see Figure 2.1) are two chemical groups that can react together. An aldohexose incorporates both groups, and their reaction folds the molecule on itself as shown in Figure 2.6. This intramolecular cyclization reaction forms a *pyranose ring* from five carbons and one oxygen atoms [9]. The cyclized hexose can adopt either of two configurations, α or β, according to whether the first carbon hydroxyl group —OH* is located below or above the pyranose ring (see Figure 2.6). The asterisk highlights the difference in the first carbon hydroxyl's position. The pyranose ring carbon atoms are not explicitly shown. α- and β-pyranoses are setereoisomers.

D-Glucose can readily shift from one conformation to another, as indicated in Figure 2.6 by the double arrows $\rightleftharpoons$. In physiological solutions, i.e. in the living organisms' cells, fluids and tissues, aldohexoses exist almost exclusively in the pyranose form. For example, at 31°C, D-Glucose exists in an equilibrium mixture of 64% β-D-Glucopyranose and 36% α-D-Glucopyranose, with only a tiny fraction in the open-chain form [21].

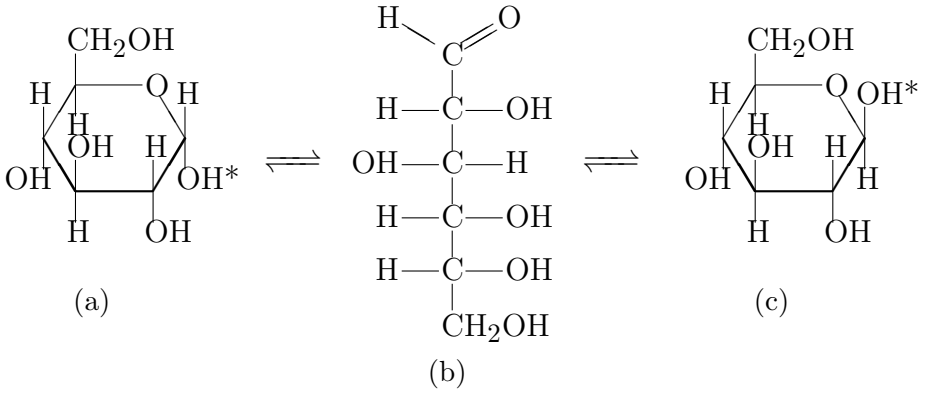


Figure 2.6: (a)α-D-Glucopyranose, (b) D-Glucose, (c) β-D-Glucopyranose.

2.2 Proteins

As previously stated, proteins are one of the major classes of organic molecules. Each cell type has a characteristic set of proteins that give the cell type its functional properties. Proteins participate in virtually every biological process. Identifying their composition, structure, chemical properties and 3-D shapes is key to understanding many biochemical functions and processes.

2.2.1 Amino Acids

Proteins are formed by one or more *polypeptide chains*, which are polymerized linear chains of *amino acids*. The amino acid structure consists of a central carbon atom C, which bonds to an amino group (NH_2), a carboxyl group (COOH), a hydrogen atom H, and a side chain R, as depicted in Figure 2.7.

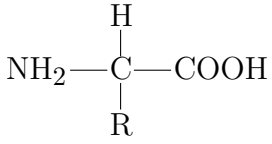


Figure 2.7: General structural formula of an amino acid.

The side chain R varies among amino acids, giving each amino acid its distinctive chemical properties. Twenty different amino acid types combine to form

all proteins in all living organisms. These 20 biological amino acids are called the "standard" amino acids [21]. Table 2.1 lists their names, three-letter and one-letter abbreviations. Aspartate and glutamate are also called aspartic acid and glutamic acid, respectively. Amino acids in a protein are also called *residues*.

Table 2.1: Standard amino acids names, three- and one-letter abbreviations. They are sorted according to biological convention based on the size and properties of the side chain.

Name	3-L	1-L	Name	3-L	1-L
Glycine	Gly	G	Cysteine	Cys	C
Alanine	Ala	A	Serine	Ser	S
Valine	Val	V	Threonine	Thr	T
Leucine	Leu	L	Aspartate	Asp	D
Isoleucine	Ile	I	Glutamate	Glu	E
Proline	Pro	P	Histidine	His	H
Phenylalanine	Phe	F	Lysine	Lys	K
Tyrosine	Tyr	Y	Arginine	Arg	R
Tryptophan	Trp	W	Asparagine	Asn	N
Methionine	Met	M	Glutamine	Gln	Q

Amino acids are subdivided into subgroups, based on the structural and chemical properties of their side chains [41]: aliphatic, aromatic, sulfur containing, alcohols, acidic, basic, amides, polar, nonpolar....

Amino acids are linked together by *peptide bonds* to form proteins. A peptide bond links the amino group (NH_2) of a residue with the carboxyl group (COOH) of another (refer to Figure 2.8). A protein is a chain of amino acids linked by peptide bonds. The protein *backbone chain* is the main chain formed by the peptide bonds, composed of a repetition of NH—C—CO—NH linked atoms.

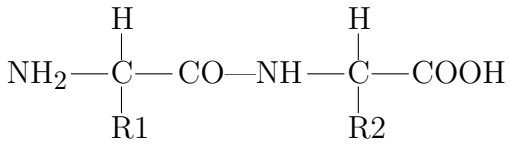


Figure 2.8: A peptide bond linking two amino acids together.

2.2.2 Protein Structure

Different proteins have different functions within the cell. The 3-D structure of the protein determines its function; the sequence of amino acids in a protein dictates the protein's 3-D shape [36]. This is an important concept in biology, where protein shape and function is determined by the sequence of its amino acid residues.

The importance of the amino acid sequence is highlighted in the structural organization of a protein, where each level directly influences the next one [36]:

1. Primary Structure: the linear sequence of amino acids within the polypeptide chain.

2. Secondary Structure: due to their different chemical properties, amino acids interact together and force the chain to fold on itself. Secondary structure formations include α-helix and β-strands.

3. Tertiary Structure: it is the 3-D overall structure resulting from the folding of different secondary structure formations.

4. Quaternary Structure: some proteins are composed of multiple polypeptide chains. Quaternary structure is the final packaging of the different polypeptide chains to form the protein.

All four structural levels of a protein directly depend on the amino acid sequence and properties. Amino acids sharing similar spatio-chemical properties can be easily interchanged in a protein. By contrast, replacing a residue with a spatio-chemically distant one results in a distortion of the protein structure and function.

2.3 Biochemical Interactions

Atoms and molecules interact together in different ways. Some are bonded together, others are subject to attracting or repelling forces. This section introduces the biochemical bonds and forces in protein-hexose interactions.

2.3.1 Chemical Bonds

The atoms of a compound are held together by *chemical bonds*. Chemical bonds are of many types and of different strengths. Two bond types are of special relevance to us: covalent bonds and hydrogen bonds.

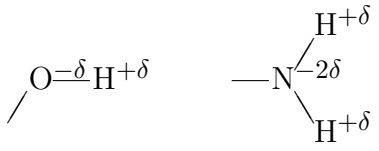
2.3.1.1 Covalent Bonds and Polarity

A covalent bond is a strong attraction force between atoms involving the sharing of electrons. The atoms of a molecule are held together by covalent bonds, like all the bonds we have seen so far. Atoms or compounds linked by covalent bonds form one molecule. As an example, consider the general structural formula of an amino acid shown in Figure 2.7. In this structure, the side chain R is covalently linked to the central carbon C, as well as the amino group (NH_2), the carboxyl group (COOH), and the hydrogen atom H. The whole amino acid is considered one molecule.

Atoms share electrons in a covalent bond. *Electronegativity* is a measure of an atom's attraction for electrons in chemical bonds [38]. When both atoms involved in a covalent bond have similar electronegativities, the electrons are shared equally and the bond is *nonpolar*. However, if the electronegativities are different, electrons are pulled closer to the more electronegative atom, resulting in a *polar covalent bond*. A polar covalent bond will have a partial negative charge on one side, and a partial positive charge on the other.

Oxygen O and nitrogen N are highly electronegative while hydrogen H is not. O—H and N—H bonds are quite abundant in amino acids and hexoses (see

Figures 2.5 and 2.7) and are polar covalent bonds as shown in Figure 2.9. A polar molecule contains one or more polar covalent bonds such that it becomes dipolar, with one pole bearing a partial positive charge and the other a partial negative charge [38].



(a) (b)

Figure 2.9: Polar molecules: (a) Hydroxyl group; (b) Amino group. δ indicates a partial charge.

2.3.1.2 Hydrogen Bonds

The hydrogen bond is the weakest of the bonds, and the most relevant in protein interactions. The hydrogen bond is an electrostatic attraction between an atom with a partial negative charge and a hydrogen atom that is covalently bonded to oxygen or nitrogen, thus bearing a partial positive charge [38] (see Figure 2.10).

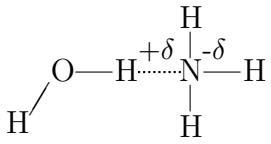


Figure 2.10: A hydrogen bond is formed between the water's partially positive hydrogen atom H and the ammonia's partially negative nitrogen atom N.

Hydrogen bonds are weak, easily formed and broken and have a specific length (0.27–0.30 nm) and orientation. Hydrogen bonds play an essential role within and between biological molecules. At the intramolecular level, hydrogen bonds determine the 3-D structure of the protein by establishing the secondary structure

formations (refer to Section 2.2.2). At the intermolecular level, protein-binding hexoses attach to the protein using hydrogen bonds.

Hexose-hydroxyl groups, being polar and highly exposed (protruding out of the molecule), are always involved in hydrogen bonding. Ideally, a hydroxyl group can donate one and accept two hydrogen bonds [33]. From the amino-acid side, most hydrogen bonds involve planar polar residues, namely asparagine (Asn), aspartate (Asp), glutamine (Gln), glutamate (Glu) and arginine (Arg) [33].

2.3.2 Intermolecular and Intramolecular Forces and Interactions

In addition to bonding, molecules may interact through other forces, such as the van der Waals and hydrophobic forces. Like hydrogen bonds, these are weak forces, but are collectively strong when present in large numbers. Both forces play an important role in protein structure and in protein-hexose binding.

2.3.2.1 Van der Waals Forces

Polar molecules (refer to Section 2.3.1.1) induce, in addition to hydrogen bonds, other weak electrostatic forces. The dipoles of polarized molecules interact together; positive poles attract negative ones while repelling each other. Polar molecules induce transient polarization in neighboring nonpolar molecules. Even in the absence of polar molecules, nonpolar molecules can be momentarily polarized due to random movement of their electrons [21]. These weak electrostatic attraction and repulsion forces, called *van der Waals forces*, operate over very short distances.

Van der Waals forces in protein-sugar binding sites mainly result from the apolar hexose pyranose ring stacking against an aromatic residue [33, 39, 40]. Stacking aromatic residues are tryptophan (Trp), tyrosine (Tyr), phenylalanine (Phe), in addition to histidine (His).

2.3.2.2 Hydrophobic Interactions

Hydrophobic, or "water-hating" interactions (as opposed to *hydrophilic*, or "water-loving") occur between nonpolar molecules in an aqueous medium [38]. These hydrophobic molecules are insoluble in water and tend to cluster together. In protein folding, hydrophobic residues aggregate away from water in the inner core of the protein, while the exposed residues are hydrophilic.

On the other hand, polar molecules tend to dissolve readily in water. Likewise, hydrophilic molecules tend to cluster together. It is important to note that the large number of polar hydroxyl groups, plus the carbonyl group, give a monosaccharide hydrophilic properties (refer to Figures 2.3 and 2.5).

2.3.3 Protein-Hexose Interaction

Binding proteins are characterized by a *binding site*: a cleft or groove in their structure where binding occurs. The molecule that binds to the protein is called a *ligand*. The binding process, known as *docking*, occurs in a key-lock fashion, where the binding site is tailored to accept and bind to this specific ligand.

Proteins that bind hexoses belong to diverse protein families that lack significant sequence or structural similarity. However, proteins within the same family have a structural similarity in the sugar binding site [26, 39]. Despite the dissimilarity in the binding site architecture between protein families, these proteins show high specificity to their hexose ligands. Assuming that common recognition principles exist for common substrate recognition [39], binding sites have unique distinguishing biochemical and spatial features identifying them.

The 3-D structure and biochemical properties of a protein directly result from its amino acid sequence and spatio-chemical properties, as seen in Section 2.2.2. Jaramillo *et al.* [25] even suggest that the amino-acid sequences at binding-sites might have been selected, at least in part, for mediating intermolecular interactions, probably at the expense of protein stability. Since the main function of a binding

protein is to bind ligands at its binding site, then the protein function is mainly determined by the few residues found at the binding site [40]. Thus, we need a computational method to properly select relevant features and predict hexose ligand types docking within a protein.

Hexoses bind tightly to proteins. Lectin, for example,is a hexose-binding protein family. The free energy of a lectin-carbohydrate binding is 1.7 times larger than that of a protein-protein binding [18]. Protein-hexose interactions have a dual nature: a polar nature that results in the formation of extensive hydrogen-bonding networks, and a hydrophobic nature that allows the stacking of the hydrophobic side of the pyranose ring against aromatic residues. A bound monosaccharide can be in contact with up to 19 residues [42]!

2.4 Previous Work

2.4.1 Support Vector Machines Applied to Biological Problems

SVMs has been used in many biological problems, including functional site recognition, structural prediction, protein remote homology detection, microarray expression data analysis and protein fold recognition. Noble [31] reviews the different SVM applications in bioinformatics, and includes a detailed description of the approach used in multiple biological domains.

Bobadilla, Nino, and Narasimhan use SVM to predict and characterize metal-binding sites [6]. They generate site descriptions based on the geometric and biochemical attributes of known metal-binding sites, and can accurately predict metal binding sites and characterize their key features. We use a similar approach to solve the glucose binding site classification problem.

2.4.2 Identification of Hexose Binding Sites

There have been many attempts to identify proteins with hexose binding sites. Most of these attempts studied the galactose binding site, galactose being the

most common hexose in biological processes. Few studies tackled protein-glucose binding.

Rao, Lam, and Qasba [35] describe a computer model of the binding of galactose and mannose to the lectin protein family. They extensively study the lectin binding cavity that is formed of four amino-acid loops (termed A through D). These invariant residues are Aspartate (Asp) in loop A, Glycine (Gly) in loop B, Asparagine (Asn) in loop C, and an aromatic residue Phenylalanine/Tyrosine (Phe/Tyr) also in loop C. These four amino acids occupy identical positions independent of their sugar specificity and interact with the hexose independent of its type.

Quiocho and Vyas [33] present a review of the biochemical characteristics of carbohydrates-binding sites. They subdivide the carbohydrate-binding proteins in two major groups. Table 2.2 summarizes the differences among the two groups. The ordered water molecules are water molecules, present in the binding-site, that take part in the binding process. The atomic thermal B-factor is a representation of the motion of atoms in a crystallized structure. It measures the variations of individual atoms from their point locations. In this work, we do not segregate glucose binding sites into Quiocho and Vyas groups. We rather characterize common features to all glucose-binding sites.

Table 2.2: Differences between Group I and Group II carbohydrate-binding proteins.

Features	Group I	Group II
Binding site location	Deep cleft	Shallow
Sugar affinity	High	Low
Substrates	Monosaccharides	Oligosaccharides
Ligand specificity	Specific	Multivalent
Ordered water molecules	Low	High
B-factor	Low	High
Hydrogen bonds per monomer	High	Low
Van der Waals contacts per monomer	High	Low
Sugar polar groups	All paired	Some unpaired

Quiocho and Vyas stress on the importance of hydrogen bonds in stabiliz-

ing the sugar binding. They identify the planar polar residues asparagine (Asn), aspartate (Asp), glutamine (Gln), glutamate (Glu) and arginine (Arg) as the most frequently involved residues in hydrogen bonding. They also describe an important common feature to all protein-hexose docking: aromatic residues stacking against the apolar sugar pyranose ring. The hexose pyranose ring has a planar apolar hydrophobic side. By hydrophobic and van der Waals interactions, the hexose apolar side stacks over hydrophobic residues, namely aromatic residues having a planar apolar hydrophobic ring within their side chain. They report the stacking aromatic residues to be tryptophan (Trp), tyrosine (Tyr), phenylalanine (Phe) and histidine (His). Finally, Quiocho and Vyas pinpoint the role of ordered water molecules and metal ions in determining substrate specificity and affinity.

Taroni, Jones and Thornton [42], in an effort to predict sugar binding sites, analyze their characteristic properties. They found that the sugar interface residue propensity parameter best discriminates the sugar binding sites from other protein surface patches. For a residue i, they define the propensity quotient P as the ratio between the residue frequency at the binding site and the average frequency of any residue at the binding site.

$$P(i) = \frac{N_c(i)/N_d(i)}{T_c/T_d} \tag{2.2}$$

where $N_c(i)$ is the number of amino acids of type i making a contact with the sugar, T_c is the overall number of residues in contact with the sugar, $N_d(i)$ is the number of surface amino acids of type i in the dataset, and $T(d)$ is the total number of surface residues in the dataset.

Table 2.3 lists the different amino acids sugar interface propensity values. A higher propensity value reflects a higher tendency of that residue to be involved in sugar binding. In total accordance with previous reviews, the residues having high propensity values are the aromatic residues, histidine and the planar polar residues.

Table 2.3: Natural logarithm of the sugar interface propensity values for the standard amino acids (reproduced from Taroni *et al.* [42]).

Residue	Propensity	Residue	Propensity
Tryptophan (Trp)	1.40	Leucine (Leu)	-0.17
Histidine (His)	0.95	Glycine (Gly)	-0.40
Tyrosine (Tyr)	0.80	Isoleucine (Ile)	-0.41
Glutamate (Glu)	0.52	Cysteine (Cys)	-0.43
Arginine (Arg)	0.50	Alanine (Ala)	-0.48
Aspartate (Asp)	0.33	Serine (Ser)	-0.51
Phenylalanine (Phe)	0.09	Valine (Val)	-0.72
Asparagine (Asn)	0.08	Lysine (Lys)	-0.82
Methionine (Met)	0.02	Threonine (Thr)	-0.83
Glutamine (Gln)	0.02	Proline (Pro)	-1.50

Finally, Taroni, Jones and Thornton find that the sugar binding sites are neither hydrophobic nor hydrophilic. This is due to the dual nature of sugar docking, composed of a polar-hydrophilic aspect establishing hydrogen bonds and a hydrophobic aspect responsible for the pyranose ring stacking.

García-Hernández, Zubillaga, Chavelas-Adame, Vázquez-Contreras, Rojo-Domínguez and Costas [18] study the structural energetics of protein-carbohydrate interactions. They model binding energetics based on the changes of entropy, enthalpy, heat capacity and surface accessibility. They found that protein-carbohydrate complexes have distinctive properties, clearly differing from other protein systems. They observe a decrease in heat capacity due to the binding structural rearrangements, implying a dehydration of polar groups. Polar groups are thus behaving hydrophobically. They also report a low degree of carbohydrates apolar surfaces hydration. Their findings are consistent with previous reports of the apolar pyranose ring stacking over aromatic groups, and the extensive network of hexose-protein hydrogen bonds.

Sujatha and Balaji [39] formulate a signature for characterizing galactose-binding sites based on solvent accessibility and secondary structure types. They implemented a 3-D structure searching algorithm, COTRAN, which identifies galactose-

binding sites in proteins that share the same fold as the known galactose-binding proteins.

Sujatha, Sasidhar, and Balaji [40] demonstrate that the interaction of the hexose with the protein docking aromatic residue (tryptophan (Trp), tyrosine (Tyr) or phenylalanine (Phe)) is determined by their relative position as well as orientation. They show that the mode of binding of galactose relative to the aromatic residue is different from that of glucose. Galactose interacts with the aromatic residue favorably in positions and orientations observed for glucose and mannose, while the reverse situation does not hold. This implies that galactose-binding positions encompass those of glucose and mannose too. They conclude that the docking aromatic residue may not play a significant role in distinguishing glucose from galactose in glucose-specific proteins. They even note the absence of the docking aromatic residue in some glucose-binding sites.

Chakrabarti, Klibanov and Friesner [10] model a glucose binding site by optimization of binding affinity, under geometric and folding free energy constraints. Their aim is to uncover the evolutionary pressure determining the binding site sequence. They note that both cysteine (Cys) and proline (Pro) are not generally involved in binding site recognition and activity. In fact, both residues score low propensity values in Table 2.3.

CHAPTER 3

PATTERN RECOGNITION

Pattern recognition is the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and assign new objects to their correct classes. The pattern recognition process includes two phases: (1) the choice of a classification algorithm and (2) feature representation.

3.1 Statistical Pattern Recognition

This section explains the basic classification concepts and details the SVM and kNN algorithms.

3.1.1 Basic Classification Concepts

3.1.1.1 Pattern Classification Approaches

Pattern classification (also called pattern recognition or pattern learning) requires the computer to learn how to correctly classify an object. Its goal is to recover the original mathematical function that generated the present patterns. The different approaches to machine learning can be grouped in three categories depending on the extent of human involvement in the learning process [16]:

1. Supervised Classification: a teacher provides a class label for each input object.
2. Unsupervised Classification: also referred to as clustering. The algorithm forms clusters of the unlabeled input patterns.
3. Reinforcement Classification: the process of learning with a critic. The feedback is binary: it just indicates if the classification is right or wrong.

The supervised classification algorithms can be grouped into four categories [24]:

1. Template Matching: a template (or prototype) of the pattern to be recognized is available. The input data is matched against the stored template.

2. Statistical Matching: the pattern is represented in terms of its features. The input object is represented as a point in a multi-dimensional space. The classifier searches for a hyperplane separating the different classes.

3. Structural Matching: the classifier uses a hierarchical approach, where a pattern is viewed as being composed of a collection of elementary subpatterns, linked together by a set of rules. In this group, classifying a pattern reduces to determining the subpatterns and the rules.

4. Neural Networks: the classifier is a dense network of nodes interconnected by weighted edges. Weights are tuned by a gradient-descent approach. A complex combination of the weight values together with the input features determine the input class.

Our approach uses supervised statistical matching to model the hexose binding site in terms of its spatio-chemical attributes. The statistical classification technique consists of two phases: (1) a *learning* or *training* phase that processes descriptions of the input data and learns how to partition the feature space; (2) a *classification* or *testing* phase, where the classifier assigns the input data to one of the classes, using the pattern identified in the learning phase. In our case, we adopt a binary recognizer classifying queried instances into either hexose-binding sites or hexose non-binding sites.

The training input consists of $(\boldsymbol{x}_i, y_i)$ pairs, where $\boldsymbol{x}$ is the input sample vector and y its binary class label. Having q attributes per sample, the training input space reduces to a subset of $\mathbb{R}^q$. The input sample $\boldsymbol{x}$ is therefore coded as a

feature or *attribute* vector $\boldsymbol{x} = (x_1, \ldots, x_q)^t$. The class label y is binary, $y \in \{-1, 1\}$. The queried binding site is either a glucose binding site, or is not.

3.1.1.2 Bayesian Learning

The statistical method's underlying concept is *Bayes rule*, which provides a probabilistic approach to classification. Each instance classification into a certain class is regarded as a probable hypothesis. Bayes theorem computes hypothesis probabilities and classifies its input according to the hypothesis's prior probability, the probabilities of observing previous input given the hypothesis, and the input object itself [28].

The *prior probability* of a hypothesis h, denoted $P(h)$, is the probability of h before observing a new training input. Let $P(\boldsymbol{x})$ be the prior probability that training data vector $\boldsymbol{x}$ will be observed. The *posterior probability* of h, termed $P(h|\boldsymbol{x})$, is the probability of h holding, knowing that $\boldsymbol{x}$ has occurred. Equation 3.1 formulates the Bayes theorem.

$$P(h|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|h)\ P(h)}{P(\boldsymbol{x})} \tag{3.1}$$

Bayes theorem computes new probabilities in accordance with newly acquired info. These new probabilities are actually conditional probabilities [20].

The maximally probable hypothesis, the one with the highest posterior probability $P(h|\boldsymbol{x})$, is termed the *maximum a posteriori* (MAP) hypothesis. All hypotheses/classes have a probability assigned to them, but the classifier will assign the input object to the class whose posterior probability is the highest. Bayesian methods require initial knowledge of many probabilities, not all of which may be available. When they are not known in advance, they are estimated based on prior knowledge, available data or some statistical assumptions [28]. By identifying prob-

ability distributions $P(h)$ and $P(\boldsymbol{x}|h)$, it is possible to generate MAP hypotheses and classify $\boldsymbol{x}$ accordingly.

Statistical pattern recognition relies on the probability distributions of the patterns belonging to each class, which must either be specified or learned. The input objects are considered to be randomly drawn from the class conditional probability function $P(\boldsymbol{x}|h_i)$. If all of the class conditional densities are specified, then Bayes theorem can be applied. A parametric decision problem arises if only the form of the densities is known, but some of the parameters of the densities are unknown. Finally, if even the densities forms are unknown, then density estimation and nonparametric modes are used [24].

3.1.1.3 Density Estimation

As Duda *et al.* explain, the probability P that an instance $\boldsymbol{x}$ will fall in a region R is [16]:

$$P = \int_R p(\boldsymbol{x}')\, d\boldsymbol{x}' \tag{3.2}$$

where $p(\boldsymbol{x})$ is the density function.

Taking n samples independently and identically distributed, the probability that exactly k of these n samples fall in R is given by the binomial law:

$$P_k = \binom{n}{k} P^k \,(1-P)^{n-k} \tag{3.3}$$

The expected value (mean) for k is:

$$\varepsilon[k] = n\,P \tag{3.4}$$

Since this binomial distribution peaks sharply around the mean, the probability P can be accurately estimated by the ratio k/n:

$$P \approx \frac{k}{n} \tag{3.5}$$

Assuming that $p(\boldsymbol{x})$ is continuous, that n is large and that R is small, then

$$P = \int_R p(\boldsymbol{x}')\, d\boldsymbol{x}' \approx p(\boldsymbol{x})\, V \tag{3.6}$$

where V is the volume of the region R.

Combining equations 3.5 and 3.6 we get the following estimate:

$$p(\boldsymbol{x}) \approx \frac{\frac{k}{n}}{V} = \frac{k}{n\, V} \tag{3.7}$$

The k-Nearest-Neighbor approach is one way to construct sequences of regions satisfying the above conditions. k is specified to be a function of the training data n; the volume V is grown until it contains k neighbors of the input object $\boldsymbol{x}$.

3.1.2 k-Nearest-Neighbor

kNN is a simple algorithm whose resulting classifying rules are easy to analyze. We first introduce the basic algorithm, then discuss the more advanced weighted version.

3.1.2.1 Simple kNN

kNN is a statistical nonparametric learning method that stores the training examples and every new query is assigned a class label according to its neighbors' classes [48]. kNN assumes that all instances are points in the $\mathbb{R}^q$ space, where each instance is described in terms of q attributes. The nearest neighbors of an instance $\boldsymbol{x}$ are the samples whose "distance" away from $\boldsymbol{x}$ is below a certain threshold, the "distance" being a metric function between instances. kNN permits to locally estimate the original mathematical function for each input instance [28].

The "distance" parameter can be computed in many ways; the Euclidean distance being one form of distance metrics, among others. Duda *et al.* [16] describe metrics as follows:

> A metric must have the following four properties, for all vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$:
>
> 1. Nonnegativity: $D(\boldsymbol{a}, \boldsymbol{b}) \geq 0$
> 2. Reflexivity: $D(\boldsymbol{a}, \boldsymbol{b}) = 0$ iff $\boldsymbol{a} = \boldsymbol{b}$
> 3. Symmetry: $D(\boldsymbol{a}, \boldsymbol{b}) = D(\boldsymbol{b}, \boldsymbol{a})$
> 4. Triangle Inequality: $D(\boldsymbol{a}, \boldsymbol{b}) + D(\boldsymbol{b}, \boldsymbol{c}) \geq D(\boldsymbol{a}, \boldsymbol{c})$

A general metric measure in the $\mathbb{R}^q$ space is the *Minkowski* distance

$$L_j(\boldsymbol{a}, \boldsymbol{b}) = \left(\sum_{i=1}^{q} |a_i - b_i|^j \right)^{1/j} \tag{3.8}$$

where L_1 is the *absolute* or *Manhattan* distance, and L_2 is the *Euclidean* distance.

Using the density estimation function $p(\boldsymbol{x}) = \frac{k}{n\,V}$ (Equation 3.7), the estimate of the joint probability $p(\boldsymbol{x}, h_i)$ is:

$$p(\boldsymbol{x}, h_i) = \frac{k_i}{n\,V} \tag{3.9}$$

where k_i samples within V space are from class h_i. The posterior probability $P(h_i|\boldsymbol{x})$ becomes:

$$P(h_i|\boldsymbol{x}) = \frac{k_i}{k} \tag{3.10}$$

This leads to the MAP hypothesis, where the new input $\boldsymbol{x}$ is estimated to be the most common class label among the k training examples nearest to $\boldsymbol{x}$. Duda *et al.* [16] convey the proof that, with an unlimited number of samples, the error rate of kNN is never worse than twice the Bayes rate, the Bayes classification being the

optimal. For an optimal classification, k should be a small fraction of the training samples n.

3.1.2.2 Distance-Weighted kNN

An improvement over the simple kNN algorithm is the Distance-Weighted k-Nearest Neighbor algorithm [19, 28]. The contributions of the different neighbors towards the posterior probabilities calculations are weighted according to their distance away from the query point. The closer the neighbor, the higher its influence on the input classification.

Many different weighting functions, called *kernels*, have been devised to convert the "distances" into weights. For a proper kernel function K and a distance d, the following properties must hold [19]:

- $K(d) \geq 0$ for all $d \in \mathbb{R}$
- $K(d)$ gets its maximum for $d = 0$
- $K(d)$ descends monotonously for $d \to \pm\infty$

Table 3.1 reproduce the Hechenbichler and Schliep [19] listing of the typical kernel functions. It is worth noting that the rectangular kernel is the standard non-weighted one and assigns equal weights to all the neighbors.

Some kernel functions require the relative distances away from the queried sample to be normalized. Since only k nearest neighbors are used, then all the distances can be normalized relatively to the distance of the $(k+1)^{th}$ neighbor:

$$d(\boldsymbol{x}, \boldsymbol{x}_i) = \frac{D(\boldsymbol{x}, \boldsymbol{x}_i)}{D(\boldsymbol{x}, \boldsymbol{x}_{k+1})} \quad \text{for} \quad i = 1, \ldots, k \tag{3.11}$$

where d is the relative distance and D the measured distance. It follows that the normalized distance will be bounded within the $[0, 1]$ interval.

Table 3.1: Typical kNN kernel functions

Kernel type	$K(d)$
Rectangular kernel	$\frac{1}{2}$
Triangular kernel	$1 - \|d\|$
Epanechnikov kernel	$\frac{3}{4}(1 - d^2)$
Biweight kernel	$\frac{15}{16}(1 - d^2)^2$
Triweight kernel	$\frac{35}{32}(1 - d^2)^3$
Cosine kernel	$\frac{\pi}{4}\cos(\frac{\pi}{2}\, d)$
Gaussian kernel	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{d^2}{2}\right)$
Inversion kernel	$\frac{1}{\|d\|}$
Rank kernel	$(k+1) - rank(d)$

Distance-weighted kNN is robust to noisy training data and minimizes the effects of outliers. With distance weighting, one can even include all training samples in the probability calculations.

3.1.3 Support Vector Machines

SVM is a parametric statistical classifier that outperforms, in many cases, other classification methods [37]. This section explains the SVM underlying concepts.

3.1.3.1 Linear Discriminants

In an $\mathbb{R}^q$ space, a linear discriminant function f for an input sample $\boldsymbol{x} = (x_1, \ldots, x_q)^t$ is of the form:

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{w}^t \cdot \boldsymbol{x} + w_0 \\ &= \sum_{i=1}^{q} w_i^t \cdot x_i + w_0 \end{aligned} \tag{3.12}$$

where $\boldsymbol{w}$ is the *weight vector* and w_0 the *bias* [13]. The corresponding binary classifier assigns a new example $\boldsymbol{x}$ to class $y = +1$ if $f(\boldsymbol{x}) \geq 0$ and to class $y = -1$ otherwise. Baldi and Brunak [4] incorporate the *posterior probability* concept used in Bayesian learning (see Section 3.1.1.2) in a proper probabilistic setting, and assign the input $\boldsymbol{x}$ to class y according to the following function:

$$y = \text{sign}(f(\boldsymbol{x})) = \text{sign}(\log \frac{P(h^+|\boldsymbol{x})}{P(h^-|\boldsymbol{x})}) \tag{3.13}$$

this linear function can be visualized as a hyperplane R, defined by $\boldsymbol{w}^t \cdot \boldsymbol{x} + w_0 = 0$, that splits the $\mathbb{R}^q$ space. The positive class region extends above the hyperplane; the negative class lies beneath it. The weight vector $\boldsymbol{w}$ is perpendicular to the hyperplane, while the bias w_0 represents the plane's distance from the origin (see Figure 3.1).

The discriminant function $f(\boldsymbol{x})$ is a measure of the distance of the samples around the separating hyperplane R. $\boldsymbol{x}$ can be written as:

$$\boldsymbol{x} = \boldsymbol{x}_p + d \frac{\boldsymbol{w}}{||\boldsymbol{w}||} \tag{3.14}$$

where $\boldsymbol{x}_p$ is the normal projection of $\boldsymbol{x}$ onto R and d its signed distance [16] (see

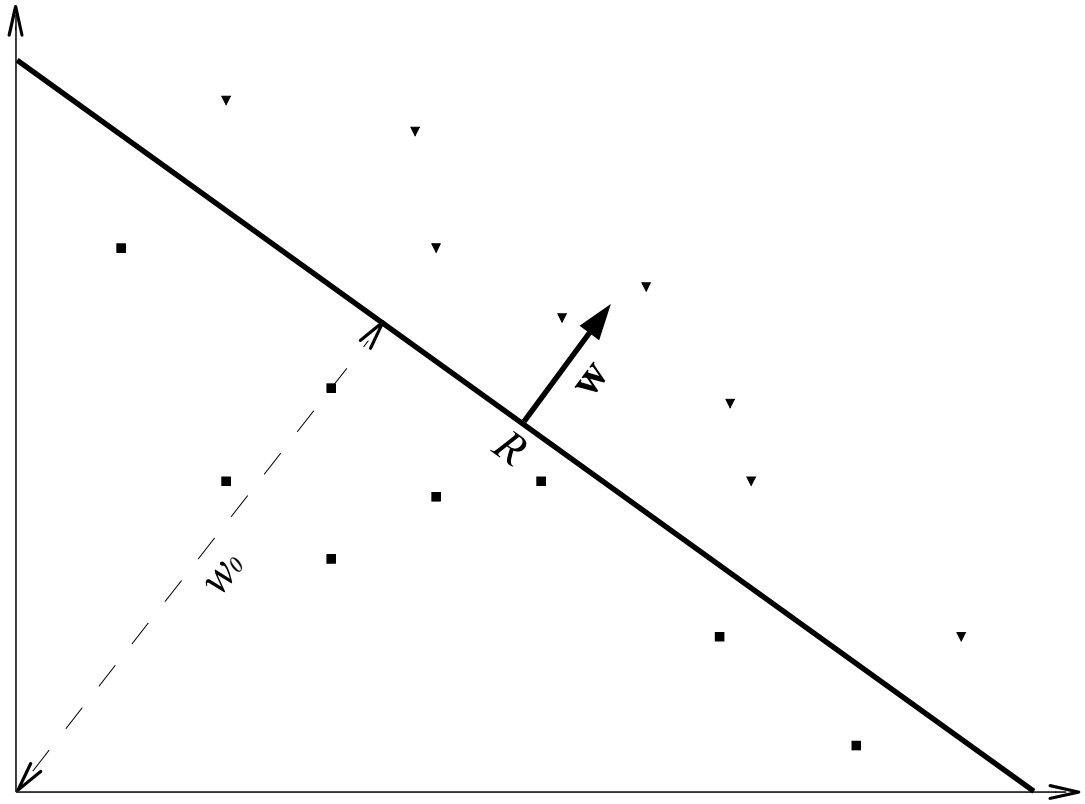


Figure 3.1: A separating hyperplane segregating two classes in a $\mathbb{R}^2$ space

Figure 3.2). Since $f(\boldsymbol{x}_p) = 0$, Equation 3.12 results in:

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{w}^t \cdot \left(\boldsymbol{x}_p + d \frac{\boldsymbol{w}}{||\boldsymbol{w}||} \right) + w_0 \\ &= d \, ||\boldsymbol{w}|| \end{aligned} \tag{3.15}$$

where d is positive if $\boldsymbol{x}$ is located in the positive region and negative otherwise.

The *functional margin* γ_i of a training instance $(\boldsymbol{x}_i, y_i)$ with respect to the hyperplane $(\boldsymbol{w}, w_0)$ is its signed distance away from the hyperplane:

$$\gamma_i = y_i \, f(\boldsymbol{x}_i) = y_i \, (\boldsymbol{w}^t \cdot \boldsymbol{x}_i + w_0) \tag{3.16}$$

where the functional margin of a linear classifier is the minimal margin of the training set. The *geometric margin* is the functional margin as computed for the normalized function.

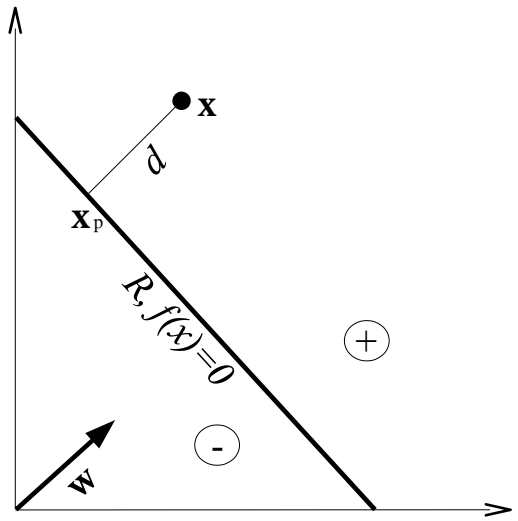


Figure 3.2: Projection of input $\boldsymbol{x}$ onto hyperplane $R : \boldsymbol{w}^t \cdot \boldsymbol{x} + w_0 = 0$

Knowing that $y \in \{-1, 1\}$, and using Equation 3.15, one can conclude that $\boldsymbol{x}_i$ is correctly classified by the decision rule f if $\gamma_i > 0$.

Perceptrons are simple classifiers that can learn these linear discriminant functions. Novikoff's theorem [13] proves that the perceptrons will converge to a hyperplane and correctly classify the data, if such a hyperplane exists. If it is the case, the data is said to be *linearly separable*. A perceptron begins with a random weight vector $\boldsymbol{w}$ and learns by iteratively updating the weights whenever it misclassifies an example. Thus, perceptrons and all other linear machines converge by maximizing the margin γ of the training sample.

For a fixed training sample S, let us assume that the initial weight vector is the zero vector and it is updated each time a misclassification occurs. The final weight vector becomes a linear combination of the number of times α_i the sample $\boldsymbol{x}_i$ was misclassified. The decision rule (Equation 3.12) takes its *dual* form:

$$f(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j y_j (\boldsymbol{x}_j^t \cdot \boldsymbol{x}) + w_0 \tag{3.17}$$

where the discriminant function is now expressed as a linear combination of the

training examples.

Dual representation This form is crucial because the data is encapsulated within the *Gramm matrix* $\boldsymbol{G} = (\boldsymbol{x}_i^t \cdot \boldsymbol{x}_j)_{i,j=1}^n$. The individual attribute values are *not* needed explicitly. The number of parameters within the dual function does not depend any more on the number of attributes (and attribute space dimension) q. It is only the scalar product of the training data with the new input sample that matters [13].

3.1.3.2 Kernel Mapping

The linear discriminant function has an inherent limitation: it is unable to converge to a hyperplane in a nonlinearly separable data set. To overcome this problem, one can perform a non-linear mapping of the input space to a new (potentially higher dimensional) feature space, to which a linear machine can be applied [4]. Let $\phi : X \rightarrow F$ be the non-linear mapping function from the input space X to the feature space F. The dual discriminant function in the new feature space will be:

$$f(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j y_j (\phi(\boldsymbol{x}_j)^t \cdot \phi(\boldsymbol{x})) + w_0 \tag{3.18}$$

As noted by Cristianini and Shawe-Taylor [13], the original quantities x_i are sometimes referred to as *attributes*, while the newly introduced quantities f_i are usually called *features*. This thesis document observes this distinction.

Instead of applying the mapping function ϕ to each instance vector and then computing the scalar product in the feature space, it is possible to merge these steps in one function K, called a *kernel* function:

$$K(\boldsymbol{a}, \boldsymbol{b}) = \phi(\boldsymbol{a})^t \cdot \phi(\boldsymbol{b}) \qquad \text{for} \quad \boldsymbol{a}, \boldsymbol{b} \in X \tag{3.19}$$

The discriminant function (Equation 3.18) becomes:

$$f(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j y_j K(\boldsymbol{x}_j, \boldsymbol{x}) + w_0 \tag{3.20}$$

where α_i measures the importance of example i, and the kernel $K(\boldsymbol{x}_i, \boldsymbol{x})$ measures the similarity of vectors $\boldsymbol{x}$ and $\boldsymbol{x}_i$.

Although kernels map input instances to a higher feature space, they perform the computation in the input space [37]. This is because the kernel function (Equation 3.19) is a Gramm matrix.

Many different kernels exist, the basic ones are [22]:

- Linear kernel: $K(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^t \cdot \boldsymbol{b}$
- Polynomial kernel: $K(\boldsymbol{a}, \boldsymbol{b}) = (\kappa(\boldsymbol{a}^t \cdot \boldsymbol{b}) + r)^m, \quad \kappa > 0$
- Radial Basis Function (RBF) kernel: $K(\boldsymbol{a}, \boldsymbol{b}) = \exp(-\kappa||\boldsymbol{a} - \boldsymbol{b}||^2), \quad \kappa > 0$
- Sigmoidal kernel: $K(\boldsymbol{a}, \boldsymbol{b}) = \tanh(\kappa(\boldsymbol{a}^t \cdot \boldsymbol{b}) + r)$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are attribute vectors and κ, r and m are kernel parameters. The sigmoidal kernel is sometimes referred to as the neural network kernel.

3.1.3.3 Support Vector Machines Implementation

Although mapping the attributes to higher dimensions extends the power of linear machines, it creates a generalization problem. A higher dimensional space increases the risk of overfitting and performs poorly on new data.

Let H be the *hypothesis space*—the class of decision rules that the learner may consider—for a fixed set S of training samples. Overfitting can be avoided by restricting the complexity of H, whereas a simple linear rule correctly segregating most of the data is preferable to a complex one (Occam's razor) [30]. Assuming independently and identically distributed input samples, the expected classification error of a certain hypothesis h depends on the random selection of the training set.

A good generalization implies a tight bound (or limit) on the difference between empirical and true estimates.

The Vapnik-Chervonenkis (VC) theory [13, 30] establishes consistent bounds on the generalization of linear machines and proves that overfitting can be prevented. It suggests that a learner for a hypothesis space H should attempt to minimize the number of training classification errors, since all other variables in the generalization bound have been fixed by the choice of H. Different bounds on the generalized error have been devised, many of which involve maximizing the margins γ_i of the training samples (refer to Equation 3.16) while the dimension of the feature space does not appear in the formulas [13]. However, the margin can be measured with the weight vector $\boldsymbol{w}$. According to the functional margin definition, if one normalizes the margin γ so that $\gamma = 1$, then

$$|\boldsymbol{w}^t \cdot \boldsymbol{x} + w_0| = 1 \tag{3.21}$$

and the geometric margin becomes

$$\gamma = \frac{1}{||\boldsymbol{w}||}. \tag{3.22}$$

Knowing that, by definition,

$$||\boldsymbol{w}|| = \sqrt{\boldsymbol{w}^t \cdot \boldsymbol{w}} \tag{3.23}$$

then maximizing the margin γ reduces to minimizing $(\boldsymbol{w}^t \cdot \boldsymbol{w})$.

If the data is linearly separable, the maximal margin hyperplane $(\boldsymbol{w}, w_0)$ would be the one solving the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w}, w_0} \quad & \boldsymbol{w}^t \cdot \boldsymbol{w} \\ \text{subject to} \quad & \gamma = y_i \, f(\boldsymbol{w}^t \cdot \boldsymbol{x}_i + w_0) \geq 1, \quad i = 1, \ldots, n. \end{aligned} \tag{3.24}$$

In the more general case of noisy or nonlinearly separable data, slack variables ξ measuring the margin distribution are introduced. The optimal margin hyperplane becomes the one solving the "soft margin" optimization:

$$\begin{aligned} \min_{\boldsymbol{w}, w_0, \xi} \quad & \boldsymbol{w}^t \cdot \boldsymbol{w} + C \sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & \gamma = y_i \, f(\boldsymbol{w}^t \cdot \boldsymbol{x}_i + w_0) \geq 1 - \xi_i, \quad i = 1, \ldots, n, \\ & \xi_i \geq 0, \quad i = 1, \ldots, n. \end{aligned} \tag{3.25}$$

where C is a cost constant. As a result, and assuming an independently and identically distributed sample set S, the margin γ indicates the generalization accuracy with a higher γ resulting in a better generalization [17]. Overfitting is thus controlled by maximizing the margin through the correct choice of the kernel function.

Support Vector Machines [44] are linear classifiers that use kernel mapping, integrate the VC theory, and optimize the margin distribution. The SVM error surface is a convex one, resulting in the absence of local minima. Cristianini and Shawe-Taylor [13] explain how the maximal margin optimization problem reduces to a quadratic programming problem with a unique solution that can be found efficiently. Therefore SVM converges to the optimal separating hyperplane R, given the examples set S and the feature space F.

According to the dual representation (see Equation 3.17), the hyperplane R can be expressed as a linear function of the training set S. However, for many optimizations, fewer number of training examples influence R. SVM compresses the data retaining only those inputs specifying the weight vector of the maximal margin hyperplane R. These selected points are called *support vectors* (see Figure 3.3). The hyperplane R is therefore a function of the support vectors training samples. A smaller number of support vectors reflects a better generalization.

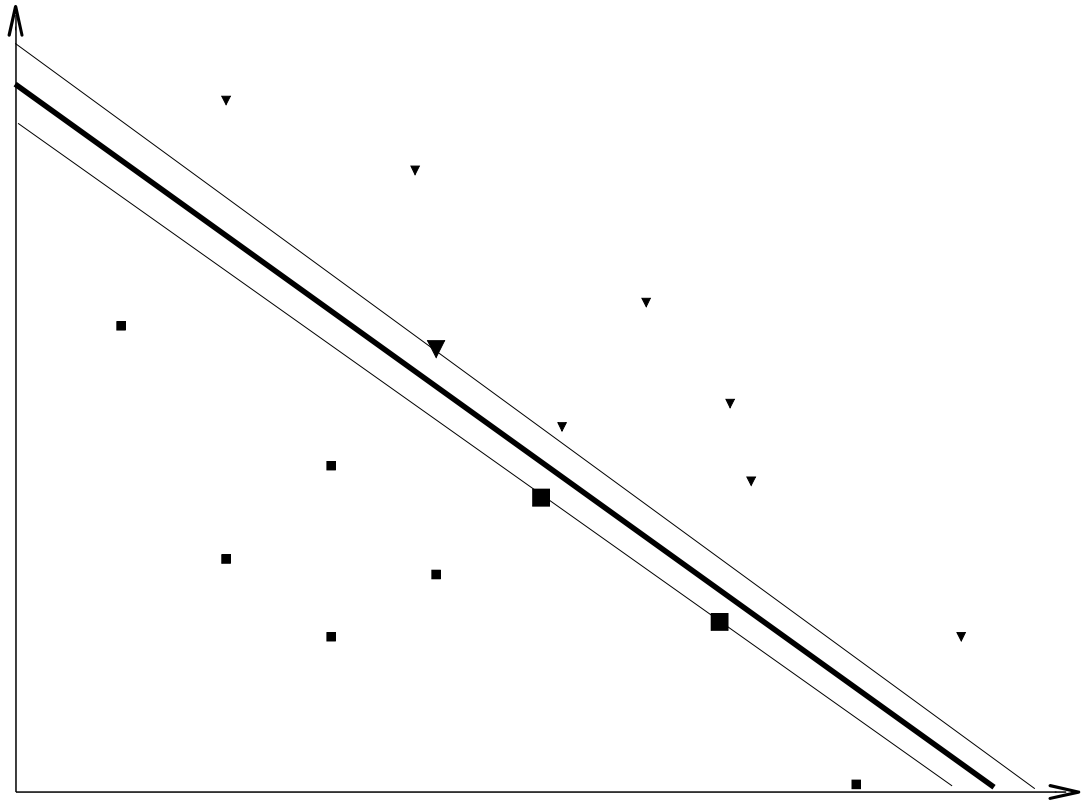


Figure 3.3: A separating hyperplane with its support vectors highlighted.

3.2 Feature Representation

Feature representation constitutes an important step in designing classifiers. Some feature combinations can effectively partition the input space, while others are completely irrelevant. Which features to use? How to represent them? How many to retain? Most of the art of pattern classification resides in the correct choice of features/attributes. Knowing that an increase in the feature dimensionality tends to increase measurement cost and decrease classification accuracy [23], the aim of a good designer is to keep the feature number minimal, while maximizing the classifier performance.

3.2.1 Standardizing Data

Data measurements usually have different scales. Various features require diverse measuring units, the range varies... The data scale has a direct impact

on classification accuracy. A feature with a high data mean will a priori influence classification more than one with a small mean, regardless of their discrimination power [16].

In order to put equal initial weight on the different features, the data is standardized by scaling and centering. To scale, the variables are divided by their standard deviation. To center, the mean is subtracted.

3.2.2 Curse of Dimensionality

The *curse of dimensionality* problem arises when a classifier maps input instances to an $\mathbb{R}^q$ space where q, the number of features, is high. Statistically, having a large number of training samples leads to convergence towards complex target functions. The required number of samples n is an exponential function of the dimensionality of the feature space q, since the classifier needs an exponentially large number of patterns to sample the space properly [30].

Pattern classification techniques classify the inputs according to all their features. However, only a subset of the q features is relevant to determine the correct decision rule. Including the irrelevant features in the classification calculations is misleading and degrades the classifier performance. There are two methods to minimize the number of used features: (1) *feature selection* selects the best subset of the input features; (2) *feature extraction* creates new features by combining original ones.

Since the classifier performance depends both on the sample size and the number of features, a rule of thumb is to provide at least ten times as many training samples per class as the number of features ($n_i/q \geq 10$) [23].

3.2.3 Random Forests

Random Forests (RF) [7] is a classification algorithm based on multiple classification trees. RF provides measures of feature importance, and is used as a

feature selection tool.

3.2.3.1 RF Implementation Overview

The Random Forests algorithm is implemented as follows [8]:

1. For an original training set S with n training samples, RF creates j *bootstrap* datasets of n samples each. A bootstrap dataset is created by randomly selecting samples with replacement. A classification tree is grown over each dataset.

2. Let Q be the original feature number, and q a number such that $q \ll Q$. At each tree node, q features are selected randomly. The node is split according to the best split among the selected q features.

3. Each tree remains unpruned, which generates low-bias trees.

When RF is used for classification, the new sample is fed to all the single classification trees. Each tree returns its class prediction, or casts its "vote". The forest returns the final classification, the one receiving the largest amount of votes.

A *bagging* classifier is one that samples bootstrap datasets, feeds each one to a different component classifier, and classifies based on majority vote [16]. RF combines bagging, random feature selection and unpruned trees to achieve both low bias and low variance [14].

3.2.3.2 Feature Importance Computation

A bootstrap dataset of n samples, sampled (with replacement) from an original set S of n observations, contains about 2/3 of set S items. Approximately 1/3 of the original cases are not included in the bootstrap set and hence are not used in the corresponding tree construction. The remaining 1/3 of S data, not used for the tree construction, is called the tree *out-of-bag* (OOB) data [7].

RF measures feature importance as follows [8], for each feature q and each tree t:

1. Let the tree t classify its own OOB data.
2. Compute the number of correctly classified samples (c_i).
3. Permute the values of feature q in the OOB samples.
4. Let t reclassify the modified OOB and recompute the new number of correctly classified samples (c_f).

The raw importance score r of feature q over tree t is:

$$r_{(q,t)} = c_i - c_f \tag{3.26}$$

The feature raw importance score returned by the forest is the average of the feature's raw importance score over each tree. It measures the decrease of classification accuracy when values of a feature are randomly permuted.

RF feature selection method is robust to noise, can be used when the number of features q is much larger than the number of observations n, incorporates feature interactions, and returns a direct feature importance measure [7]. RF matches or outperforms other feature selection approaches [14].

CHAPTER 4

PROBLEM REPRESENTATION AND IMPLEMENTATION

A good classifier must be trained on representative data and on the right features. This chapter explains how the data was obtained, and details the feature representation and how the data is used.

4.1 Data

4.1.1 Database Availability

The Protein Data Bank (PDB) [5] is a repository of experimentally determined and hypothetical three-dimensional structures of biological macromolecules. The protein configurations are uncovered using X-Ray Crystallography or Nuclear Magnetic Resonance (NMR). A PDB file contains, among others, atomic coordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic or NMR experimental data.

A PDB file follows a strict format and is stored as a sequence of records in a flat file. Table 4.1 lists the various sections of a PDB file. Each section has a certain number of specific records, and each record follows a strict format [1]. Perl (Practical Extraction and Report Language) [43, 46] is an interpreted language suitable for scanning arbitrary text files, extracting information from these text files, and printing reports based on the extracted information. Perl has emerged as a leading language for processing structured flat files [29]. An extension of Perl, BioPerl [43], includes functionality for bioinformatics data processing and analysis. It has special modules that parse a PDB file, represent it as an object and query its fields.

Table 4.1: The various sections of a PDB file (reproduced from Abola *et al.* [1]).

Section	Description
Title	Summary descriptive remarks.
Remarks	Bibliography, refinement, annotation...
Primary Structure	Peptide and/or nucleotide sequence and the relationship between the PDB sequence and that found in the sequence database(s).
Heterogen	Description of non-standard groups.
Secondary structure	Description of secondary structure.
Connectivity annotation	Chemical connectivity.
Miscellaneous features	Features within the macromolecule.
Crystallographic	Description of the crystallographic cell.
Coordinate transformation	Coordinate transformation operators.
Coordinate	Atomic coordinate data.
Connectivity	Chemical connectivity.
Bookkeeping	Summary information, end-of-file marker.

The PDB is a data repository for the work of thousands of labs over many years and, as such, contains some errors and many redundancies. Proteins are crystallized with different resolutions, some of which are obsolete. Caveat entries are deemed incorrect by an outside editorial board, and contain severe errors. In addition, the PDB contains hypothetical structures modeled by computer simulations and energy minimization.

4.1.2 Data mining step

The Protein Data Bank is the largest protein 3-D structure repository. However, it contains many redundant, obsolete, caveat or hypothetical structures. As such, any computational query result may return many potentially erroneous hits, and should be double checked. Data filtering through a thorough result check and an extensive literature review was performed. Data filtering is discussed throughout the current and the following sections.

In our case, we mine for both proteins crystallized with hexose docked into their active site and for proteins known to bind hexoses, but crystallized without

their ligand (*i.e.* the hexose). For the first case, we directly mine the PDB. However, for the second case, we must concentrate on literature research in order to identify such proteins.

To mine the PDB, we use the keywords of the required HET (heterogen) groups and perform an iterative matching search. HET records describe non-standard residues, such as ligands, prosthetic groups, inhibitors, solvent molecules, and ions for which coordinates are supplied. Furthermore, many of the obtained hits are glycoproteins, where the hexose is covalently bound to the protein backbone. No receptor-ligand interaction occurs between the protein and the sugar in this case, since the sugar is an integral part of the protein molecule itself! Each one of the obtained hits must be checked for consistency, to avoid glycoproteins, redundancies, and structures flagged as obsolete, caveat, or hypothetical.

4.1.3 PDB Filtering

To perform the data mining, we used the PDB repository present at the Bioinformatics Core lab in the Faculty of Medicine at AUB. This databank was last updated on October 2004 and contains 28353 proteins.

D-Glucose is able to take many forms (refer to Figure 2.6), each having a specific name and a corresponding HET name. An extensive search throughout the entire HET Dictionary revealed the existence of the following D-Glucose HETs:

1. GLC: D-Glucose
2. AGC: α-D-Glucopyranose
3. BGC: β-D-Glucopyranose

Parsing the whole PDB repository for the mentioned HETs, while discarding obsolete, caveat or hypothetical structures, revealed 331 glucose-containing proteins (292 GLC, 3 AGC, 36 BGC). As explained in Section 2.1.2.2, the D-Glucose

GLC occurrences are most likely crystallized as either α-D-Glucopyranose or β-D-Glucopyranose, but this distinction is not reported in the GLC-containing PDB file.

The glucoses present in the PDB files can be covalently bound with the protein forming a glycoprotein, or docking within a binding-site, or just floating around in the medium. This thesis is only concerned with the glucose docking within a binding site category. Computationally, the first category can be eliminated, while the dissociation of the docking from the floating glucoses needs a literature review and a trained biologist's intervention. After eliminating the glycoproteins, 311 proteins remained (273 GLC, 3 AGC, 35 BGC). Some of these proteins follow old PDB formats, older than format 2.1 (25 October 1996) and, as such, are not compatible with our feature extraction algorithm.

Colleagues at the University of Miami (Khuri and Al-Ali, personal communication) conducted a thorough biological check on the remaining hits and generated a list of 59 resolved protein-glucose binding sites following the compatible formats (format 2.1 and newer). Table 4.2 shows the selected proteins segregated into α- and β-D-Glucopyranoses. Some proteins contain more than one valid glucose-binding site.

Table 4.2: Inventory of the D-Glucose-binding proteins.

Glucose type	PDB entries
Distinct α-D-Glucopyranose	1BDG 1GJW 1GWW 1H5U 1HIZ 1HSJ 1K1W 1NSZ 1PWB 1S5M 1UA4 1V2B 2BQP
Distinct β-D-Glucopyranose	1EX1 1HKC 1I8A 1ISY 1J0Y 1JG9 1KME 1MMU 1NF5 1Q33 1WOQ 2BVW
Redundant α-D-Glucopyranose	1CZA 1DGK 1E1Y 1E6X 1K72 1LWN 1LWO 1MXD 1OFC 1QHA 1V4S 2GPA 2SKC 2SKD 2SKE 3AMV
Redundant β-D-Glucopyranose	1BG3 1CQ1 1HGY 1HKB 1IEQ 1JS4 1JSW 1JWY 1JX2 1NS4 1NS7 1NSR 1NSS 1NSV 1OSE

Mr. Al-Ali used the PISCES server [47] to perform a redundancy check. He clustered the entries according to homology, with a 30% maximum allowed homology between representative structures in the set. That is any two proteins in different sets are less than 30% structurally similar. After selecting one representative of each cluster, the training positive set is reduced to 29 distinct binding sites, as shown in Table 4.2. Table 4.3 shows, for each binding site, the PDB entry and the docked D-Glucopyranose.

Table 4.3: Inventory of the positive training set binding sites.

PDB entry	α-D-Glucopyranose	PDB entry	β-D-Glucopyranose
1BDG	GLC-501	1EX1	GLC-617
1GJW	GLC-701	1HKC	GLC-915
1GWW	GLC-1371	1I8A	GLC-189
1H5U	GLC-998	1ISY	GLC-1461
1HIZ	GLC-1381	1ISY	GLC-1471
1HIZ	GLC-1382	1J0Y	GLC-1601
1HSJ	GLC-671	1JG9	GLC-2000
1HSJ	GLC-672	1KME	GLC-501
1K1W	GLC-653	1MMUA	GLC-1
1NSZ	GLC-1400	1NF5	GLC-125
1PWB	GLC-405	1Q33	GLC-400
1S5M	AGC-1001	1WOQA	GLC-290
1UA4	GLC-1457	2BVW	GLC-602
1V2B	AGC-1203	2BVW	GLC-603
2BQP	GLC-337		

We generate the testing positive set in a similar way to the training positive set, while parsing the PDB repository for entries newer than October 2004. We identify 7 distinct glucose-binding entries. Table 4.4 lists the testing positive set.

Table 4.4: Inventory of the positive testing set binding sites.

PDB entry	D-Glucopyranose	PDB entry	D-Glucopyranose
1RYD	GLC-601	1S5M	AGC-1001
1SZ2	BGC-1001	1SZ2	BGC-2001
1U2S	GLC-1	1Z8D	GLC-901
2F2E	AGC-401		

4.1.4 Negative Entries

A classifier needs to process both positive and negative samples in its learning phase. The entries in Table 4.3 are the positive inputs for classification. Good prediction requires an equal number of inputs per class; colleagues at the University of Miami (Khuri and Al-Ali, personal communication) identified the negative set, composed of non-binding sites and binding-sites that do not bind hexoses (refer to Tables 4.5, 4.6, 4.7 and 4.8). The cavity center is computed as the centroid of the given atom numbers.

Table 4.5: Inventory of the set 1 negative training sites. The cavity center is computed as the centroid of the given atoms. The atoms are listed by their PDB atomic number.

PDB entry	Cavity center non-hexose binding site	Ligand	PDB entry	Cavity center non-binding site
11GS	1672 - 1675	MES-3	11AS	5132
1A42	2054 - 2055	BZO-555	1BSI	103 - 114
1A50	4939 - 4940	FIP-270	1C3P	1089 - 1576
1A53	2016 - 2017	IGP-300	1C5K	605 - 871
1AA1	4472 - 4474	3PG-477	1DXJ	867 - 1498
1AJN	6074 - 6079	AAN-1	1EVT	2149 - 2229
1AJS	3276 - 3281	PLA-415	1FSZ	2048 - 2190
1D09	7246	PAL-1311	1KLM	4373 - 4113
1F8I	13237	MG-451		
1FI2	1493	MN-202		
1IOL	2674 - 2675	EST-400		
1JTV	2136 - 2137	TES-500		
1KF6	16674 - 16675	OAA-702		
1NX8	6104 - 6109 - 6110	N7P-290		
1UK6	2142	PPI-1300		
3PCB	3421 - 3424	3HB-550		
1EQY	3831	ATP-380		
1AL8	2652	FMN-360		

The classification we use distinguishes between non-hexose binding sites and non-binding sites. Non-hexose binding sites are actual binding sites that do not bind hexoses. These proteins are crystallized with their non-hexose substrates.

Table 4.6: Inventory of the set 2 negative training sites.

PDB entry	Cavity center non-hexose binding site	Ligand	PDB entry	Cavity center non-binding site
2PAH	5318	FE-453	2BG9	1237
2BIW	15171	FE-1492	1A7W	351
1DY1	1423	ZN-401	1KWP	1212
1J1L	2246	FE2-1001		
1BOB	2566	ACO-400		
1TVO	2857	FRZ-1001		

Non-binding sites are protein areas that are crystallized without any ligand. They are random parts of the protein. Most of them are grooves located at the protein surface and look as if they may bind something. The literature does not report the binding of any substrate to these non-binding sites.

We divide the negative-input entries into four sets. The first two sets (Tables 4.5 and 4.6) are used in the training phase. Set 1 consists of 18 non-hexose binding sites and 8 non-binding sites. Set 2 adds 6 and 3 sites, respectively. The third set from Table 4.7, solely composed of non-binding sites, is used to analyze the hydrogen bond property. The fourth and last set from Table 4.8 constitutes the testing phase negative set.

Table 4.7: Inventory of the set 3 negative training sites. This set is solely composed of non-binding sites.

PDB entry	Cavity center	PDB entry	Cavity center
1A04	541	1A0I	1689 - 799
1A0P	65 - 2100	1A22	2927
1A33	604	1AA7	579
1AF7	631 - 1492	1AM2	1277
1ARO	154 - 1663	1BO4	208 - 1386 - 1929
1B6B	1871 - 2341	1BGF	284
1C3G	630 - 888	1C44	92 - 880
1YVB	1546 - 1814	2D7S	3786
1QZ7	3592 - 2509		

Table 4.8: Inventory of the set 4 negative testing sites.

PDB entry	Cavity center	PDB entry	Cavity center
1YQZ	4458 - 4269	1ZT9	1056 - 1188
2A1K	2758 - 3345	2C9Q	777
2CL3	123 - 948	2DN2	749 - 1006
2F1K	316 - 642	2G50	26265 - 31672
2G69	248 - 378	2GAX	326
2GRK	369 - 380	2GSE	337 - 10618
2GSH	6260	2GSV	1126
2GTR	400		

4.2 Solution Approach

This work implements both a SVM and a kNN classifier to detect hexose-binding sites. SVM is a parametric classifier that provides state-of-the-art performance in many domains [31]. kNN is a non-parametric classifier which remains accurate to small training samples and whose resulting classifying rules are easy to analyze [48].

The first step is to data mine and preprocess data from PDB files. The next step is a feature extraction step that describes each candidate site as a feature vector. The classifier is consequently trained on the generated feature vectors. Figure 4.1 outlines the classifier algorithm, similar for both SVM and kNN classifiers. The analysis step analyzes and compares the results of both classifiers. A modeling and feature selection step finalizes our work whereby the key features of each hexose-binding site type are characterized. A simulation of the docking receptor-ligand interaction can thus take place.

4.2.1 Feature Extraction

A binding site is characterized by its biochemical, biophysical and geometric properties. We devise a list of the main hexose-binding site properties and characteristics from literature scrutiny. These features are grouped into two categories,

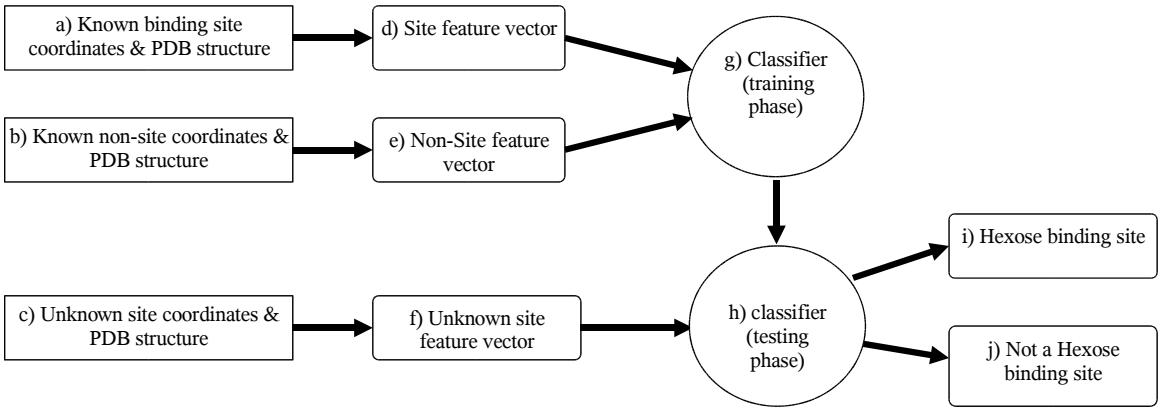


Figure 4.1: The classifier algorithm outline.

atomic features and residue features, which are described below.

4.2.1.1 Atomic Features

Based on biochemical knowledge and literature review, we identified three atomic properties that most likely determine hexose recognition and binding. The chosen atomic properties are:

1. Charge: the partial atomic charge (refer to Section 2.3.1.1).
2. Hydrogen Bond: the capacity to establish an atomic bond (refer to Section 2.3.1.2).
3. Hydrophobicity: the relative hydrophobicity level (refer to Section 2.3.2.2).

Table 4.9 details the atomic feature values used. It lists all the atoms that the algorithm accounts for along with their type, name, functional group and location. It first lists the amino acid atoms grouped by atom type (oxygen, nitrogen, carbon and sulfur), followed by the different HET atoms in the fifth and last group. All the atoms of the amino acids are listed, and their functional group identified.

Atomic properties are assigned nominal values. Charge is either positive, neutral or negative. Atoms are either capable of forming hydrogen bonds, or are

Table 4.9: Atomic features.

Atom Type	Functional Group	Location	Residue	PDB Atom Symbol	Chrg	Hydrophob	H Bond
Oxygen	Amide peptide linkage	Backbone	All	O	0	-1	H Bond
Oxygen	Carboxyl – C terminus	Backbone	All	OXT	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	GLU	OE1	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	GLU	OE2	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	ASP	OD1	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	ASP	OD2	-ve	-1	H Bond
Oxygen	Amide	Side Chain	GLN	OE1	0	-1	H Bond
Oxygen	Amide	Side Chain	ASN	OD1	0	-1	H Bond
Oxygen	Hydroxyl	Side Chain	SER	OG	0	-1	H Bond
Oxygen	Hydroxyl	Side Chain	THR	OG1	0	-1	H Bond
Oxygen	Hydroxyl - Phenolic	Side Chain	TYR	OH	0	-1	H Bond
Nitrogen	Amide peptide linkage	Backbone	All except PRO	N	0	-1	H Bond
Nitrogen	Amide peptide linkage	Backbone	PRO	N	0	-1	--
Nitrogen	Amide	Side Chain	GLN	NE2	0	-1	H Bond
Nitrogen	Amide	Side Chain	ASN	ND2	0	-1	H Bond
Nitrogen	Amine	Side Chain	LYS	NZ	+ve	-1	H Bond
Nitrogen	Guanidino	Side Chain	ARG	NE	+ve	-1	--
Nitrogen	Guanidino	Side Chain	ARG	NH1	+ve	-1	H Bond
Nitrogen	Guanidino	Side Chain	ARG	NH2	+ve	-1	H Bond
Nitrogen	Imidazole	Side Chain	HIS	ND1	0	-1	--
Nitrogen	Imidazole	Side Chain	HIS	NE2	0	-1	H Bond
Nitrogen	Indole	Side Chain	TRP	NE1	0	0	--
Carbon	Amide peptide linkage	Backbone	All	C	0	0	--
Carbon	C-alpha	Backbone	All	CA	0	0	--
Carbon	Aliphatic – neutral	Side Chain	Set A (See below)	CB, CG, CD, CE	0	0	--
Carbon	Aliphatic – hydrophobic	Side Chain	LEU, VAL, ILE, MET	CB, CG, CD, CE	0	1	--
Carbon	Aliphatic – Branch	Side Chain	LEU, VAL, ILE	CG1, CG2, CD1, CD2, CD1	0	1	--
Carbon	Phenyl - aromatic	Side Chain	PHE, TYR	CG,CD1, CD2, CE1, CE2, CZ	0	1	--
Carbon	Imidazole	Side Chain	HIS	CG, CD2, CE1	0	1	--
Carbon	Aromatic	Side Chain	TRP	CG,CD1, CD2,	0	1	--
Carbon	Aromatic	Side Chain	TRP	CE2, CE3, CZ2, CZ3, CH2	0	1	--
Sulfur	Sulfhydril	Side Chain	CYS	SG	0	-1	H Bond
Sulfur	Thioether	Side Chain	MET	SD	0	0	--
Oxygen	Sulfate	HET Group	SO4	O1, O2, O3, O4	-ve	-1	H Bond
Oxygen	Phosphate	HET Group	2HP	O1, O2, O3, O4	-ve	-1	H Bond
Oxygen	Water	HET Group	HOH	O	0	-1	H Bond
Calcium	Ion	HET Group	CA	CA	+ve	-1	H Bond
Magnesium	Ion	HET Group	MG	MG	+ve	-1	H Bond
Zinc	Ion	HET Group	ZN	ZN	+ve	-1	H Bond

Set A = ALA, SER, THR, CYS, ASP, ASN, GLU, GLN, ARG, LYS, PRO

not. Hydrophobicity levels are +1 (hydrophobic), 0 (neutral) and -1 (hydrophilic). For example the central carbon of an amino acid, portrayed in Figure 2.7, is called "carbon alpha". It is listed in the second carbon record of Table 4.9. C_α is a backbone atom present in all residues, its PDB atom symbol is "CA", it does not bear a partial atomic charge, is not capable of establishing a hydrogen bond, and has a neutral hydrophobicity level.

Due to their contribution in hexose binding [33], water molecules and some ions are taken into account while sampling atomic features (refer to Table 4.9). The sampled ions are: Sulfate, Phosphate, Calcium, Magnesium and Zinc.

Finally, due to the different PDB file resolutions, hydrogen atoms are not accounted for. In fact, most PDB files do not report hydrogen atom positions.

4.2.1.2 Residue Features

As discussed earlier, amino acids are grouped based on their spatio-chemical properties. The residue's side chain characterizes the amino acid's functionality. Some side chains are closely related, giving their respective amino acids similar biochemical aspects. Amino acid groupings differ depending on the relative importance given to the different side chains properties.

We generate different residue subgrouping schemes that incorporate biochemical knowledge and hexose-protein interaction studies findings. Out of general biochemical knowledge, we first devised a detailed scheme, called "detailed1" and listed in Table 4.10.

Table 4.10: The two detailed residue subgrouping schemes, "detailed1" and "detailed2".

Scheme Name	Subgroups	Residues
Detailed1	Aromatic	Phe, Tyr, Trp
	Neutral	Gln, Asn, Ser, Thr, Pro, Gly, Cys
	Carboxylate	Glu, Asp
	Aliphatic	Ala, Val, Leu, Ile, Met
	Positively Charged	Lys, Arg
	Histidine	His
Detailed2	Aromatic	Phe, Tyr, Trp, His
	Neutral	Gln, Asn, Ser, Thr, Pro, Gly, Cys
	Carboxylate	Glu, Asp
	Aliphatic	Ala, Val, Leu, Ile, Met
	Positively Charged	Lys, Arg

Quiocho and Vyas [33] stress the importance of both the planar polar residues (asparagine (Asn), aspartate (Asp), glutamine (Gln), glutamate (Glu) and

arginine (Arg)) and the aromatic residues (tryptophan (Trp), tyrosine (Tyr), phenylalanine (Phe) and histidine (His)) for hexose docking and binding. Sujatha *et al.* [40] study the glucose docking on top of aromatic residues, reporting the involvement of tryptophan (Trp), tyrosine (Tyr) and phenylalanine (Phe). This difference in identifying aromatic residues is due to the fact that tryptophan, tyrosine and phenylalanine have a benzine-ring derivative as their side chain, while histidine has an imidazole-ring. Since imidazole is biochemically distant from benzene, Quiocho and Vyas justify their grouping by the common hydrophobic properties of these residues, whereas both ring types offer a stacking interaction to the hydrophobic faces of hexose. The "detailed2" scheme adds histidine to the aromatic subgroup. Residue schemes "simplified1", "simplified2" and "simplified3" test for the relevance of the above findings (refer to Table 4.11).

Planar polar residues are involved in glucose-protein hydrogen bonding. The "simplified1"scheme groups planar polar residues together. It tests for the discrimination power of hydrogen-bonding residues. "Simplified2"adds the aromatic subgroup of Sujatha *et al.* [40] while "simplified3"adds the histidine containing aromatic subgroup of Quiocho and Vyas [33]. The aromatic subgroup residues take part in hydrophobic and van der Waals interactions with the docking glucose.

Table 4.11: The different simplified residue subgrouping schemes.

Scheme Name	Subgroupings	Residues
Simplified1	Planar Polar	Asn, Asp, Gln, Glu, Arg
	Other	Remaining residues
Simplified2	Planar Polar	Asn, Asp, Gln, Glu, Arg
	Aromatic	Phe, Tyr, Trp
	Other	Remaining residues
Simplified3	Planar Polar	Asn, Asp, Gln, Glu, Arg
	Aromatic	Phe, Tyr, Trp, His
	Other	Remaining residues

The simplified schemes divide the amino acids into 2 or 3 subgroupings and, as such, are too simple to capture the biochemical complexity of the different

residues. Further detailed schemes were formulated, as shown in Table 4.10, that incorporate previous research findings.

4.2.2 Binding Site Representation

Our implementation views a binding site as a set of concentric spherical layers centered at the hexose pyranose centroid. This section describes the binding site representation.

4.2.2.1 Concentric Layers Sphere

We represent the binding site as a sphere centered at the ligand, as portrayed in Figure 4.2. The sphere is subdivided into concentric shells, as suggested by Bagley and Altman [3].

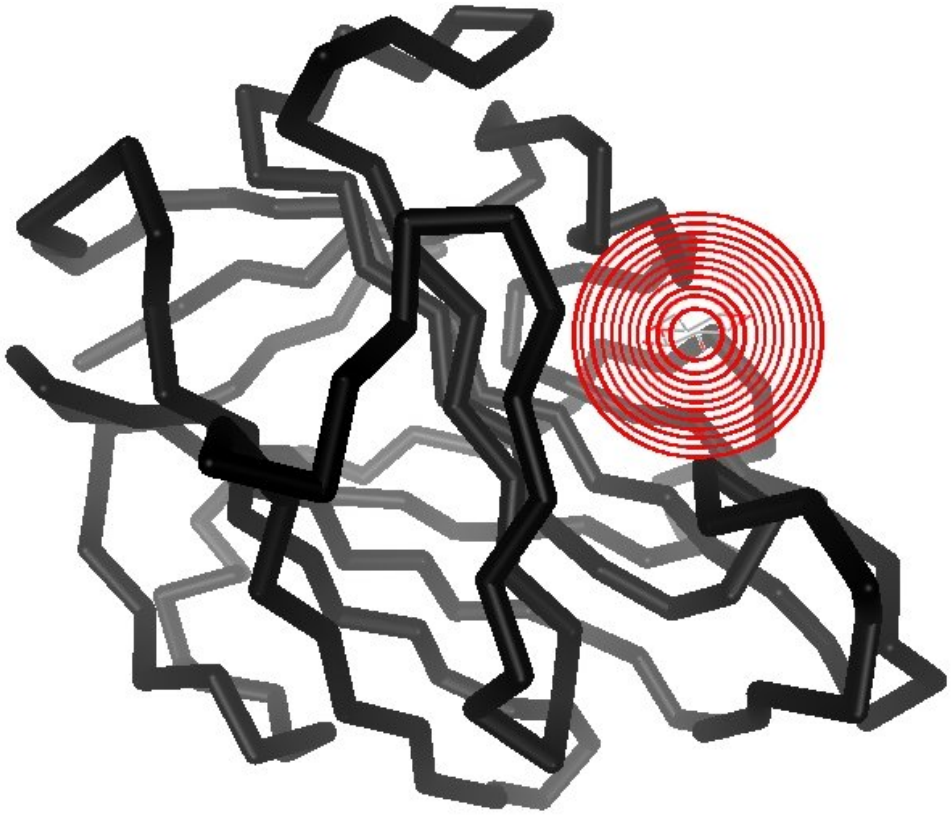


Figure 4.2: Glucose bound to a hydrolase, PDB entry 1I8A. The glucose pyranose ring's centroid is the center of the binding site's concentric spherical layers.

The algorithm obtains the spatial distribution of the different atoms in a spherical region of a certain radius, and divides them into concentric layers (as portrayed in Figure 4.2). For each layer, the program samples all the atoms and residues contained in that layer, Algorithm 4.1 then samples all the properties of each atom and creates a feature vector for every layer. As such, each feature has a measure in each concentric layer.

Algorithm 4.1 Compute layer feature vector

```
procedure computeLayerFeatureVector(layer)
    N ← layer.getAtomNumber()
    P ← layer.getPropertiesNumber()
    for i ← 1 to N do
        atom ← layer.getAtom(i)
        for j ← 1 to P do
            layerVector[j] ← layerVector[j] + atom.property(j)
        end for
    end for
    return layerVector
end procedure
```

The binding site feature vector is the concatenation of the ordered layer feature vectors (see Algorithm 4.2). The radius of the sphere itself and the width of the shells are parameters to be optimized.

Algorithm 4.2 Compute sphere feature vector

```
L ← sphere.getLayerNumber()
for i ← 1 to L do
    layer ← sphere.getLayer(i)
    sphereDoubleVector[i] ← computeLayerFeatureVector(layer)
end for
return sphereDoubleVector
```

4.2.2.2 Sphere Center

The center of the glucose-binding site is the center of the glucose pyranose ring and is computed as the centroid of the coordinates of the ring's 6 atoms (see Figures 2.6 and 4.2). The non-hexose binding-sites center is the cavity center, or

the ligand central point. For the non-binding site negatives, the center is a random point within the protein or what appears to be the center of the groove. The centers of the different positive and negative samples were listed previously.

4.2.3 SVM and kNN Classifiers

SVM and kNN are pattern recognition techniques. The classifier involves a learning and a recognition phase. Each instance in the training set contains one class label (true if it is a glucose-binding site, false if not) and a vector of features.

The training phase consists of the following two steps (refer to Section 3.1.3):

1. The input to the SVM classifier is a vector of numbers. Hence data preprocessing is required to convert categorical attributes to numerical ones.

2. Given a training set of instance-label pairs $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$ where $\mathbf{x}_i \in \mathbb{R}^q$ and $y \in \{1, -1\}^n$, the SVM constructs a hyperplane separating the positive examples from the negative ones in the space representation. To avoid overfitting, SVM chooses the Optimal Separating Hyperplane that maximizes the margin of error.

The feature vectors for each binding site are the input to the classifier. During SVM's learning phase, the system processes positive examples—feature vectors that characterize known hexose binding sites. It also processes negative examples—non-binding sites and sites that do not bind hexoses.

To achieve a meaningful validation of our experiments, we use a leave-one-out cross-validation method. The classifier is trained n times, each time with a different input sample held out as a validation set.

The SVM results are compared to a kNN classifier results. Both classifiers operate on the same feature vectors and rely on a leave-one-out cross-validation process.

Statistical analysis is performed using the "R" statistical computing environment [34]. Simple kNN runs the "class" library [45]. Distance-weighted kNN is computed by the "kknn" library [19]. For SVM classification, we use the R LIBSVM implementation [11, 27]. of package "e1071" [15].

4.2.4 Error Estimation

We estimate the classifier performance by a leave-one-out cross-validation method. This method requires large computational power but derives an unbiased error estimate [24]. The reported error is the classification error rate and is an estimate of the classifier generalization error.

Let $\mathcal{P}$ and $\mathcal{N}$ be the sets of positive and negative samples, respectively. Let $\hat{\mathcal{P}}$ and $\hat{\mathcal{N}}$ be the sets of predicted positive and negative samples, respectively. As Equation 4.1 states, we define P and N to be the number of positive and negative samples respectively; and $\hat{P}$ and $\hat{N}$ to be the number of predicted positive and negative samples respectively.

$$\begin{array}{ll} P = |\mathcal{P}| & \hat{P} = |\hat{\mathcal{P}}| \\ & \text{and} \\ N = |\mathcal{N}| & \hat{N} = |\hat{\mathcal{N}}| \end{array} \tag{4.1}$$

Let TP be the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. Their corresponding definitions are:

$$\begin{array}{ll} TP = |\hat{\mathcal{P}} \cap \mathcal{P}| & FP = |\hat{\mathcal{P}} \cap \mathcal{N}| \\ & \text{and} \\ TN = |\hat{\mathcal{N}} \cap \mathcal{N}| & FN = |\hat{\mathcal{N}} \cap \mathcal{P}| \end{array} \tag{4.2}$$

The basic performance measure estimates the classification error rate as:

$$error = \frac{FP + FN}{P + N} \tag{4.3}$$

Other measures include accuracy, precision, sensitivity and specificity. They are estimated as:

$$\begin{aligned} accuracy &= \frac{TP+TN}{P+N} & sensitivity &= \frac{TP}{P} \\ precision &= \frac{TP}{TP+FP} & specificity &= \frac{TN}{N} \end{aligned} \tag{4.4}$$

Accuracy is the rate of correct classification, precision reflects the ability to reject false positives, sensitivity to detect true positives, and specificity to reject true negatives.

Error estimation is performed in the "R" statistical computing environment [34] using the ipred package [32].

4.2.5 Modeling Glucose-Binding Sites

This process intends to identify the key features of the glucose-binding sites. Of all the different possible biochemical and geometrical features of a binding site, only some are essential for correct classification. In addition, recognizing these features as part of a dimension-reduction step will improve the efficiency of our method [24].

The use of Random Forests as a feature selection tool is ideal in our case. The use of RF for feature selection, coupled with SVM for classification, outperforms SVM alone [12]. In addition, the number of features we are monitoring is larger than the number of PDB entries.

Feature selection is performed in the "R" statistical computing environment [34] using the varSelRF package [14].

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Introduction

Many biochemical and spatial factors influence glucose binding. In addition, classifiers have multiple parameters to tune. Trying all possible property and parameter combinations is computationally prohibitive. For our set of experiments and in order to narrow down the grid of possible solutions, we fixed some of the parameters.

The pyranose ring radius is 1.5 Å, the farthest glucose atom (O6) is 3.5 Å away, the molecular interactions are significant to a range of 7 Å as suggested by Bobadilla *et al.* [6]. Therefore, the radius of the interaction is fixed at 10 Å. The first layer width is fixed to 3 Å while the width of the subsequent 7 layers is 1 Å each.

Our first testing data is made of 55 samples. The positive set is the 29 "distinct" α- and β-D-Glucopyranoses from Table 4.3. The negative set consists of the 26 entries of "set 1" from Table 4.5. The data is scaled and centered before being fed to the classifiers.

For testing the SVM classifier, we use the nonlinear soft margin implementation (see Equation 3.25) together with the RBF kernel (see Section 3.1.3.2), as suggested by Hsu *et al.* [22]. The parameters C and κ are incremented exponentially:

$$C = \{2^{-4}, 2^{-2}, \ldots, 2^{14}\} \tag{5.1}$$

$$\kappa = \{2^{-14}, 2^{-12}, \ldots, 2^{2}\} \tag{5.2}$$

The smaller the number of support vectors, the better the classifier's generalization.

For the simple kNN approach, $k = \sqrt{n}$ is a good approximation of k [16] and should be odd to break ties. Consequently, the k values used are:

$$k = \{1, 3, 5, 7, 9, 11\} \tag{5.3}$$

For distance weighted kNN, we increase k to include all training samples, and k acquires even values. Weighted k values are:

$$k = \{1, 2, \ldots, n-1\} \tag{5.4}$$

where the neighbors of one sample are the remaining $n-1$ samples. Regarding the distance metric, we use the Minkowski distance L_j (see Equation 3.8), where

$$j = \{1, 2, 3, 4, 5, 10, 20\} \tag{5.5}$$

Finally, we test all the kernel types listed in Table 3.1.

5.2 Learning Phase

We conduct different experiments during the learning phase, in order to select the best classification parameters. We first run experiments under the restrictions stated above, analyze results, and proceed to explore other paths.

5.2.1 Primary Findings

Table 5.1 presents the first experimental results. The atomic properties considered are: charge, hydrogen bonds and hydrophobicity. In order to check their relative discriminative powers, we considered each one of these properties alone, together with all their possible combinations. The percentages are the minimal classification error rates achieved by each classifier. We do not report simple kNN results for all experiments. The residue scheme used is "detailed1". SV stands for

support vectors. We report the percentage of support vectors over the total number of samples, 55.

Table 5.1: Misclassification rates using the "detailed1" residue scheme. SV reports the percentage of support vectors.

Properties	kNN	Weighted kNN	SVM	SV
Charge	N/A	14.55%	14.55%	78.18%
H-Bond	N/A	21.82%	16.36%	92.73%
Hydrophobicity	N/A	21.82%	20.00%	92.73%
Charge + H-Bond	N/A	14.55%	14.55%	89.09%
Charge + Hydro	N/A	12.73%	14.55%	47.27%
H-Bond + Hydro	34.55%	21.82%	18.18%	100%
Charge + H-Bond + Hydro	37.27%	16.36%	16.36%	60.00%
Residue (detailed1 scheme)	27.27%	16.36%	14.55%	81.82%
Residue + Charge	N/A	18.18%	10.91%	100%
Residue + H-Bond	N/A	16.36%	10.91%	94.55%
Residue + Hydro	N/A	16.36%	10.91%	90.91%
Residue + Charge + H-Bond	N/A	16.36%	10.91%	100%
Residue + Charge + Hydro	N/A	20.00%	09.09%	100%
Residue + H-Bond + Hydro	23.64%	18.18%	10.91%	98.18%
Residue + Charge + H-Bond + Hydro	30.91%	18.18%	09.09%	100%

5.2.1.1 Atomic Properties Comparison

Based on the Table 5.1 results, we notice some interesting trends in the data. Considering the atomic single properties, charge outperforms hydrophobicity and hydrogen bond. This is surprising, since we were expecting hydrogen bond to be more important for glucose docking than charge. In addition, charge and hydrophobicity are two closely linked properties. Nevertheless, they diverge in their classification accuracy.

Charge, linked with another property, yields results similar to or slightly better than charge alone. As for the combination of the three atomic factors, it yields moderate results relative to charge (16.36% and 14.55% respectively). "Charge + hydro" is the best atomic combination, surprisingly excluding hydrogen bond. It

gives a better result for kNN than SVM (SVM may need more fine tuning), and the support vector percentage is the lowest, 47.27%, implying better generalization and scalability.

When residue features are added, SVM performs better while weighted kNN deteriorates. The addition of the residue property limits the differences between the multiple atomic property combinations. They differ by just one sample misclassification. "Charge + hydro", combined with the "residue" property, gives the best result (9.09%). The addition of the hydrogen bond property does not improve the result.

The residue scheme to use, the residue-atomic property correlation, and the weak discrimination power of hydrogen bond, are fields we will investigate later on.

5.2.1.2 Comparison of Classifiers

Simple kNN performs very poorly, as can be seen from Table 5.1 and Figure 5.1. Weighted kNN consistently outperforms the simple version. We discard simple kNN from further experiments.

Weighted kNN and SVM give similar results when atomic properties are used, but SVM clearly outperforms kNN when a combination of residue and atomic properties are included. In the latter case, SVM overfits the data: most samples are support vectors.

5.2.2 Water and Ions Inclusion

In our first experiments, we include water molecules and sulfate, phosphate, calcium, magnesium and zinc ions in the atomic feature vector computation (listed in Table 4.9). Quiocho and Vyas [33] pinpoint the role of ordered water molecules and metal ions in determining substrate specificity and affinity. To test for this hypothesis, we run experiments on the best kNN and SVM property combinations of Table 5.1, "charge + hydro" and "detailed1 residue + charge + hydro" respectively.

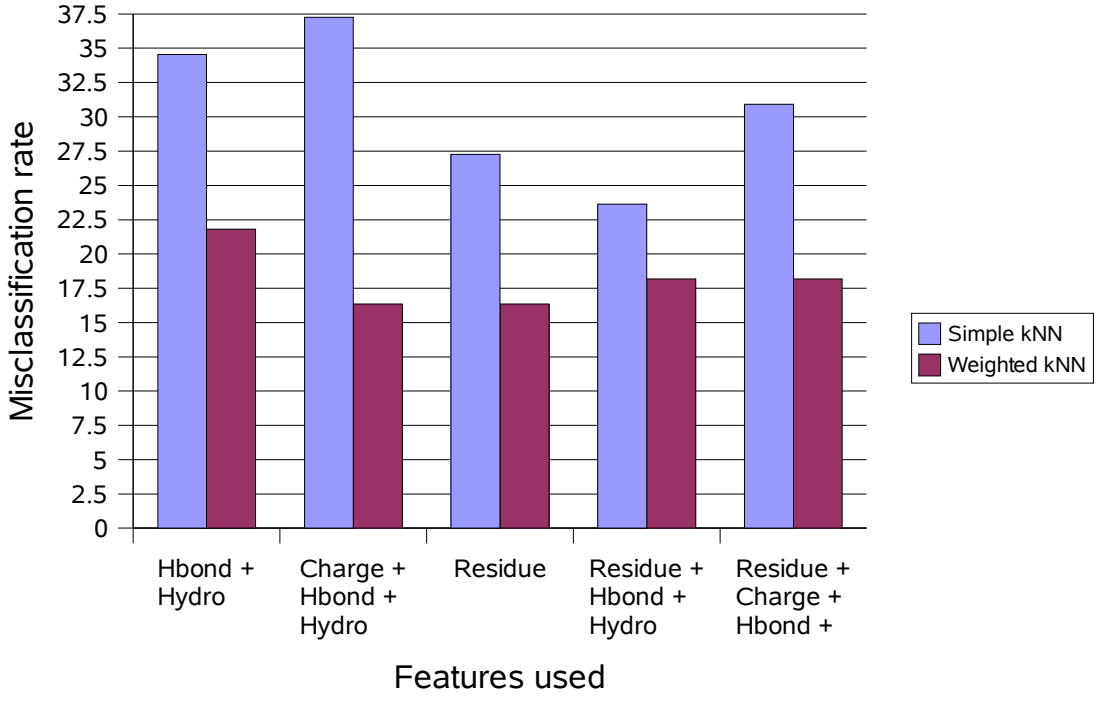


Figure 5.1: Simple *vs.* weighted kNN plot.

We first discard water molecules from the atomic feature vector computation. We then exclude the ions, and finally we discard both the water molecules and the aforementioned ions. The results are tabulated in Table 5.2.

Table 5.2: Testing the importance of water and ions to glucose specificity.

Properties	Weighted kNN	SVM	SV
Include Water and Ions			
Charge + Hydro	12.73%	14.55%	47.27%
Residue + Charge + Hydro	20.00%	09.09%	100%
Discard Water			
Charge + Hydro	14.55%	16.36%	83.64%
Residue + Charge + Hydro	20.00%	10.91%	100%
Discard Ions			
Charge + Hydro	16.36%	18.18%	83.64%
Residue + Charge + Hydro	20.00%	10.91%	100%
Discard Water and Ions			
Charge + Hydro	14.55%	16.36%	52.73%
Residue + Charge + Hydro	20.00%	12.73%	100%

It is clear that the inclusion of water and ions in the computation yields similar or better results than the exclusion of either one. This observation applies to both property sets and to the kNN, SVM and the support vector (SV) measures. Our results confirm the findings of Quiocho and Vyas [33].

5.2.3 Residue Schemes Analysis

As described earlier in Section 4.2.1.2, we have devised multiple residue schemes to test for different hypothesis. The "simplified " schemes test for the effect of certain precise subgroupings while the "detailed " schemes incorporate further biochemical knowledge.

5.2.3.1 Residue Features Comparison

We first check the discrimination power of the residue property taken alone. Table 5.3 lists the results.

Table 5.3: Comparison of the different residue schemes.

Residue scheme	Weighted kNN	SVM	Support vectors
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

The detailed schemes, which incorporate a higher level of biochemical knowledge, slightly outperform the simplified ones while having a higher percentage of support vectors. The detailed schemes yield similar results to the atomic properties (compare with Table 5.1), suggesting that they incorporate in themselves, due to their high detail level, some of the atomic properties.

Adding histidine (His) to the aromatic subgroup gives slightly better results. "Simplified3", which incorporates histidine, improves the "simplified2" classification rate by one hit.

The aromatic residues do not seem to contribute to the classification accuracy. In fact, the "simplified2" scheme gives similar results to "simplified1" while adding the aromatic subgroup to the latter. Aromatic residues are the amino acids mostly involved in glucose hydrophobic stacking. The low discrimination capacity of the aromatic residues and of the hydrophobicity property might be connected.

5.2.3.2 Atomic and Residue Properties Combination

The combination of atomic and residue properties increases the number of features. In order to increase the samples to features ratio (n_i/q) while keeping equal positive and negative sets, we perform the atomic and residue properties combination experiments with 64 samples, by addition of the Table 4.6 "set 2" entries. The negative set now consists of 24 non-hexose binding sites and 11 non-binding sites. The positive set retains its 29 entries.

Combining atomic and residue property levels, the classification accuracy improves (see Table 5.4). All SVM classification errors are less than 13%. For all property combinations used, SVM outperforms kNN.

Table 5.4: Comparison of atom and residue properties.

Properties and Schemes	Weighted kNN	SVM	SV
Simplified1 + Charge + Hydro	14.06%	12.50%	56.25%
Simplified2 + Charge + Hydro	15.63%	07.81%	96.88%
Simplified3 + Charge + Hydro	14.06%	12.50%	60.94%
Detailed1 + Charge + Hydro	17.19%	10.94%	98.44%
Detailed2 + Charge + Hydro	17.19%	09.38%	57.81%
Simplified1 + Charge + H-Bond + Hydro	15.63%	12.50%	62.50%
Simplified2 + Charge + H-Bond + Hydro	15.63%	09.38%	60.94%
Simplified3 + Charge + H-Bond + Hydro	15.63%	12.50%	65.63%
Detailed1 + Charge + H-Bond + Hydro	15.63%	07.81%	89.06%
Detailed2 + Charge + H-Bond + Hydro	14.06%	07.81%	53.13%

The "detailed2" scheme performs equally to or better than "detailed1" (refer to Figure 5.2). In addition, "detailed2" has about 40% fewer support vectors and has a shorter feature vector, since we merged the aromatic and histidine subgroups

(refer to Table 4.10). A short feature vector and a low number of support vectors are signs of good scalability and generalization. These results support Quiocho and Vyas [33] outcome that histidine and aromatic residues are chemically similar regarding hexose docking.

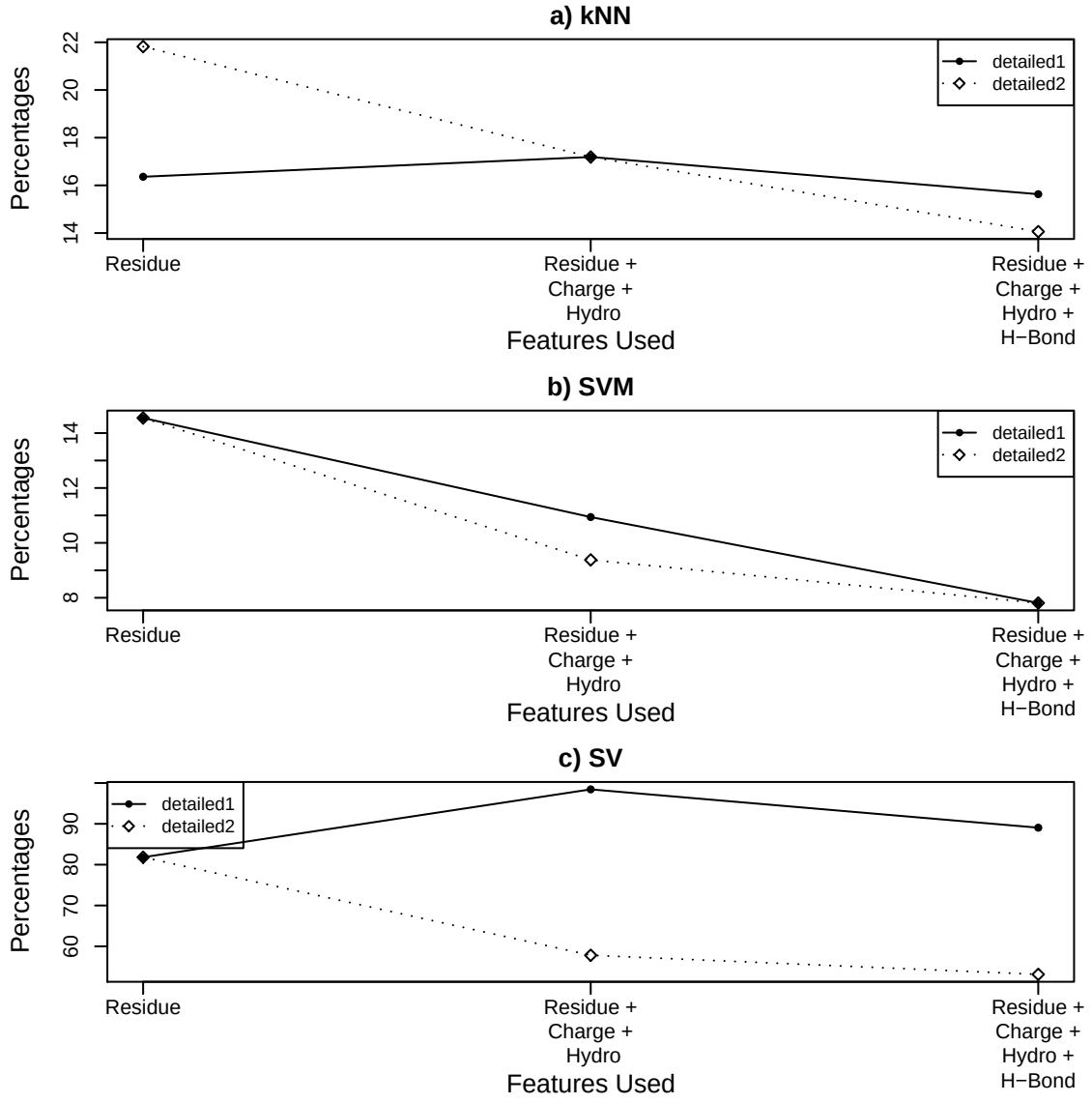


Figure 5.2: Comparison of detailed1 and detailed2 schemes.

"Simplified3"incorporates histidine in the aromatic subgroup; it surpasses "simplified2" when only residue features are used. When we couple residue and atomic properties, "simplified2" yields the best results (refer to Figure 5.3(b)). The best result, 7.81% error rate using "residue + charge + hydro", is coupled with

a very high support vector percentage of 96.88% (see Figure 5.3(c)), which implies overfitting. Nevertheless, the use of "residue + charge + hydro + hbond" properties favors "simplified2" both for SVM error rate (9.38% *vs.* 12.50%) and support vector numbers (60.94% *vs.* 65.63%).

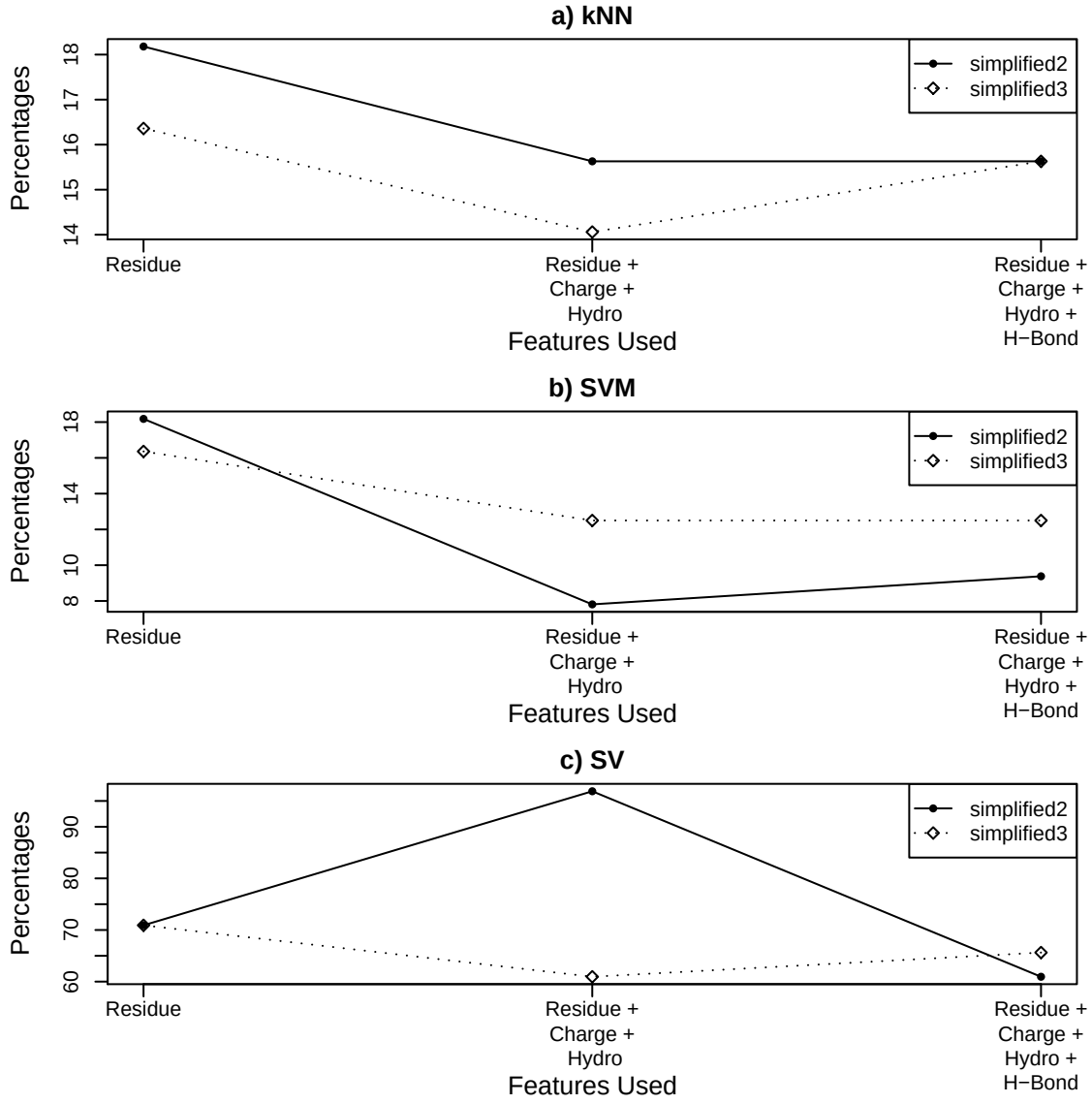


Figure 5.3: Comparison of simplified2 and simplified3 schemes.

The relation of histidine (His) to glucose docking is an issue that requires further wet lab studies. Our results are not clear cut toward this issue. Sujatha *et al.* [40] note the absence of the docking aromatic residue in some glucose-binding sites and its crucial importance for galactose docking. The inconsistent importance

of aromatic residues toward glucose docking, coupled with the common hydrophobic properties of histidine and the other aromatic residues, may explain the different outcomes of adding histidine to the simplified and detailed residue schemes.

5.2.4 Atomic Properties Analysis

This section analyzes the status of each atomic property. These properties are charge, hydrogen bonding and hydrophobicity.

5.2.4.1 The Hydrogen Bond Atomic Property

The hydrogen bond performed poorly as a discriminating property in the first set of experiments (Section 5.2.1). We know that glucose-binding relies heavily on hydrogen bonds (refer to Section 2.3.3). This section further investigates the weak discrimination power of the hydrogen bond property.

In the first set of experiments (Section 5.2.1 and Table 5.1), we use a negative training set composed mainly of binding sites that do not bind hexoses. The negative input set comprises 18 sites that do not bind hexoses and only 8 non-binding sites listed in Table 4.5.

The non-hexose binding sites are actual binding sites. Most protein-ligand binding requires the establishment of hydrogen bonds between the protein and the ligand. Therefore atoms and residues capable of establishing hydrogen bonds abound in most binding-site grooves. This means that the non-hexose binding sites, like the glucose binding sites, contain many hydrogen-bonding atoms. Hence the hydrogen bond property can not discriminate between such sets!

To test this hypothesis, we build a classifier using an exclusively non-binding sites negative set. Non-binding sites are not biased regarding hydrogen bonding atoms, and such a classifier should mainly discriminate according to the hydrogen bond property. We use the same 29 positives from Table 4.3 as in the previous experiments. We use the non-binding site negatives of sets 1, 2 and 3 in Tables 4.5, 4.6

and 4.7. The negative set adds up to 28 samples. Our training sample totals to 57 inputs. We perform the runs on each of the atomic properties, namely charge, hydrogen bond and hydrophobicity, and on all their possible combinations. Table 5.5 lists the results.

Table 5.5: Classifier training using an exclusively non-binding sites negative set.

Properties	Weighted kNN	SVM	SV
Charge	05.26%	05.26%	73.68%
H-Bond	03.51%	03.51%	61.40%
Hydrophobicity	05.26%	05.26%	68.42%
Charge + H-Bond	07.02%	01.75%	66.67%
Charge + Hydro	08.77%	03.51%	54.39%
H-Bond + Hydro	03.51%	03.51%	71.93%
Charge + H-Bond + Hydro	08.77%	03.51%	68.42%

As expected, the hydrogen bond feature outperforms charge and hydrophobicity, for kNN and SVM classification accuracy and for the support vector percentage. The "charge + hbond" combination yields the best result, misclassifying only one sample to reach an error of 1.75%.

A comparison between Tables 5.1 and 5.5 reveals a great improvement in classification accuracy. This is due to the choice of the negative samples. The former negative samples, composed mainly of non-hexose binding sites, are much more biochemically similar to the glucose-binding sites than the latter non-site negatives.

Both findings, the hydrogen bond high discrimination capacity and the sharp drop in the classification error, confirm our suggestion. Hydrogen bond is a key property in glucose-binding sites, but it is not a good discrimination criteria vis-a-vis other binding sites.

5.2.4.2 The Hydrophobicity Atomic Property

From the primary results of Section 5.2.1, we inferred that the "charge + hydro" pair yields good classification results, either alone or coupled with the residue property.

Hydrophobicity alone is the worst discriminator, with an SVM error rate of 20.00%. Coupled to charge or residue properties, it significantly improves to 14.55% and 10.91%, with an accompanying decrease in support vector percentages (refer to Figure 5.4).

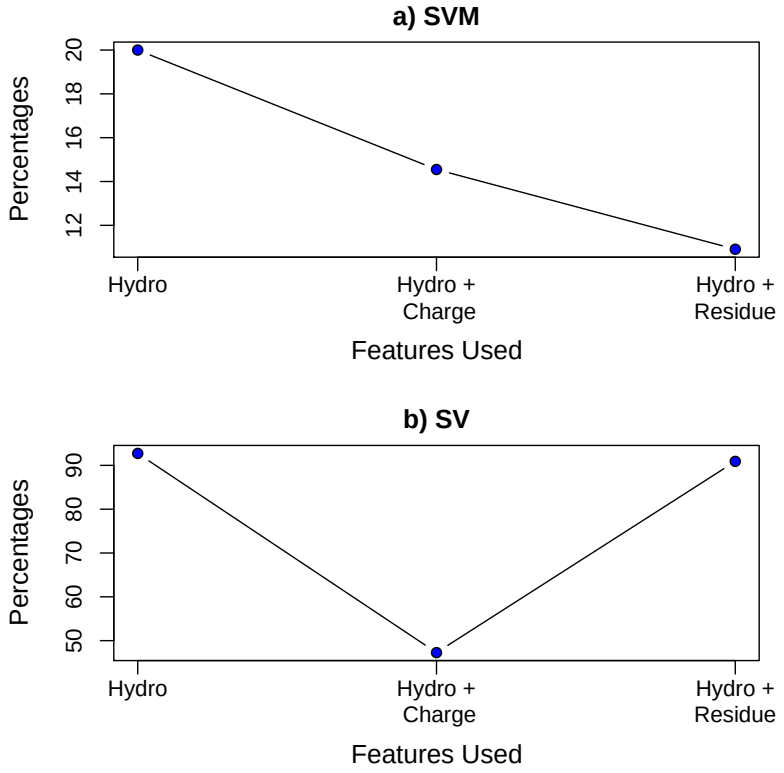


Figure 5.4: Analysis of the hydrophobicity property.

As Taroni *et al.* [42] report, the sugar binding sites are neither hydrophobic nor hydrophilic. Hexoses exhibit a dual hydrophobic-hydrophilic nature, both antagonist properties are involved in hexose docking. This fact may explain why hydrophobicity alone yields bad discrimination results. Low error rates result from coupling hydrophobicity with another discriminating property. This means that hydrophobicity fine tunes the classifier, offering a secondary, more subtle, discrimination axis.

5.2.4.3 The Charge Atomic Property

Charge is the property with the highest discrimination power, amidst a negative set composed mainly of non-hexose binding sites. Most binding sites are rich in hydrogen bonding atoms and residues, thus subtle properties, like charge, permit to discriminate between them.

5.3 Feature Selection

RF feature selection depends on some parameters. We use the settings suggested by Díaz-Uriarte and de Andrés [14]. The number of features to select at each split is $mtry = \sqrt{q}$ where q is the feature number. The number of trees to grow is $ntree = 5000$. The minimum size of the tree terminal nodes is $nodesize = 1$. The fraction of features to drop at each iteration is $fraction.dropped = 0.2$. Finally, we choose the solution with the smallest number of features whose error rate is within 1 standard errors of the minimum error rate of the forest ($c.sd = 1$). This practice, called the "1 standard error rule", is common in classification trees literature [14]. We perform feature selection on a training set of 55 samples. The positive set is the 29 α- and β-D-Glucopyranoses from Table 4.3. The negative input is the 26 entries of "set 1" from Table 4.5. The data is scaled and centered before being fed to the feature selector.

5.3.1 Atomic Properties

We plot the different features of each atomic property according to their RF importance score. Table 4.9 lists the different atomic properties features. We select the best feature combination based on its classification accuracy. The selected feature subset has the highest *information gain*, which is a measure of its classification accuracy [28]. Table 5.6 presents the results, where false RF means classifying the data using all features, and true RF refers to selecting features prior to classification. We use both SVM and kNN classifiers and compute the feature number, sensitivity

and specificity percentages for the best classifier. As expected, the classification error for both classifiers and the support vector percentage decrease.

Table 5.6: Comparison of classifiers' performance on atomic data with and without RF feature selection.

Property	RF	Feature Number	kNN Error	SVM Error	Sensitivity	Specificity	SV
Charge	false	24	14.55%	14.55%	96.55%	73.08%	78.18%
	true	6	09.09%	05.45%	93.10%	96.15%	41.82%
H-Bond	false	16	21.82%	16.36%	86.21%	80.77%	92.73%
	true	5	09.09%	07.27%	96.55%	88.46%	16.36%
Hydro	false	24	21.82%	20.00%	79.31%	80.77%	92.73%
	true	5	09.09%	10.91%	96.55%	84.62%	34.55%

5.3.1.1 Charge Feature Selection

In accordance to previous findings, the charge atomic property possesses the best discrimination capacity (refer to Table 5.6). After feature selection, SVM reaches a very low error rate of 5.45%, coupled with a good generalization capability reflected by 41.82% of support vectors. Without feature selection, the charge classifier is biased towards detecting true positives (high sensitivity and low specificity). When RF is used, both sensitivity and specificity percentages are high and close to each other.

The 6 features selected to achieve the lowest error are: layer 1 neutral atoms number, layer 2 neutral atoms number, layer 3 negative atoms number, layer 4 negative atoms number, layer 5 negative atoms number and layer 8 negative atoms number. Figure 5.5 plots their importance measure as returned by RF.

These findings suggest a negatively charged glucose binding site. The negative atoms are preponderant from layer 3 and above. It is important to note that the layer 3 negative atoms feature is the most discriminating feature. Layer 3, together with layers 1 and 2, is in direct contact with the glucose molecule. Layers 1 and 2 overlap with the glucose molecule's own space. These two layers tend to have fewer

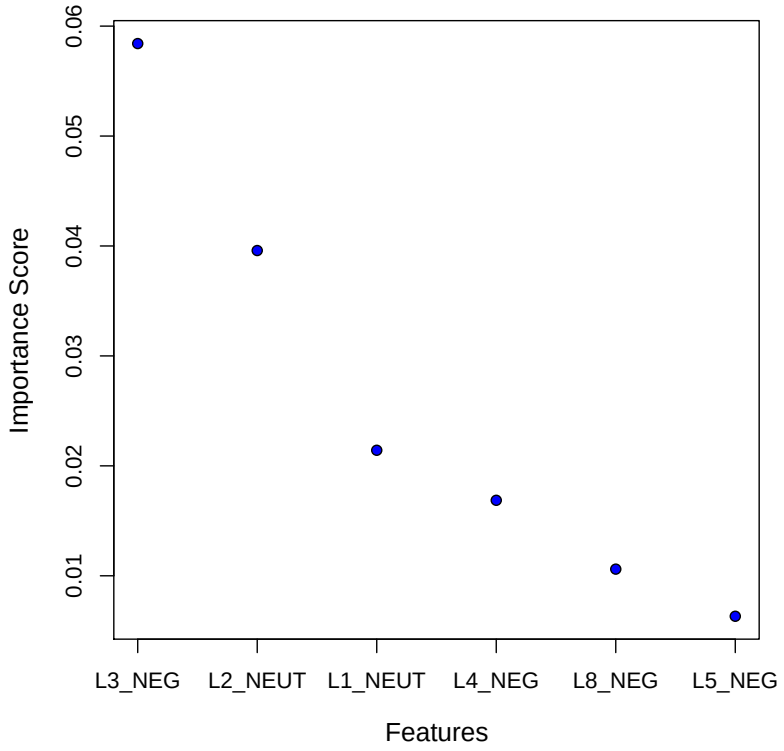


Figure 5.5: Importance of charge features according to RF; NEUT stands for neutral and NEG for negative.

atoms in the positive set than the negative set. Since the neutral atoms are much more abundant in a protein than the charged ones, a high neutral atom number in layers 1 and 2 implies a negative entry.

5.3.1.2 Hydrogen Bond Feature Selection

After feature selection, the hydrogen bond atomic property returns good results, as shown in Table 5.6. The SVM error drops to 7.27% and the number of support vectors reaches a record 9/55. Sensitivity and specificity percentages are also high with a slight bias toward detecting true positives.

Feature selection using RF returned layers 1, 2 and 3 hydrogen-bonding atom numbers, and layers 2 and 7 non-hydrogen-bonding atom numbers. Figure 5.6 plots their importance measure as returned by RF.

All the positive and most of the negative training entries are actual binding

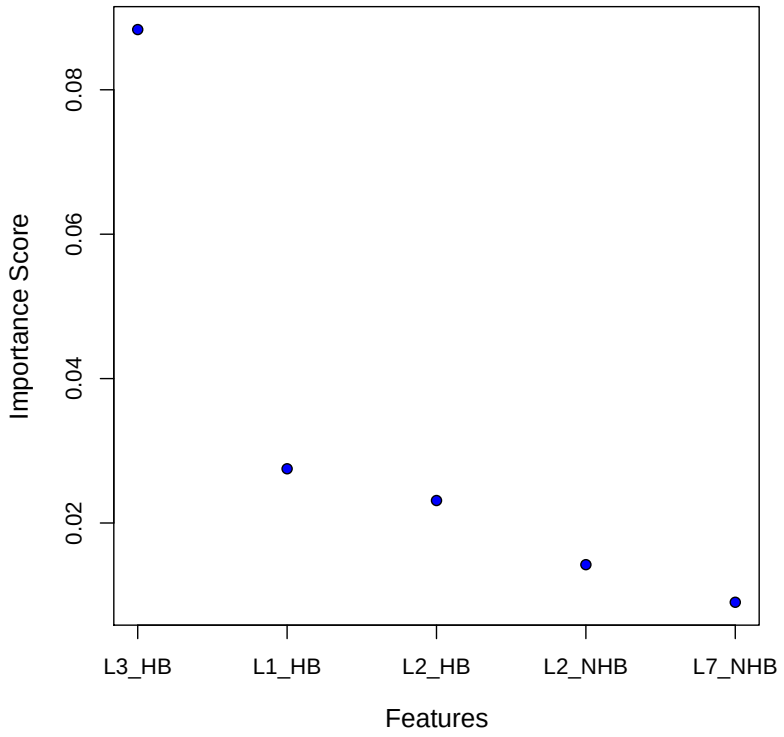


Figure 5.6: Importance of hydrogen bond features according to RF; HB stands for hydrogen-bonding and NHB for non-hydrogen-bonding.

sites. Thus both sets heavily rely on hydrogen bonding. This is why the selected features of highest importance are hydrogen-bonding atom numbers from the layers adjacent to the ligand. The positive and negative sets differ by their hydrogen-bonding atoms distribution. Glucose-binding entries have a high hydrogen-bonding atoms concentration at layer 3, which explains the high importance score (and discrimination capacity) of layer 3 hydrogen-bonding feature. High hydrogen-bonding atoms concentration at layers 1 and 2 specify negative entries.

5.3.1.3 Hydrophobicity Feature Selection

Unlike the previous two properties, kNN slightly outperforms SVM after feature selection (refer to Table 5.6). With an error rate of 9.09%, hydrophobicity remains the worst discriminating atomic property. The kNN classifier is biased in favor of detecting true positives.

The selected features are layers 1, 2 and 3 hydrophilic atom numbers, layer 2 neutral atom numbers and layer 7 hydrophobic atom numbers. Figure 5.7 plots their importance measure as returned by RF.

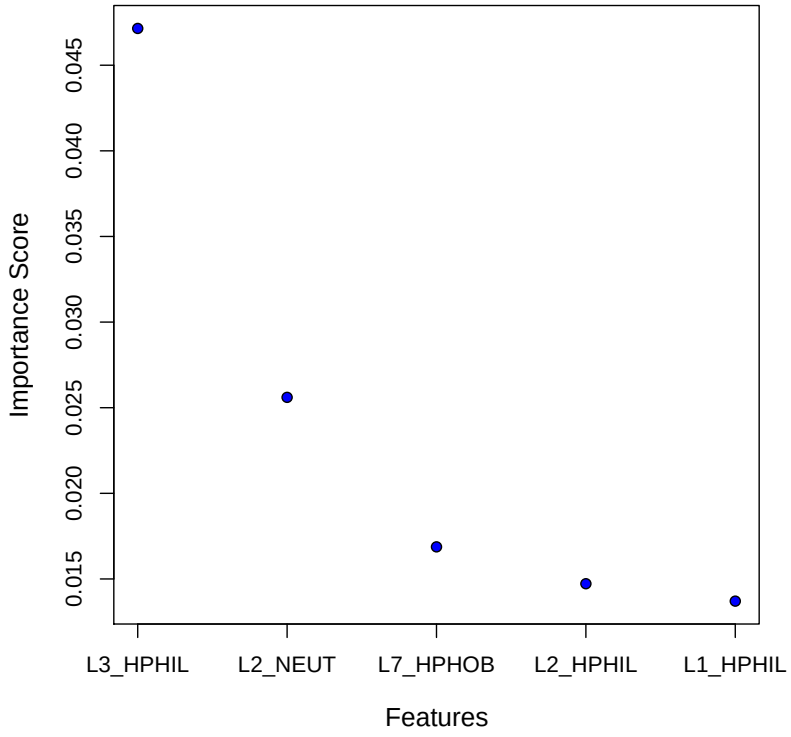


Figure 5.7: Importance of hydrophobicity features according to RF; HPHIL stands for hydrophilic, NEUT for neutral and HPHOB for hydrophobic.

The positive set has a high concentration of hydrophilic atoms in layer 3, which gives layer 3 hydrophilic feature a high discrimination capacity. The negative set has more neutral and hydrophilic atoms in layers 1 and 2. It is surprising to note that the hydrophobic atoms have a low importance score which reflects a low information gain. Even the reported hydrophobic feature of layer 7 seems to characterize negative entries. The aromatic hydrophobic atoms, which are involved in hexose's hydrophobic ring stacking, do not seem to correctly classify our data.

5.3.2 Residue Properties

For each residue scheme, we plot the different features according to their RF importance score. The residue features are tabulated in Tables 4.10 and 4.11. We select the best feature combination based on its classification accuracy. Table 5.7 presents the results. The residue scheme feature selection does not perform as well as the atomic properties. A comparison with Table 5.6 shows that charge and hydrogen bond properties outperform the residue schemes. Both classifiers accuracies improves with feature selection, as well as the generalization capability. The sensitivity and specificity values become more balanced.

Table 5.7: Comparison of classifiers' performance on residue data with and without RF feature selection.

Property	RF	Feature Number	kNN Error	SVM Error	Sensitivity	Specificity	SV
Detailed1	false	48	16.36%	14.55%	89.66%	80.77%	81.82%
	true	19	09.09%	12.73%	93.10%	88.46%	56.36%
Detailed2	false	40	21.82%	14.55%	96.55%	73.08%	81.82%
	true	3	10.91%	12.73%	89.66%	88.46%	56.36%
Simplified1	false	16	16.36%	18.18%	96.55%	69.23%	76.36%
	true	6	12.73%	12.73%	93.10%	80.77%	67.27%
Simplified2	false	24	18.18%	18.18%	86.21%	76.92%	70.91%
	true	4	10.91%	10.91%	96.55%	80.77%	54.55%
Simplified3	false	24	16.36%	16.36%	96.55%	69.23%	70.91%
	true	6	12.73%	12.73%	96.55%	76.92%	74.55%

5.3.2.1 Simplified Schemes Feature Selection

For the three simplified schemes of Table 5.7, kNN and SVM yield exactly the same error, sensitivity and specificity. All the simplified scheme classifiers are biased towards detecting true positives.

The three schemes yield similar feature selection results. "Simplified2" includes layers 5 and 7 planar polar residues, and layers 1 and 2 remaining residues.

"Simplified1" and "simplified3" add layers 1 and 6 planar polar residues. Figure 5.8 plots the feature importance of "simplified3" residue scheme.

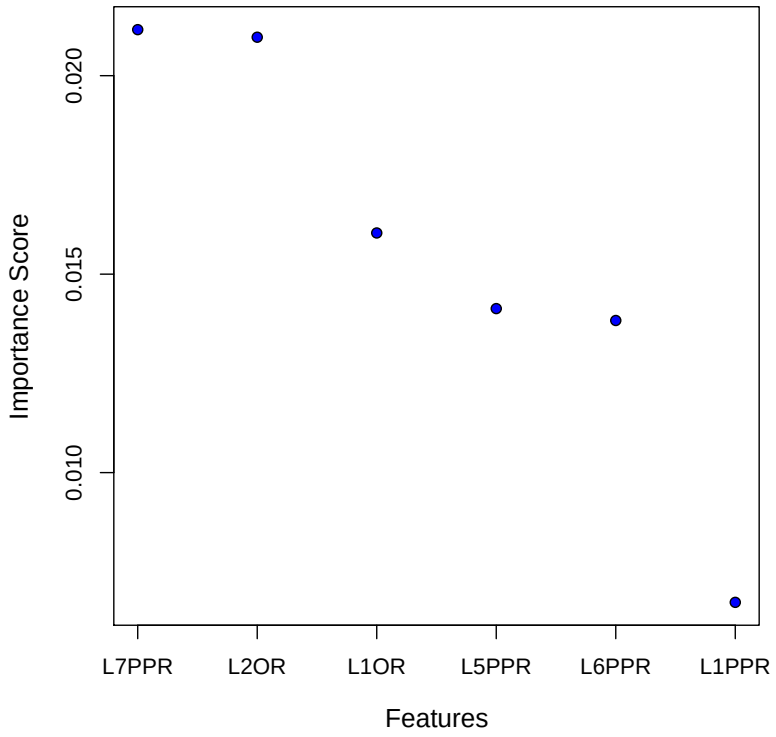


Figure 5.8: Importance of "simplified3" features according to RF; PPR stands for planar polar residues and OR for other residues.

The results show the primordial importance of planar polar residues in glucose binding sites, whereas their presence at layers 5 to 7 characterizes the positive set. It is worth to note the low information gain of the aromatic residues. In both "simplified2" and "simplified3" schemes, the aromatic residue features score low, reflecting a weak classification capacity.

5.3.2.2 Detailed Schemes Feature Selection

After feature selection, detailed residue schemes still perform better than their simplified homologues, while scoring worse than atomic properties (refer to Tables 5.6 and 5.7). kNN outperforms SVM for both schemes. Detailed residue classifiers incorporating feature selection have equal specificity and sensitivity scores.

"Detailed2" residue scheme feature selection yields layers 4, 6 and 8 carboxylate-bearing residues. As for "detailed1", it slightly outperforms "detailed2" to the expense of selecting 19 features. Nevertheless, layers 4 to 8 carboxylate-bearing residues are ranked first. Figure 5.9 plots the 5 most important features of "detailed1" residue scheme.

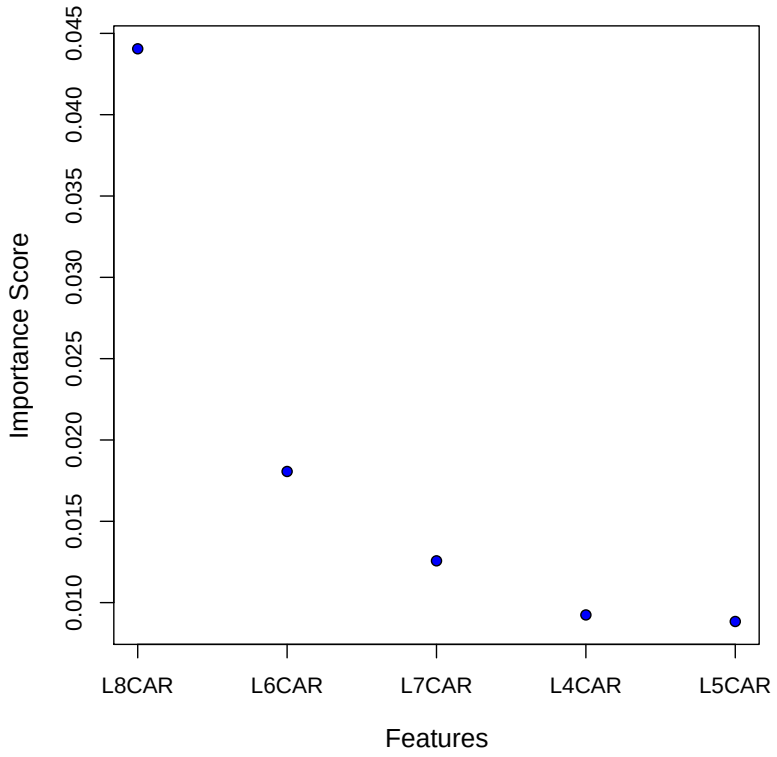


Figure 5.9: The 5 highest "detailed1" features importance measure as returned by RF; CAR stands for carboxylate-bearing residues.

The high information gain of carboxylate-bearing residues suggest a high density of the negatively charged glutamate (Glu) and aspartate (Asp) amino acids in the glucose binding sites. As we noted earlier for the simplified schemes, the aromatic residue features tend to bear low information gain values.

5.3.3 Combining Atomic and Residue Properties

From previous experiments, we found that combining atomic and residue properties improves the classifier's accuracy, where the "charge + hbond + hydro

+ detailed2" association yields the best classification results. We use the same 55 samples training set and perform feature selection for the above property group. Table 5.8 shows the results.

Table 5.8: Comparison of classifiers' performance on combined atomic and residue data with and without RF feature selection.

Property	RF	Feature Number	kNN Error	SVM Error	Sensitivity	Specificity	SV
Detailed2 + Charge	false	104	14.06%	07.81%	93.10%	91.43%	53.13%
+ H-Bond + Hydro	true	15	03.64%	03.64%	96.55%	96.15%	69.09%

Combining multiple properties improves the classifiers' accuracy. Different (although roughly similar) feature combinations return a classification error rate as low as 3.64%, the minimum achieved throughout this work. A distinctive combination of 15 features reaches the smallest error rate for both kNN and SVM. Table 5.8 refers to this outstanding combination, misclassifying just one positive and one negative entries. Table 5.9 lists the 15 features selected. It is important to note that the only residue subgroup present is the carboxylate-bearing residues, a subgroup common to both "detailed1" and "detailed2" schemes. Thus the selected feature combination is identical for both detailed schemes.

Figure 5.10 compares the selected features importance score. It shows the supremacy of layer 8 carboxylate residues, followed by layer 3 negative charge atoms, layer 2 neutral charge atoms, and layer 3 hydrogen bonding atoms. A bit further lie layer 3 hydrophilic atoms and layer 6 carboxylate residues. Finally, the remaining 9 features range close to each other.

The atomic and residue combination results confirm our previous findings, whereby the highest scoring features of the different single atomic and residue properties combine to yield the best discriminating feature subset. Their combination has a synergistic effect, dropping the error rate to a minimum of 03.64%.

Table 5.9: The feature combination achieving the minimal error for both kNN and SVM classifiers.

Property	Features	L1	L2	L3	L4	L5	L6	L7	L8
Charge	Negative			X					
	Neutral	X	X						
H-Bond	H-Bonding			X					
Hydrophobicity	Hydrophilic	X	X	X					
	Neutral		X						
	Hydrophobic							X	X
Residues	Neutral						X		
	Carboxylate					X	X		X
	Aliphatic			X					

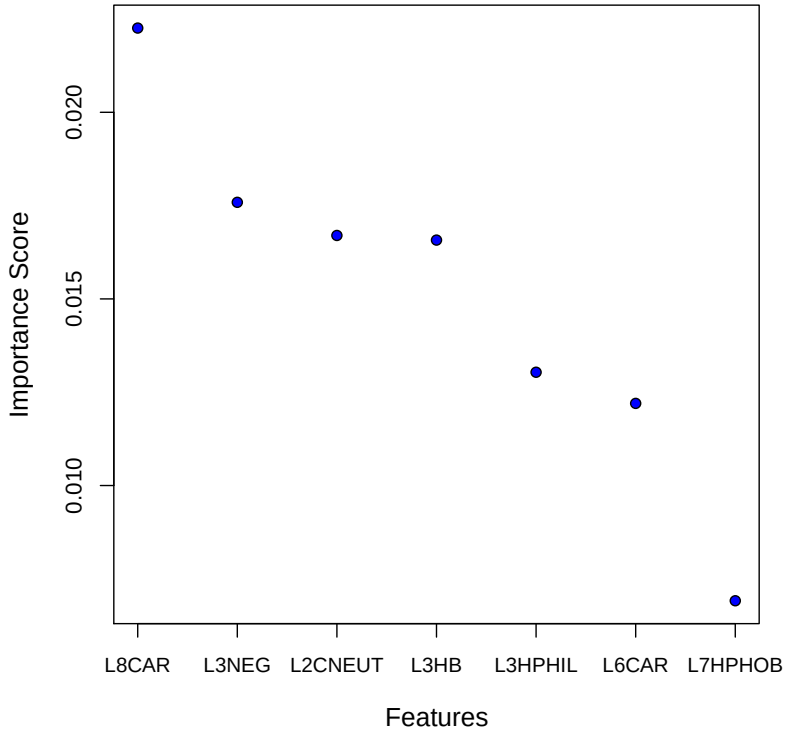


Figure 5.10: The 7 highest "charge + hbond + hydro + detailed" features importance measure as returned by RF. CAR stands for carboxylate-bearing residues, NEG for negative charge, CNEUT for neutral charge, HB for hydrogen-bonding, HPHIL for hydrophilic and HPHOB for hydrophobic.

5.3.4 Feature Selection Findings

Feature selection pinpoints the difference in spatial configuration between the positive and negative binding sites. Negative binding sites have a higher atomic presence in layers 1 and 2, both of which overlap with the bound glucose molecule's own space. Positive binding sites tend to have most of their discriminating atomic features in layer 3, in direct contact with the bound glucose.

Charge turns out to be the property having the highest information gain. Using selected charge features alone, SVM's error drops to 5.45. Feature selection confirms the learning phase findings about charge's classifying supremacy. RF shows that the glucose-binding site third layer has a relative negative charge. Hydrogen bond and hydrophobicity feature selection adds weight to this finding: layer 3 is rich with hydrogen-bonding and hydrophilic atoms. In fact, a negatively charged atom tends to be both hydrophilic and hydrogen bonding.

Charge feature selection actually reveals the relative negativity of layers 3 and above. This finding is confirmed by the high propensity of the negatively charged carboxylate-bearing residues in layers 4 to 8. It is important to note that both carboxylate-bearing residues, glutamate (Glu) and aspartate (Asp), are identified by Taroni *et al.* [42] to have a high sugar interface propensity level (refer to Table 2.3).

The simplified schemes feature selection discloses the relevance of planar polar residues in glucose specificity. Planar polar residues abound in layers 5 to 7. Carboxylate-bearing residues being planar polar residues, both residue scheme findings correlate. All planar polar residues have a high sugar interface propensity level (refer to Table 2.3).

Although glucose is reported to stack over a hydrophobic aromatic residue, the latter's features are removed by feature selection. None of the residue scheme attributes to the aromatic features a high information gain. This fact is also highlighted by the hydrophobicity property, where hydrophobic features do not seem to characterize glucose binding sites. The aromatic features absence might be explained

by their mode of action. A pyranose ring requires just one correctly placed aromatic residue to stack against [33, 40]. Since our algorithm computes and compares the total number of aromatic residues, it is unable to discover this one-correctly-placed relationship. In addition, Sujatha *et al.* [40] report the absence of the docking aromatic residue in some glucose-binding sites.

5.4 Testing Phase

In order to validate our classifiers, we test them using a separate testing set, not used previously in the training phase. We parse the PDB repository for non-redundant glucose binding sites, limiting the query to entries newer than October 2004. The generated positive testing set from Table 4.4 amounts to 7 entries. The negative set consists of the 15 entries testing data from Table 4.8. We use the best reduced feature representation of Table 5.9, and classify the validation data using SVM and kNN.

Table 5.10 shows the results. TP are the true positive sites that bind glucose. FP are the false positive sites that do not bind glucose but are labeled as glucose-binding by the classifier. TN are the true negatives that do not bind glucose. FN are the false negative sites that do bind glucose and are mislabeled by the classifier.

Table 5.10: Testing phase results.

Classifier	TP	FP	TN	FN	Error	Sensitivity	Specificity	Precision
SVM	6	0	15	1	04.55%	85.71%	100%	100%
kNN	6	2	13	1	13.64%	85.71%	86.67%	75.00%
COTRAN	94	27	633	12	05.09%	88.68%	95.91%	77.69%

SVM performs better than kNN regarding all the performance measures. It correctly classifies all the negatives and misclassifies one positive entry. kNN misclassifies 2 positives and one negative entries.

Table 5.10 compares our results with the glucose binding site identifier COTRAN [39]. COTRAN is validated using a large testing set, where the negative

testing samples greatly outnumber the positive ones ($660 \gg 106$). This discrepancy in choosing the testing data explains the high specificity and low error rates achieved by COTRAN. Despite that fact, our SVM classifier reports higher specificity and precision values than COTRAN and a lower error rate. COTRAN achieves a slightly higher sensitivity, mainly due to our small positive testing set.

Finally, it is worth noting that we encountered a tricky PDB entry, 2ESR, where the glucose GLC-300 mistakingly seems, at first sight, to be docked to its binding-site. The SVM classifier rejected this binding site. Upon examination, this site turns out to be a pocket where glucose, floating in the crystallizing medium, happened to fit in. The site doesn't have any known glucose-affinity. We hereby report an example where our algorithm helped us to correctly classify a surface groove!

CHAPTER 6

CONCLUSION

6.1 Summary

This thesis work attempts to improve our understanding of the key features and aspects of glucose docking in proteins. Given the groove center, our algorithm classifies potential glucose binding sites in non-annotated proteins. In this work, we design predictors and descriptors for glucose binding sites, identifying their key features.

We build a classifier to recognize glucose binding sites among different binding-sites that do not bind hexoses and non-binding sites. We determine different atomic and residue properties that characterize a glucose binding site. We compare different SVM and kNN statistical pattern recognizers to select the best parameters that properly classify our data.

Given the center of a protein surface groove, our system uses SVM to correctly detect a glucose binding site 85.71% of the times and correctly rejects a non-glucose binding site 100% of the times. Using kNN, the system correctly detects a glucose binding site 85.71% of the times and correctly rejects a non-glucose binding site 86.67% of the times.

We show that glucose binding sites can be modeled using a limited number of features. SVM classification using atomic charge property alone gives a training error as low as 5.45%.

Our results support the relevance of ordered water molecules and ions in determining glucose specificity. We report the importance of planar polar residues in glucose binding, and specifically the carboxylate residues. Finally we note a high concentration of negatively charged atoms in direct contact with the bound glucose.

6.2 Future Work

This work can be extended in several directions. First, the system can be tested and validated using a larger experimental set. The used parameters can be fine tuned using a finer grid. The radius of the sphere and the number of the layers can be varied. We sample the different properties for all the layers, while some properties have a limited effective range. For example, hydrogen bond can not be established between atoms more than 4 Å apart. Sampling for hydrogen bonds at layer 8 is irrelevant and adds noise to the data. The range of the different atomic forces can be incorporated to the algorithm. In addition, further properties may be sampled and tested, such as secondary structure, solvent accessibility, evolutionary data...

This thesis uncovered the importance of certain glucose-binding sites features. Biological wet lab studies can be conducted to validate our findings by cellular and biochemical evidences.

A direct application of this work is to predict potential glucose binding sites. Our system can form the core of a glucose binding-site predictor. A program that identifies protein surface groove centers may feed the center coordinates to our program for functional prediction. Such a predicting system can even parse the whole PDB to predict and annotate potential glucose binding sites. A glucose docking simulation can be performed on some of our predicted glucose-binding motifs. Such a molecular dynamics simulation can be performed using modeling softwares.

This study focuses on glucose. A possible extension may include other hexoses, or even other biochemical families. One may even think of exploring the selective features that characterize each of the different hexose binding sites.

REFERENCES

[1] E. Abola, F. Bernstein, N. Manning, R. Shea, D. Stampf, and J. Sussman. *The Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description*, December 1996.
URL: http://www.rcsb.org/pdb/docs/format/pdbguide2.2/.

[2] Accelrys Inc., San Diego, United States of America. *Insight II, Molecular Modeling Environment, Release 2000.1*, 2002.
URL: http://www.accelrys.com/insight/.

[3] S. C. Bagley and R. B. Altman. Characterizing the microenvironment surrounding protein sites. *Protein Science*, 4:622–635, 1995.

[4] P. Baldi and S. Brunak. *Bioinformatics, the Machine Learning Approach*. MIT Press, Cambridge, Massachusetts, second edition, 2001.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[6] L. Bobadilla, F. Nino, and G. Narasimhan. Predicting and Characterizing Metal-Binding Sites Using Support Vector Machines. In *Proceedings of ICBA'04*, pages 307–318, 2004.

[7] L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

[8] L. Breiman and A. Cutler. *Random Forests*, 2004.
URL: http://www.stat.berkeley.edu/~breiman/RandomForests.

[9] F. A. Carey. *Organic Chemistry*. McGraw-Hill, 5th edition, 2003.

[10] R. Chakrabarti, A. M. Klibanov, and R. A. Friesner. Computational prediction of native protein ligand-binding and enzyme active site sequences. In *Proc. Natl. Acad. Sci.*, volume 102, pages 10153–10158, USA, 2005.

[11] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines.*, 2001.
URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[12] Y.-W. Chen and C.-J. Lin. Combining SVMs with Various Feature Selection Strategies. In I. M. Guyon, S. R. Gunn, M. Nikravesh, and L. Zadeh, editors, *Feature Extraction, Foundations and Applications*. Springer, 2006.

[13] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2002.

[14] R. Díaz-Uriarte and S. A. de Andrés. Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinformatics*, 7:3, 2006.

[15] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2006. R package version 1.5-13.

[16] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, second edition, 2001.

[17] G. M. Fung, O. L. Mangasarian, and J. W. Shavlik. Knowledge-Based Support Vector Machine Classifiers. In *NIPS*, pages 521–528, 2002.

[18] E. García-Hernández, R.A. Zubillaga, E.A. Chavelas-Adame, E. Vázquez-Contreras, A. Rojo-Domínguez, and M. Costas. Structural Energetics of Protein-Carbohydrate Interactions: Insights Derived from the Study of Lysozyme Binding to its Natural Saccharide Inhibitors. *Protein Science*, 12:135–142, 2003.

[19] K. Hechenbichler and K. Schliep. *Weighted k-Nearest-Neighbor Techniques and Ordinal Classification*. Discussion Paper 399, SFB 386, Ludwig-Maximilians University Munich, 2004.
URL: http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper399.ps.

[20] R. V. Hogg, J. W. McKean, and A. Craig. *Introduction to Mathematical Statistics*. Prentice Hall, 6th edition, 2004.

[21] H. R. Horton, L. A. Moran, R. S. Ochs, J. D. Rawn, and K. G. Scrimgeour. *Principles of Biochemistry*. Prentice Hall, prentice-hall/pearson education edition, 2002.

[22] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. *A Practical Guide to Support Vector Classification*. National Taiwan University, 2003.
URL: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[23] A. K. Jain and B. Chandrasekaran. Dimensionality and Sample Size Considerations in Pattern Recognition Practice. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 835–855. North-Holland, Amsterdam, 1982.

[24] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, January 2000.

[25] A. Jaramillo, L. Wernisch, S. Héry, and S. J. Wodak. Folding free energy function selects native-like protein sequences in the core but not on the surface. In *Proc. Natl. Acad. Sci.*, volume 99, pages 13554–13559, USA, 2002.

[26] S. Khuri, F. T. Bakker, and J. M. Dunwell. Phylogeny, Function and Evolution of the Cupins, a Structurally Conserved, Functionally Diverse Superfamily of Proteins. *Mol. Biol. Evol.*, 18:593–605, 2001.

[27] D. Meyer. Support Vector Machines. The Interface to LIBSVM in Package e1071. *R-News*, 1(3), 9 2001.

[28] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Singapore, 1997.

[29] M. Moorhouse and P. Barry. *Bioinformatics, Biocomputing and Perl*. Wiley, England, 2004.

[30] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–202, March 2001.

[31] W. S. Noble. Support Vector Machine Applications in Computational Biology. In B. Schoelkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 71–92. MIT Press, 2004.

[32] A. Peters and T. Hothorn. *ipred: Improved Predictors*, 2004. R package version 0.8-3.

[33] F. A. Quiocho and N. K. Vyas. Atomic Interactions Between Proteins/Enzymes and Carbohydrates. In S. M. Hecht, editor, *Bioorganic Chemistry: Carbohydrates*, chapter 11, pages 441–457. Oxford University Press, 1999.

[34] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL: http://www.R-project.org.

[35] V. S. R. Rao, K. Lam, and P. K. Qasba. Architecture of the Sugar Binding Sites in Carbohydrate Binding Proteins—a Computer Modeling Study. *International Journal of Biological Macromolecules*, 23:295–307, 1998.

[36] P. J. Russell. *Genetics*. Benjamin Cummings, fifth edition, 1998.

[37] B. Schölkopf. SVMs—a Practical Consequence of Learning Theory. *IEEE Intelligent Systems*, pages 18–21, July/August 1998.

[38] E. Solomon, L. Berg, D. W. Martin, and C. Villee. *Biology*. Saunders College Publishing, 4th edition, 1996.

[39] M. S. Sujatha and P. V. Balaji. Identification of Common Structural Features of Binding Sites in Galactose-Specific Proteins. *PROTEINS: Structure, Function, and Bioinformatics*, 55:44–65, 2004.

[40] M. S. Sujatha, Y. U. Sasidhar, and P. V. Balaji. Energetics of Galactose- and Glucose-Aromatic Amino Acid Interactions: Implications for Binding in Galactose-Specific Proteins. *Protein Science*, 13:2502–2514, 2004.

[41] S. A. Sullivan and D. Landsman. Characterization of Sequence Variability in Nucleosome Core Histone Folds. *PROTEINS: Structure, Function, and Genetics*, 52:454–465, 2003.

[42] C. Taroni, S. Jones, and J. M. Thornton. Analysis and Prediction of Carbohydrate Binding Sites. *Protein Engineering*, 13(2):89–98, 2000.

[43] J. Tisdall. *Beginning Perl for Bioinformatics*. O'Reilly, Sebastopol, CA, United States of America, 2001.

[44] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, United States of America, 1998.

[45] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.

[46] L. Wall and R. L. Schwartz. *Programming Perl*. O'Reilly & Associates, Sebastopol, CA, United States of America, 1992.

[47] G. Wang and R. L. Dunbrack. PISCES: a protein sequence culling server. *Bioinformatics*, 19:1589–1591, 2003.

[48] M. Zaki and L. Wong. Data Mining Techniques. In L. Zhang and L. Wong, editors, *Selected Topics in Post-Genome Knowledge Discovery*, chapter 4, pages 131–169. Singapore University Press, Singapore, 2004.