Shiraz University
Faculty of Sciences

M.Sc. Thesis in Analytical Chemistry

APPLICATION OF CHEMOMETRICS METHODS IN STUDYING THE EFFECTS OF SOLUTE STRUCTURES ON THE ELECTROCHEMISTRY OF STEROIDS AND ELECTRONIC ABSORPTION SPECTRA OF SUBSTITUTED SCHIFF BASES

By

MAHDIEH YAZDANI

Supervised by:

B. HEMMATEENEJAD Ph.D.

September 2009

In the Name of GOD

Declaration

I, Mahdieh Yazdani, a chemistry student, majored in analytical chemistry from the faculty of science declare that this thesis is the result of my research and I had written the exact reference and full indication wherever I used other's sources. I also declare that the research and the topic of my thesis are not reduplicative and guarantee that I will not disseminate its accomplishments and not make them accessible to others without the permission of the university. According to the regulations of the mental and spiritual ownership, all rights of this belong to Shiraz University.

Name: Mahdieh Yazadni
Date: 2009.9.5

IN THE NAME OF GOD

APPLICATION OF CHEMOMETRICS METHODS IN STUDYING THE EFFECTS OF SOLUTE STRUCTURES ON THE ELECTROCHEMISTRY OF STEROIDS AND ELECTRONIC ABSORPTION SPECTRA OF SUBSTITUTED SCHIFF BASES

BY

MAHDIEH YAZDANI

THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (M.Sc.)

IN

ANALYTICAL CHEMISTRY
SHIRAZ UNIVERSITY
SHIRAZ
ISLAMIC REPUBLIC OF IRAN

EVALUATED AND APPROVED BY THE THESIS COMMITTEE AS: EXCELLENT

.........................B. HEMMATEENEJAD, Ph.D., ASSOC. PROF. OF CHEMISTRY (CHAIRMAN)

..A.Safavi.....A. SAFAVI, Ph.D., PROF. OF CHEMISTRY

..........................H. SHARGHI, Ph.D., PROF. OF CHEMISTRY

..........................G. ABSALAN, Ph.D., ASSOC. PROF OF CHEMISTRY

SEPTEMBER 2009

Dedicated to:

My Parents

&

My Siblings

&

In memory of:

My Grandfather

ACKNOWLEDGMENTS

I thank the almighty GOD who always helped me in my life. I would like to express my deep appreciation to Assoc. Prof. B. Hemmateenejad for his excellent guidance, remarkable support, continuous encouragement and of course his brilliant kindness.

I would like to express my sincere gratitude to Prof. H. Sharghi for his matchless generosity and excellent support through this work.

Special thanks are due to the committee members; Prof. A. Safavi, Prof. H. Sharghi and Assoc. Prof. G. Absalan for their careful review of the manuscript, helpful suggestions and corrections.

I would like to show my unique thankfulness to my parents and my siblings for their endless support and lifelong kindness.

I would also like to state special thanks to all of my friends especially my classmates, my laboratory colleagues and my group members for their supporting behavior and affectionate relationships.

ABSTRACT

Application of chemometrics methods in studying the effects of solute structures on the electrochemistry of steroids and electronic absorption spectra of substituted schiff bases

BY

MAHDIEH YAZDANI

A data set consists of a diverse set of steroids have been used to develop QSER models for their $E_{1/2}$ by means of MLR and PCR analyses. In MLR analysis, the QSER models were constructed by first grouping descriptors and then stepwise selection of variables from each group (MLR1) and stepwise selection of predictor variables from the pool of all calculated descriptors (MLR2). Similar procedure was used in PCR analysis. In overall, the R^2 of CV and prediction of the QSER models resulted from MLR1, MLR2 and PCR1 approaches were higher than 90%, which show the high ability of the models to predict $E_{1/2}$ of the steroids.

A series of Schiff bases were studied for their delicate changes in absorption electronic spectra by changing substituents and solvents. Schiff base derivatives of different substituents in the aromatic ring were used. At the same time, different solvents having different solvatochromic parameters were used. Linear relationships were established to show how the variations of ν_{max} were related to molecular descriptors and solvatochromic parameters by changing substituents and solvent, respectively. MLR and factor analysis (FA) were applied to find the meaningful chemical factors and provide the regression models. It was found that the ν_{max} was mainly controlled by the solvent's polarity and the descriptors of Mor30p, Mor22v and E2v were selected for *para*, *meta* and *ortho* positions, respectively.

TABLE OF CONTENT

CONTENT **PAGE**

1 CHAPTER ONE: INTRODUCTION 1
1.1 Introduction 2
1.2 Chemometrics 3
1.3 Historical Development of Chemometrics 5
1.3.1 Development of Chemometrics 5
1.3.2 Chemometrics in Iran 6
1.4 Chemometrics Methods in Analytical Chemistry 7
1.5 Multiple Linear Regression 8
1.6 Principal Component Analysis 9
1.6.1 Pre-Treatment of Data 10
1.6.1.1 Scaling and Centering 10
1.6.1.2 Data Correction and Compression 12
1.6.1.3 Other Transformations 12
1.6.2 Principal Component Analysis: The Method 12
1.6.2.1 Rank and Eigenvalues 14
1.7 Factor Analysis 16
1.8 Principal Component Regression 17
1.9 Overview of Quantitative Structure Activity/Property Relationships 18
1.9.2 Underlying Principle 20
1.9.3 QASR/QSPR Model 20
1.9.4 QSAR/QSPR Steps 21
1.9.4.1 Formulation of Classes of Similar Compounds 21
1.9.4.2 Quantitative Description of Structural Variation (Descriptor Generation) 22
1.9.4.3 Selection of the Training and Test Sets (Series Design) 23
1.9.4.4 Variable Selection 23
1.9.4.5 Model Development 23
1.9.4.6 Evaluation of the Predictive Capacity of the Model: (Validation Techniques) 24
1.10 Solvent and Substituent Effects on Electronic Absorption Spectra 26
1.11 Chemometrics in Spectroscopy 27
1.12 Half-Wave Reduction Potential 29
1.13 Compounds under Investigation 29
1.13.1 Steroids 29
1.13.2 Schiff Bases 30

2 CHAPTER TWO: LITERATURE REVIEW 32
2.1 The Birth of QSAR/QSPR; Origins and Evolution 33

2.1.1 Linear Free-Energy Relationships 34
2.1.2 Hansch Analysis.. 36
2.2 Application of QSPR for Modeling the Half-Wave Reduction Potential... 37
2.3 Solvent and Substituent Effects on Absorption Spectra 39
2.4 Research Objective 41
2.4.1 QSPR Modeling of Half-Wave Reduction Potential of Steroids......... 41
2.4.2 A Chemometrics Study on the Solvent & Substituent Effects on Electronic Absorption Spectra of Some Substituted Schiff Bases 41

3 CHAPTER THREE: EXPERIMENTAL 42
3.1 QSPR Modeling of the Half-Wave Reduction Potential of Steroids.......... 43
3.1.1 Data Set 43
3.1.2 Descriptor Generation 46
3.1.3 Data Pre-Processing 46
3.1.4 Model Development.. 48
3.2 A Chemometrics Study on the Solvent & Substituent Effects on Electronic Absorption Spectra of Some Substituted Schiff Bases.................................... 49
3.2.1 Reagents 49
3.2.2 Spectral Measurements Procedure... 50
3.2.3 The Methodology for Deriving Maximum absorption Wavelengths and Molar Absorptivity Coefficients... 51
3.2.4 Computational Methods 52
3.2.4.1 Software.. 52
3.2.4.2 Descriptor Generation 52
3.2.5 Model Development.. 52
3.2.5.1 MLR Modeling for Calculation of Substituent Effect.................. 53
3.2.5.2 MLR Modeling for Calculation of Solvent Effect 54
3.2.5.3 Factor Analysis Procedure... 54

4 CHAPTER FOUR: RESULTS AND DISCUSSION.................................. 56
4.1 QSPR Modeling of the Half-Wave Reduction Potential of Steroids.......... 57
4.1.1 A Strategy for Solving Chance Correlation.. 59
4.1.2 MLR Analysis.. 60
4.1.3 PCR Analysis... 66
4.1.4 Robustness and Applicability Domain of the Models......................... 75
4.1.5 Conclusion... 76
4.2 A Chemometrics Study on the Solvent & Substituent Effects on Electronic Absorption Spectra of Some Substituted Schiff Bases.................................... 78
4.2.1 The Recorded Spectra 78
4.2.2 The Calculated Pure Spectra.. 93
4.2.3 The Methodology for Accurate Determination of the Maximum Absorbance.. 100
4.2.4 Effect of Conjugation of Chromophores 116
4.2.5 MLR Results.. 117

4.2.6 FA Results 123
4.2.7 Comments on the Selected Solvatochromic Parameter (A Discussion about Solvent Effect) 130
4.2.8 Comments on the Selected Descriptors (A Discussion about Substituent Effect) 131
4.2.8.1 *Para* Position.. 131
4.2.8.2 *Meta* Position 133
4.2.8.3 *Ortho* Position 133
4.2.9 Conclusion 135

REFERENCES 136

LIST OF FIGURES

FIGURES **PAGE**

Figure 1.1: How chemometrics relates to other disciplines.................................. 4

Figure 1.2: A typical data set 9

Figure 1.3: Auto-scaling 11

Figure 1.4: Principal components analysis 13

Figure 1.5: Relationship between PCA and FA with an example 17

Figure 1.6: A graphical representation of the first two PCs 17

Figure 1.7: A Gaussian together with its first and second derivatives................ 27

Figure 1.8: Three peaks forming a cluster 28

Figure 1.9: General structure of Steroids... 29

Figure 1.10: Basic structure of Schiff bases 30

Figure 3.1: General structures of the steroids used in this study 43

Figure 3.2: General structure of the synthesized Schiff bases used in this work.. 49

Figure 3.3: The shematic diagram of the data matrix obtained for each molecule50

Figure 3.4: The shematic diagram of the data matrix obtained for each set......... 53

Figure 4.1: Plot of the first PC of the descriptors data matrix against the second PC for selecting calibration and prediction molecules....................................... 59

Figure 4.2: Changes in the CV correlation coefficients of MLR1....................... 61

Figure 4.3: Changes in the CV correlation coefficients of the MLR2 64

Figure 4.4: Changes in the CV correlation coefficients of the PCR1 68

Figure 4.5: The correlation coefficient between the extracted PCs in PCR2 approach and the $E_{1/2}$ 72

Figure 4.6: Changes in the CV correlation coefficients of the PCR2 73

Figure 4.7: The recorded spectra... 78

Figure 4.8: The calculated pure spectra... 93

Figure 4.9: A selection of the employed first and second derivatives 101

Figure 4.10: The recorded spectra for *ortho*-NO_2 substituent 115

Figure 4.11: Changing in λ_{max} of *ortho* substituents against pi* 121

Figure 4.12: Changing in λ_{max} of *meta* substituents against pi^* 121

Figure 4.13: Changing in λ_{max} of *para* substituents against pi^* 122

Figure 4.14: Changing in λ_{max} of *ortho*-NO_2 against pi^* 122

Figure 4.15: Resonance states O=*Ortho*, M=*Meta*, P=*Para* 134

LIST OF TABLES

TABLES **PAGE**

Table 2.1: QSPR papers on the subject of modeling the $E_{1/2}$ 38

Table 3.1: Substituent patterns of the steroids 44

Table 3.2: List of Dragon descriptors used in this work 47

Table 3.3: List of *ortho*, *meta* and *para* substituents used in this work 49

Table 3.4: Solvatochromic parametes of the solvent used 54

Table 4.1: The experimental half-wave potentials and the cross-validated predicted values of the training set 57

Table 4.2: Statistical results and the selected descriptors of the MLR models obtained from each set of descriptors (MLR1 approach) 62

Table 4.3: Different groups of descriptors used in this study and the number of descriptors remained after preprocessing and the extracted PCs in each group ... 67

Table 4.4: Statistical parameters of the resulted models 69

Table 4.5: The selected PCs in the PCR1 model and the list of descriptors represented significant correlation with the PCs.. 69

Table 4.6: Correlation between the selected PCs in PCR1 and the selected descriptors in MLR1 and MLR2 71

Table 4.7: External prediction results of $E_{1/2}$ 74

Table 4.8: The resulted Q^2 in test of robustness of the proposed models............ 75

Table 4.9: Leverage (h) of the external test set molecules for different models... 76

Table 4.10: λ_{max} of the longest wavelength for the *ortho* substituted derivatives in different solvents 112

Table 4.11: λ_{max} of the longest wavelength for the *meta* substituted derivatives in different solvents 112

Table 4.12: : λ_{max} of the longest wavelength for the *para* substituted derivatives in different solvents 113

Table 4.13: Molar absorptivity coefficients of *ortho* substituents 113

Table 4.14: Molar absorptivity coefficients of *meta* substituents 114

Table 4.15: Molar absorptivity coefficients of *para* substituents 114

Table 4.16: λ_{max} of the longest wavelength for the *ortho*-NO_2 substituted derivatives in different solvents 115

Table 4.17: MLR models for accounting substituent effect in *ortho* position.... 117

Table 4.18: MLR models for accounting substituent effect in *meta* position..... 118

Table 4.19: MLR models for accounting substituent effect in *para* position..... 118

Table 4.20: MLR models for accounting solvent effect in *ortho* position 119

Table 4.21: MLR models for accounting solvent effect in *meta* position 120

Table 4.22: MLR models for accounting solvent effect in *para* position 120

Table 4.23: Target testing results for *ortho* set 124

Table 4.24: Target testing results for *meta* set 124

Table 4.25: Target testing results for *para* set 124

Table 4.26: TFA model for *ortho* substituents... 126

Table 4.27: TFA model for *meta* substituents.. 126

Table 4.28: TFA model for *para* substituents.. 126

Table 4.29: Target testing results for *ortho* set 127

Table 4.30: Target testing results for *meta* set 127

Table 4.31: Target testing results for *para* set 127

Table 4.32: TFA model for *ortho* substituents... 128

Table 4.33: TFA model for *meta* substituents.. 129

Table 4.34: TFA model for *para* substituents.. 129

GLOSSARY OF ABBREVIATIONS:

AET	Apparent Error in Target
CV	Cross Validation
$E_{1/2}$	Half-wave reduction potential
FA	Factor Analysis
FE	Feature Extraction
FS	Feature Selection
GA	Genetic Algorithm
HPLC	High Performance Liquid Chromatography
HPLC-DAD	HPLC-Diode Array Detector
LFER	Linear Free Energy Relationships
LMO	Leave Many Out
LOO	Leave One Out
MLR	Multiple Linear Regression
NIPALS	Nonlinear Iterative Partial Least Squares
NIR	Near Infrared
PCA	Principal Component Analysis
PCR	Principal Component Regression
PCs	Principal Components
PLS	Partial Least Squares
QSAR	Quantitative Structure Activity Relationship
QSER	Quantitative Structure Electrochemistry Relationship
QSPR	Quantitative Structure Property Relationship
$RMSE_{\mathbf{cv}}$	Root Mean Squared Error of Cross Validation
$RMSE_{\mathbf{p}}$	Root Mean Squared Error of Prediction
SARs	Structure Activity Relationships
SSPE	Sum of Squares of Prediction Errors
Std	Standard Deviation
SVD	Singular Value Decomposition
TFA	Target Factor Analysis

CHAPTER ONE

INTRODUCTION

1.1 Introduction

The field of chemistry is currently facing major changes. Optical, mechanical, and microelectronic technologies have advanced rapidly in recent years. Computer power has increased dramatically as well. All these developments, together with other factors, provide a new opportunity but also challenge to chemists in research and development [1].

Sound chemical information that forms the basis of many of humanity's important decision-making processes depends on three critical properties of the measurement process, including its (1) chemical properties, (2) physical properties, and (3) statistical properties. The conditions that support sound chemical measurements are like a platform supported by three legs. Credible information can be provided only in an environment that permits a *thorough understanding and control* of these three critical properties of a chemical measurement:

1. Chemical properties, including stoichiometry, mass balance, chemical equilibria, kinetics, etc.
2. Physical properties, including temperature, energy transfer, phase transitions, etc.
3. Statistical properties, including sources of errors in the measurement process, control of interfering factors, calibration of response signals, modeling of complex multivariate signals, etc.

If any one of these three legs is missing or absent, the platform will be unstable and the measurement system will fail to provide reliable results, sometimes with catastrophic consequences. It is the role of statistics and chemometrics to address the third critical property. It is this fundamental role that provides the primary motivation for developments in the field of chemometrics. Sound chemometric methods and a well-trained work force are necessary for providing reliable chemical information for humanity's decision-making activities [2, 3].

1.2 Chemometrics

The term chemometrics was first coined in 1971 to describe the growing use of mathematical models, statistical principles, and other logic-based methods in the field of chemistry and, in particular, the field of analytical chemistry. Chemometrics is an interdisciplinary field that involves multivariate statistics, mathematical modeling, computer science, and analytical chemistry. Some major application areas of chemometrics include (1) calibration, validation, and significance testing; (2) optimization of chemical measurements and experimental procedures; and (3) the extraction of the maximum of chemical information from analytical data [2, 4]. A reasonable definition of chemometrics remains as how do we get chemical relevant information out of measured chemical data, how do we represent and display this information, and how do we get such information into data? As mentioned by Wold [5]. Chemometrics is considered by some chemists to be a subdiscipline that provides the basic theory and methodology for modern analytical chemistry [1].

In many respects, the field of chemometrics is the child of statistics, computers, and the "information age." Rapid technological advances, especially in the area of computerized instruments for analytical chemistry, have enabled and necessitated phenomenal growth in the field of chemometrics over the past 30 years. The relationship of chemometrics to different disciplines is indicated in Figure 1.1. On the left are the enabling sciences, mainly quite mathematical and not laboratory based. Statistics, of course, plays a major role in chemometrics. Statistical approaches are based on mathematical theory, so statistics falls between mathematics and chemometrics. Computing is important as much of chemometrics relies on software. Engineers, especially chemical and process engineers, have an important need for chemometric methods in many areas of their work, and

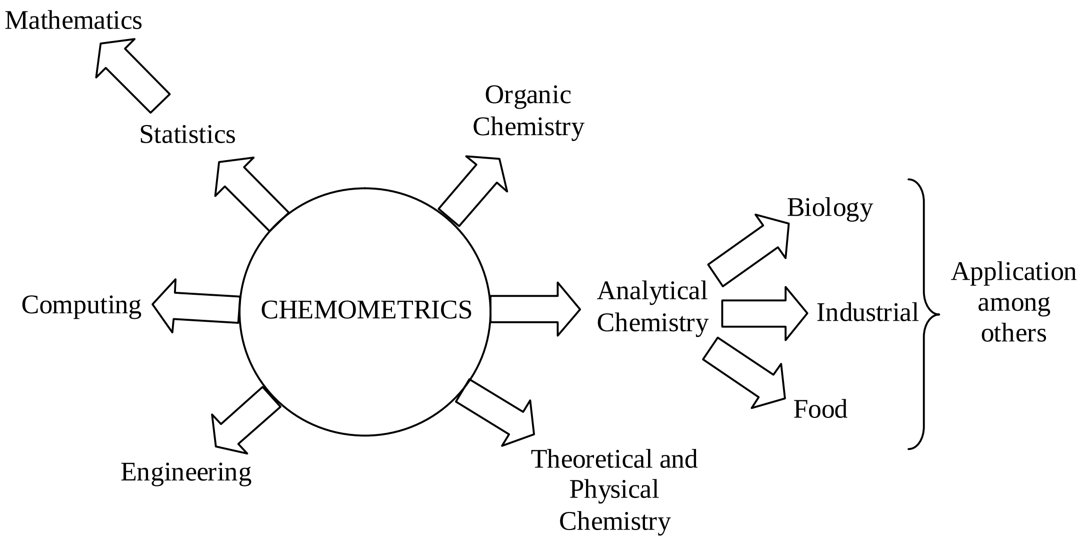

Figure 1.1: How chemometrics relates to other disciplines

have a quite different perspective from the mainstream chemist. On the right are the main disciplines of chemistry that benefit from chemometrics. Analytical chemistry is probably the most significant area. Environmental chemists, biologists, food chemists as well as geochemists, chemical archaeologists, forensic scientists and so on depend on good analytical chemistry measurements and many routinely use multivariate approaches especially for pattern recognition, and so need chemometrics to help interpret their data. These scientists tend to identify with analytical chemists. The organic chemist has a somewhat different need for chemometrics, primarily in the areas of experimental design (e.g. optimising reaction conditions) and QSAR (quantitative structure–analysis relationships) for drug design. Finally, physical chemists such as spectroscopists, kineticists and materials scientists often come across methods for signal deconvolution and multivariate data analysis [6].

To finalize the point, it should be said that chemometrics is the interface between chemistry and mathematics. As Kowalski [7-10] clearly stated "Chemometric tools are vehicles that can aid chemists to move more efficiently on the path from measurements to information to knowledge"[11].

1.3 Historical Development of Chemometrics

The term *chemometrics* was introduced by Svante Wold [5] and Bruce R. Kowalski in the early 1970s. Terms like *biometrics* and *econometrics* were also introduced into the fields of biological science and economics. Afterward, the International Chemometrics Society was established. Since then, chemometrics has been developing and is now widely applied to different fields of chemistry, especially analytical chemistry.

1.3.1 Development of Chemometrics

Chemometrics has developed over the past decade from a fairly theoretical subject to one that is applied in a wide range of sciences. Early development went hand in hand with the development of scientific computing, and primarily involved using multivariate statistical methods for the analysis of analytical chemistry data. Early chemometricians were likely to be FORTRAN programmers using mainframes and statistical libraries of subroutines. The earliest applications were primarily to analytical chemical datasets, often fairly simple in nature, for example a high performance liquid chromatography (HPLC) cluster of two or three peaks or a set of UV/visible mixture spectra.

Chemometrics as a discipline became organized in the 1980s, with the first journals, meetings, books, societies, packages and courses dedicated to the subject. Historically HPLC and near infrared (NIR) spectroscopy provided particularly important growth points in the 1980s, partly due to the economic driving forces at the time and partly due to the research interest and contacts of the pioneers. Industrial applications were particularly important in the developing phase. In the 1990s the application of chemometrics started expanding, with special emphasis on certain industries especially the pharmaceutical industry, where large companies provided significant funding for developments. A new and exciting phase has emerged as from the late 1990s, involving very complex datasets. This new and exciting phase is possible due to the capacity of analytical instruments to acquire large amounts of data rapidly, for example via autosamplers, and chemometrics becomes a type of data mining. All the building blocks are available, for example signal processing, chromatographic alignment, data scaling and pattern recognition but the problems are now much more complex. In addition many of the newer applications are biologically driven and so there is emerging a new interface between chemometrics and bioinformatics, especially the relationship between analytical data as obtained by spectroscopy or chromatography and genetics, requiring some appreciation of gene or protein sequence and similarities.

Chemometrics became much more widespread today. Individual researchers in academic positions, some with prior training in some of the more established groups, and others who read papers and books that had been written over this early period, dabbled with chemometrics, and so the applications spread not just to dedicated research groups but to investigators working in different environments [12].

1.3.2 Chemometrics in Iran

Recently, Hemmateenejad has published a paper [13] to represent the activity of the Iranian chemometrics community and to increase the awareness about the studies of chemometrics in Iran. In this article a list of the publication of the Iranian scientists in the chemometrics was collected from the ISI web of science database. A rapid growth is observed in the