Università degli Studi di Milano-Bicocca
Facoltà di Scienze Matematiche Fisiche e Naturali
Dipartimento di Scienze dell'Ambiente e del Territorio

Corso di Dottorato di Ricerca in Scienze Chimiche
Ciclo XX

tesi di Dottorato di Ricerca

Protein and peptide multivariate characterisation using a molecular descriptor based approach

CHIM01

Andrea Mauri

Tutor: Prof. Roberto Todeschini
Coordinatore del Dottorato: Prof.ssa Franca Morazzoni

Anno Accademico, 2006/2007

Cover illustration: ”proletarian protein” by labio
PhD thesis:
Protein and peptide multivariate characterisation using a molecular descriptor based approach
2007, Andrea Mauri

Special thanks to Roberto for teaching me chemometrics and for guiding me in the wide world of molecular descriptors.

I would like to acknowledge Prof. Marjana Novič for both his help, hospitality and suggestions.

Many thanks to Labio, and to its graphics skills and to Alberto for helping me in the development of the web-based application.

Thanks to all the people I met in the lab these years, in particular: Cristian, Manuela, Viviana and, again, Alberto and Labio.

Finally, thanks to all my friends and my family. Especially to all the tristian family, and to Chiara, the woman I love.

a mio padre

Contents

part I

Theory

chapter 1

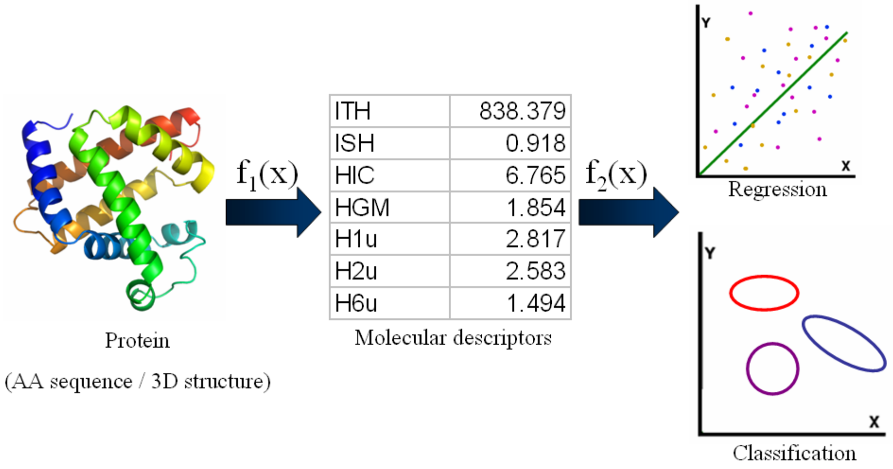
Introduction

1.1 Introduction

Large databases of protein sequences and structures are now freely available (http://www.pdb.org). Analogously the increasing number of peptide sequences with different lengths, available from synthesised peptide libraries and sequenced proteins are potentially valuable for evaluating structure-activity relationships [Gallop *et al.* (1994)].

However, in order to apply multivariate regression and classification searching for Quantitative Structure-Activity Relationship (QSAR) on such sequences, it is necessary to have a preprocessing method that translates them into a uniform set of variables.

In order to characterise and predict properties of this kind of molecular structure a molecular descriptor based approach is suitable. Unfortunately a traditional molecular descriptor based approach is not always applicable to molecule with thousands of atoms, such as proteins. During the last years different methodology in order to describe peptides and proteins have been published, the most used are the methods based on z-scores [Hellberg *et al.* (1987), Sjstrm *et al.* (1995), Sandberg *et al.* (1998), Andersson *et al.* (1998), Edman *et al.* (1999), Nystrm *et al.* (2000), Doytchinova *et al.* (2002), Doytchinova and Flower (2003), Guan *et al.* (2005), Doytchinova and Flower (2005, 2006b,a, 2007b,a)].



ITH	838.379
ISH	0.918
HIC	6.765
HGM	1.854
H1u	2.817
H2u	2.583
H6u	1.494

Figure 1.1: The information content of a molecular descriptor depends on the kind of molecular representation that is used and on the defined algorithm for its calculation. Obtained molecular descriptors are then suitable for modelling.

During this PhD thesis a novel methodology has been deeply evaluated, this methodology is based on some traditional molecular descriptors calculated on a simplified representation of peptides and proteins. This representation avoid problems related to molecular size and information redundancy due to the common structural features of every amino acid. The proposed methodology has been applied both on peptide and protein data sets.

1.2 Thesis structure

This thesis is focused on the study of a novel characterisation of proteins and peptides using a descriptors-based approach. The calculated descriptors values are then analysed by means of multivariate analysis. Consequently, chemometric methods applied in this thesis have been deepened described and analysed.

Great attention has been also given to variable (or feature) selection methods, since the great number of obtained molecular descriptors needs to be reduced in order to obtain reliable regression models.

Summarising, the structure of the thesis can be outlined with the following points:

1. in the first part of the thesis a brief introduction to chemometrics and QSAR is presented in chapter 2, in the same chapter proteins and peptides are shortly described in order to characterise the studied molecular structures. In chapter 3 the protein representation applied in order to calculate the molecular descriptors is presented. Afterwards the molecular descriptors chosen to represent protein and peptide structures are described and explained in chapter 4. Finally the chemometric methods used to analyse the data matrices obtained from the calculation of molecular descriptors on three different data sets studied in this thesis are described in chapter 5;

2. in the second part of the thesis, applications of the proposed approach on three different data sets are presented. In chapter 6 a brief resume of all the applications is showed; in chapters 7 a sensitivity analysis of the proposed method on a computationally generated data set of protein mutants is presented; in chapter 8 a cluster analysis of two different protein folds collected from the SCOP [Murzin *et al.* (1995), LoConte *et al.* (2002), Andreeva *et al.* (2004)] database is described showing how different characterisation of amino acids drive to different highlighted information; in chapter 9 a QSAR analysis based on 20 peptide sequences of different lengths is presented.

In order to give the possibility to test and trial the molecular descriptors-based approach described in this thesis a web-based application has been also developed. The presentation of this application is collected in chapter 10.

All the descriptors calculated during this PhD thesis have been calculated using an ongoing implementation of dragonX, software for molecular descriptors calculation [Mauri *et al.* (2006), dra (2007)].

chapter 2

Chemometrics, QSAR, Peptides and Proteins

A brief introduction on chemometrics and Quantitative Structure Activity Relationships (QSAR) is given in this chapter. Chemometric methods have been used in order to analyse data useful to build QSAR models. Peptides and proteins are also briefly described in order to understand the molecular structure of these kind of molecules whose analysis is the scope of this thesis.

2.1 Chemometrics

Chemometrics has been defined in broad terms as the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods according to the International Chemometrics Society, 2002.

However, the definition of the word chemometrics has been a subject of discussion and no exact consensus is available, despite of the fact that two international scientific journals and numerous of international and national scientific societies are dedicated to chemometrics and use the word in their titles. It is known that Svante Wold invented the word chemometrics in 1972 to describe the discipline of extracting chemically relevant information from chemical experiments

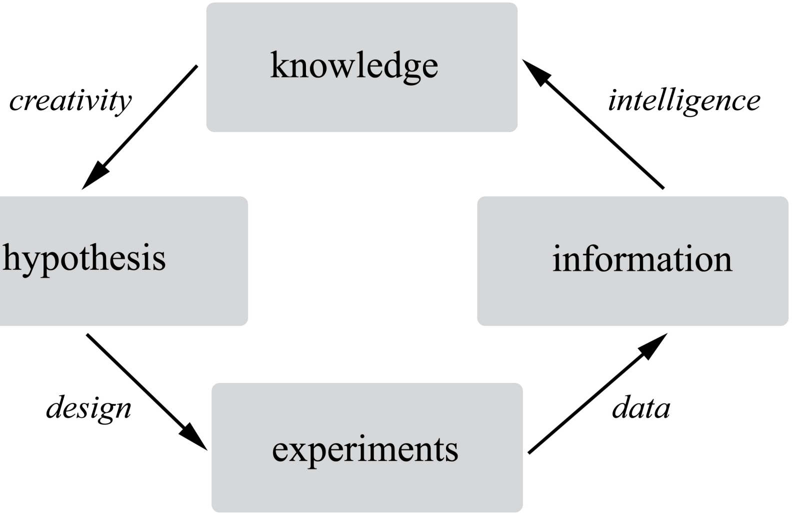


Figure 2.1: Chemometric rule in the knowledge circle

[Wold (1972)]. He tried to re-define the word as how to get chemically relevant information out of measured chemical data, how to represent and display this information, and how to get such information into the data [Wold (1990)]. A more precise definition can be found in a textbook by Massart et al. [Massart *et al.* (1997)], stating that chemometrics is the chemical discipline that uses mathematics, statistics and formal logic (a) to design or select optimal experimental procedures; (b) to provide maximum relevant chemical information by analysing chemical data; and (c) to obtain knowledge about chemical systems (Figure 2.1). This definition is very close to the formulation used by Svante Wold and Bruce Kowalski when founding the first Chemometrics Society in 1974.

The use of chemometrics also implies the use of multivariate data analysis, in which several related molecules are analysed simultaneously. A multivariate approach when handling and exploring complex chemical data and designing experiments is certainly part of the foundation of chemometrics. Multivariate data analysis as opposed to using only one or a few variables in the data analysis is based on the fact that complex problems - by nature - need multiple variables to be described. Thus, by using and combining more variables, more information about the chemical system can be retrieved. In standard multivariate data analysis, data are arranged in a two-way structure, a table or a matrix. An example is a table in which each row corresponds to a sample and each column to a variable describing the complex system. This is the typical input for multivariate techniques: when these matrices are analysed by means of chemometrics, all the variables are considered at the same time and consequently the extracted information represent a global overview of the system.

Since chemometrics proved to be able to handle large amounts of data and to extract useful information, it has been successfully applied in different fields. During the last years it has so increased in uses and applications that now modern analytical techniques are usually combined with chemometric methods.

2.2 QSAR and Molecular Descriptors

Chemometrics has been most successfully applied in four areas, namely:

1. multivariate calibration;
2. quantitative structure-activity relationship (QSAR) studies;
3. pattern recognition, classification and discriminant analysis;
4. multivariate modeling and monitoring processes.

In QSAR the aim is correlate chemical data series of compounds (i.e. compounds contained in a "chemical space") to a biological activity. "Biological activity"relates to the strength of interaction of a compound with a target, whatever the target is, e.g. an organism, a cell, or a protein.

The essence of the QSAR methodology is developing a relationship between an observed property and structural features of a molecule. By considering a set of molecules, a predictive model is developed that can then be used to predict the activity of other molecules. The key words here are "structural features". The approach depends on being able to represent the structure of a molecule in numerical form. The numerical representations of molecules are termed descriptors, and a wide variety of descriptors can be calculated. These include simple forms such as molecular weight and atom counts or more complex types such as partition coefficients and surface-property descriptors.

Given a set of descriptors, a QSAR model can be built by defining a relationship between these descriptors (also known as the independent variables) and

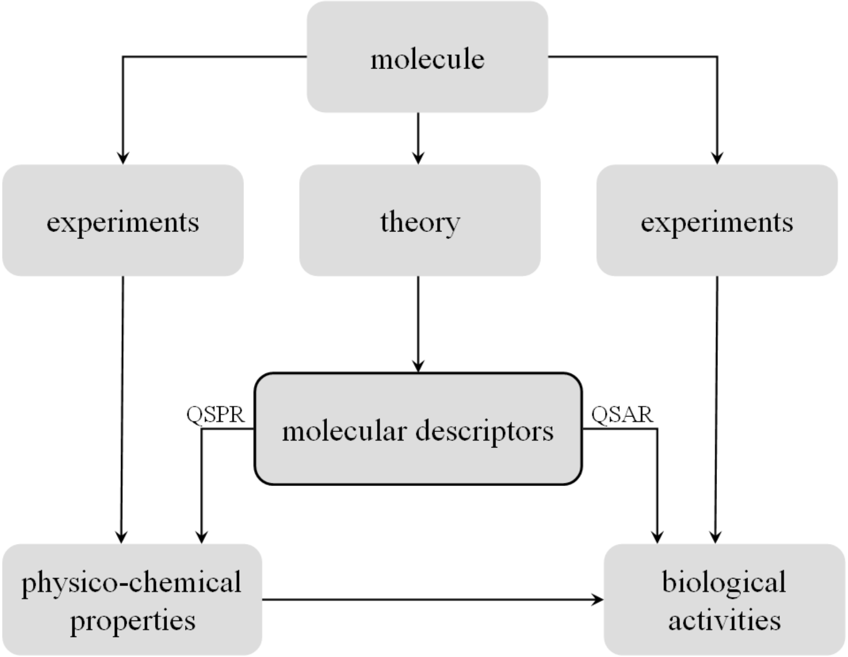


Figure 2.2: Role of molecular descriptors in Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR)

the observed property (termed the dependent variable). The first QSAR models, developed by Hansch [Hansch and Fujita (1964), Hansch (1969)] and Free-Wilson [Free and Wilson (1964)] specified linear relationships. Even now, linear models are widely used owing to their simplicity and ease of development.

2.3 Peptides and proteins

Peptides are short polymers formed from the linking, in a defined order, of α amino acids. The link between one amino acid residue and the next is known as an amide bond or peptide bond.

Proteins are polypeptide molecules (or consist of multiple polypeptide subunits). The distinction is that peptides are short and polypeptides/proteins are long. Proteins and peptides share the same basic structure, both are made by sequences of amino acids. All amino acids share common structural features in-

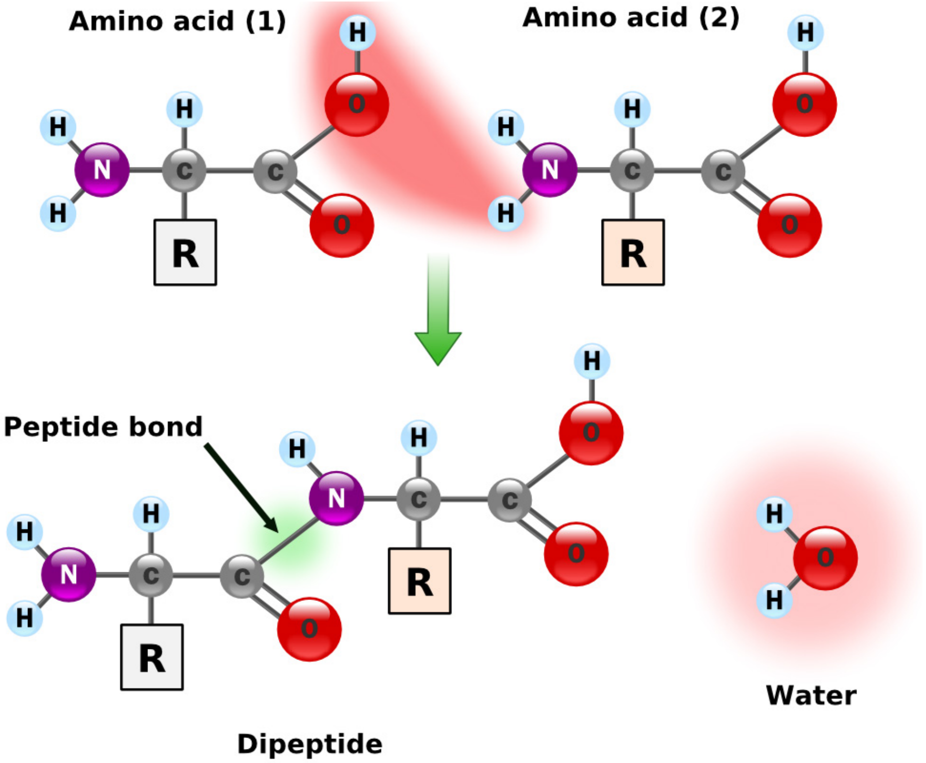


Figure 2.3: The general structure of an α amino acid, with the amino group on the left and the carboxyl group on the right (1) and the condensation of two amino acids to form a peptide bond

cluding an α-carbon to which an amino group, a carboxyl group, and a variable side chain are bonded. Only proline differs from this basic structure, as it contains an unusual ring to the N-end amine group, which forces the CO-NH amide moiety into a fixed conformation [Nelson and Cox (2005)]. The side chains of the standard amino acids, detailed in the list of standard amino acids (Table 2.1), have different chemical properties that produce proteins' three-dimensional structure and are therefore critical to protein function. The amino acids in a polypeptide chain are linked by peptide bonds formed in a dehydration reaction. Once linked in the protein chain, an individual amino acid is called a residue and the linked series of carbon, nitrogen, and oxygen atoms are known as the main chain or protein backbone.

The general formula of an α amino acid is H2NCHRCOOH, where R is an organic substituent. In the α amino acids, the amino and carboxylate groups are

Table 2.1: The twenty standard amino acids and their side chain (SC) chemical properties.

Amino Acid	**3-Letter**	**1-Letter**	**SC polarity**	**SC acidity or basicity**
Alanine	Ala	A	nonpolar	neutral
Arginine	Arg	R	polar	basic (strongly)
Asparagine	Asn	N	polar	neutral
Aspartic acid	Asp	D	polar	acidic
Cysteine	Cys	C	polar	neutral
Glutamic acid	Glu	E	polar	acidic
Glutamine	Gln	Q	polar	neutral
Glycine	Gly	G	nonpolar	neutral
Histidine	His	H	polar	basic (weakly)
Isoleucine	Ile	I	nonpolar	neutral
Leucine	Leu	L	nonpolar	neutral
Lysine	Lys	K	polar	basic
Methionine	Met	M	nonpolar	neutral
Phenylalanine	Phe	F	nonpolar	neutral
Proline	Pro	P	nonpolar	neutral
Serine	Ser	S	polar	neutral
Threonine	Thr	T	polar	neutral
Tryptophan	Trp	W	nonpolar	neutral
Tyrosine	Tyr	Y	polar	neutral
Valine	Val	V	nonpolar	neutral

attached to the same carbon, which is called the α-carbon. The various α amino acids differ in which side chain (R group) is attached to their α-carbon. They can vary in size from just a hydrogen atom in glycine, through a methyl group in alanine, to a large heterocyclic group in tryptophan. α amino acids are the building blocks of proteins. A protein forms via the condensation of amino acids to form a chain of amino acid "residues"linked by peptide bonds.

Proteins are defined by their unique sequence of amino acid residues; this sequence is the primary structure of the protein. Just as the letters of the alphabet can be combined to form an almost endless variety of words, amino acids can be linked in varying sequences to form a huge variety of proteins. Twenty amino acids are encoded by the standard genetic code and are called standard amino acids. In the structure shown in Figure 2.3, the R represents a side chain specific to each amino acid. The central carbon atom called C_α is a chiral central carbon atom (with the exception of glycine) to which the two termini and the R-group are attached. Amino acids are usually classified by the properties of the side chain into four groups. The side chain can make them behave like a weak acid, a weak base, a hydrophile if they are polar, and hydrophobe if they are nonpolar. The chemical structures of the 20 standard amino acids is shown in Figure 2.4, along with their chemical properties, are catalogued in the list of standard amino acids in table Table 2.1.

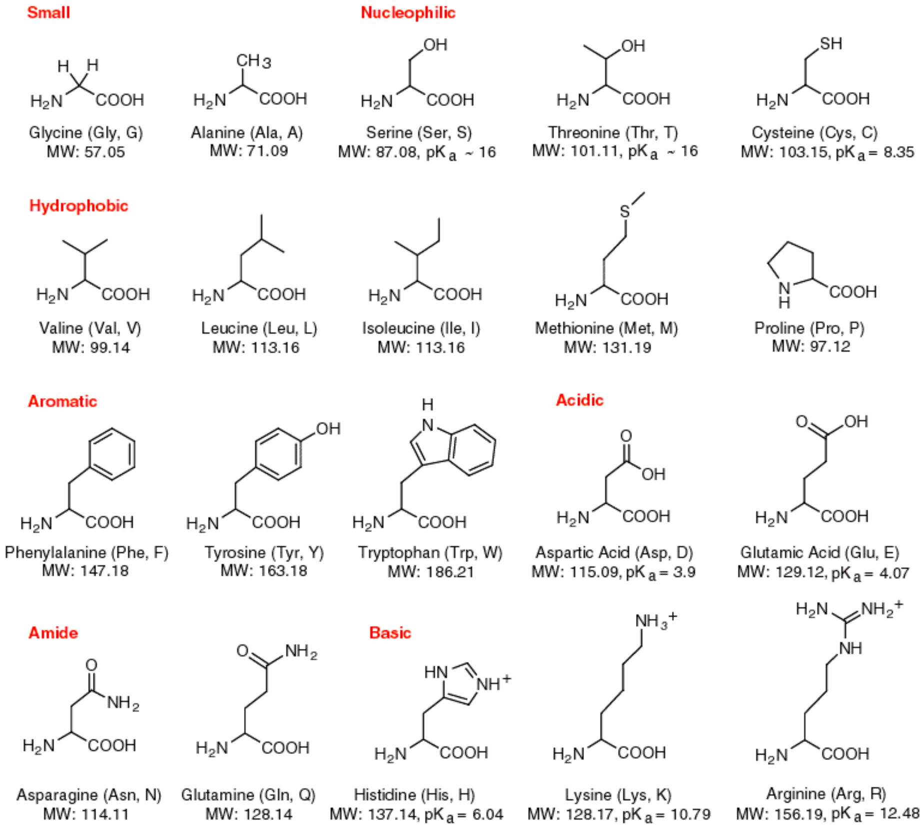

Small
Nucleophilic
Glycine (Gly, G) MW: 57.05
Alanine (Ala, A) MW: 71.09
Serine (Ser, S) MW: 87.08, pKa ~ 16
Threonine (Thr, T) MW: 101.11, pKa ~ 16
Cysteine (Cys, C) MW: 103.15, pKa = 8.35
Hydrophobic
Valine (Val, V) MW: 99.14
Leucine (Leu, L) MW: 113.16
Isoleucine (Ile, I) MW: 113.16
Methionine (Met, M) MW: 131.19
Proline (Pro, P) MW: 97.12
Aromatic
Acidic
Phenylalanine (Phe, F) MW: 147.18
Tyrosine (Tyr, Y) MW: 163.18
Tryptophan (Trp, W) MW: 186.21
Aspartic Acid (Asp, D) MW: 115.09, pKa = 3.9
Glutamic Acid (Glu, E) MW: 129.12, pKa = 4.07
Amide
Basic
Asparagine (Asn, N) MW: 114.11
Glutamine (Gln, Q) MW: 128.14
Histidine (His, H) MW: 137.14, pKa = 6.04
Lysine (Lys, K) MW: 128.17, pKa = 10.79
Arginine (Arg, R) MW: 156.19, pKa = 12.48

Figure 2.4: Chemical structures of the 20 standard amino acids. Name, abbreviations and molecular weights are reported.

Depending on the polarity of the side chain, amino acids vary in their hydrophilic or hydrophobic character. These properties are important in protein structure and protein-protein interactions. The importance of the physical properties of the side chains comes from the influence this has on the amino acid residues' interactions with other structures, both within a single protein and between proteins. The distribution of hydrophilic and hydrophobic amino acids determines the tertiary structure of the protein, and their physical location on the outside structure of the proteins influences their quaternary structure. For example, soluble proteins have surfaces rich with polar amino acids like serine and threonine, while integral membrane proteins tend to have outer ring of hydrophobic amino acids that anchors them into the lipid bilayer, and proteins anchored to the membrane have a hydrophobic end that locks into the membrane.

Similarly, proteins that have to bind to positively-charged molecules have surfaces rich with negatively charged amino acids like glutamate and aspartate, while proteins binding to negatively-charged molecules have surfaces rich with positively charged chains like lysine and arginine.

chapter 3

Protein and Peptide representation

Chemical compounds are usually represented as molecular graphs [Harary (1971)], i.e. non-directed, connected graphs in which vertices correspond to atoms and edges represent covalent bonds between atoms. The molecular graph model of the chemical structure emphasises the chemical bonding pattern of atoms [Balaban (1976)]. The molecular graph model is appropriate for prediction of physical, chemical or biological properties of the studied molecules.

3.1 Introduction

In QSAR (Quantitative Structure-Activity Relationships) molecular descriptors are generally calculated considering all the atoms belonging to a molecule, or to an H-depleted representation of the molecule. As introduced in chapter 1 the traditional approach in order to calculate molecular descriptors on long peptides and proteins is often inapplicable due to the huge amount of atoms constituting big molecules such as polypeptides and proteins. We need to take into considerations that molecules usually studied with QSAR methodology and molecular descriptors are usually constituted of tens of atoms, rarely molecules with hundreds of atoms are studied introducing longer computational time. Anyway traditional

QSAR methodology can be applied to molecules with hundreds of atoms but proteins usually being constituted of thousands of atoms are not suitable to the traditional approach.

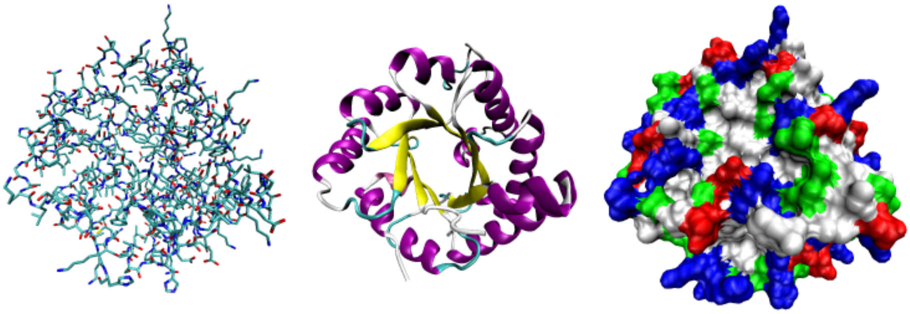


Figure 3.1: Three possible representations of the three-dimensional structure of the protein triose phosphate isomerase. Left: all-atom representation colored by atom type. Middle: simplified representation illustrating the backbone conformation, colored by secondary structure. Right: Solvent-accessible surface representation colored by residue type (acidic residues red, basic residues blue, polar residues green, nonpolar residues white).

3.2 Proteins and peptides representation: state of the art

The most used approach in order to characterise peptides is a method that model biological properties of small peptides as a function of amino acid principal properties. This approach has been introduced by Kidera et al. [Kidera *et al.* (1985)] that first coded the natural amino acids through 10 orthogonal factors derived from principal component analysis (PCA) of 188 reported properties. This line of research was followed by Hellberg et al. [Hellberg *et al.* (1987), Jonsson *et al.* (1989), Hellberg *et al.* (1991), Sandberg *et al.* (1998)] who developed principal properties, or z-scores, for each of 20 natural amino acids and for a series of unnatural ones. These were derived by carrying out principal components analysis (PCA) of numerous amino acid properties like HPLC retention times, pKas, NMR-derived properties, and other measurable variables related to hydrophobicity, size, and electronic features. The authors called the first three principal component scores of each amino acid its z_1, z_2, and z_3 scores or principal properties. These were interpreted to represent largely hydrophilicity, side chain bulk/molecular size, and electronic properties, respectively. The three principal properties for the amino acid in each position in a peptide were then used to construct models. With the three z-scales it is possible to numerically quantify the structural variation within a series of related peptides, by arranging the z-scales according to the amino acid sequence. The general formula for this approach can be written as the summation:

$$y = \sum_{i=1}^{N} \sum_{j=1}^{3} b_{ij} z_{ij} \tag{3.1}$$

Where N is the number of amino acids constituting the peptide sequences, b_{ij} is the regression coefficient and z_{ij} is the z-score associated to the i-th amino acid. Through z-scores and multivariate statistical regressions, successful models have been provided in QSAR studies for peptides active on oxytocin, bradykinin, and substance P receptors or in QSPR studies on sweetener peptides [Hellberg *et al.* (1986, 1987), Jonsson *et al.* (1989)].

Similar results were obtained by Cocchi et al. [Cocchi and Johansson (1993)] with another parametrization of amino acid side chains. In this approach the scores derived from a PCA of the interaction energies calculated with program GRID [Goodford (1985)], here defined as t-scores, turned out to be effective when applied in a QSAR study of a set of dipeptide ACE inhibitors. In 1995 Collantes et al. [Collantes and Dunn III (1995)] showed that two computable *3*-dimensional descriptors, Isotropic Surface Area (ISA) and Electronic Charge Index (ECI), may be usefully applied as side-chain descriptors. While ISA correlates well with z-score values and with Fauchere and Pliskas hydrophilicity scale [Fauchere and Pliska (1983)], ECI showed good correlation with amino acid free energy of vaporization [Wolfenden *et al.* (1981)].

In addition to these representation of peptides Zaliani et al. [Zaliani and E. (1999)] proposed new descriptors for the natural amino acids which have been derived from the principal component analysis (PCA) applied on the MS-WHIM 3D-description matrices [Todeschini *et al.* (1994), Bravi *et al.* (1997), Gancia *et al.* (2000)]. MS-WHIM indexes are a collection of 36 statistical indexes aimed at extracting and condensing steric and electrostatic 3D-properties of a molecule. These descriptors have been developed both on extended side-chain conformation and on rotamer library of natural amino acids.

These approaches produced good models for small peptides but has the disadvantage for those larger than a few amino acids than the number of terms to fit, and as a result, the number of peptides needed to construct a model is large [Siebert (2001, 2003)].

3.3 Proteins and peptides as sequence of α carbons

The descriptor-based approach could be compared to a peptide pictorial representation. As all pictorial representations of molecules are simplified versions of our current model of real structures (see Figure 3.1), analogously the descriptor-based representation is a simplified, but holistic, mathematical representation of the peptide. In both cases the peptide representation becomes clearer as much as our point of interest is simplified and highlighted in some way.

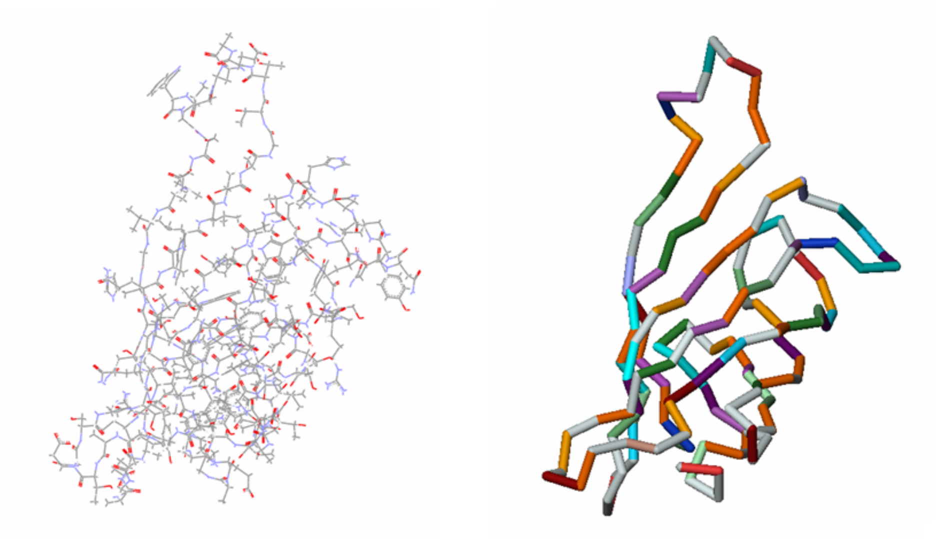


Figure 3.2: Two different representation of the same protein. On the left an atom-based representation displaying all the 1743 atoms, on the right a simplified representation visualising the 120 amino acids using the Cα stick coloured by residue type.

Due to the fact that the information content of a molecular descriptor depends on the kind of molecular representation and not only on the defined algorithm for its calculation, a good choice for the protein representation is indispensable. Considering a protein as a 3-dimensional molecular graph, the more immediate way to represent a protein is an atom based representation. Using this representation all the atoms belonging to a protein are considered. The atom-based representation raises one big problem, complex descriptors cannot be calculated on structures constituted of thousands of atoms. Another issue is that not necessarily all the information brought by the atom-based representation is directly connected to proteins' properties.

Considering that the physico-chemical properties of the amino acids are responsible for the 3D structure and the functionality of the protein and all amino acids share common structural features including an α-carbon to which an amino group, a carboxyl group, and a variable side chain are bonded; an amino acid based representation has been studied during this PhD thesis.

The amino acid based representation permits to reduce the complexity of the studied structures, in fact the number of amino acids in a protein is more than ten times lower than the number of atoms. A topological representation of a protein using an atom-based approach is a complex molecular graph where atoms are connected to the others by the molecular bonds, while the same protein using an amino acid-based approach is just a sequence of amino acid types. Analogously considering the 3-dimensional structure of the protein the atom-based approach brought to a 3-dimensional graph with thousands of vertexes while the amino acid based approach proposed in this PhD thesis considers only the Cα stick, the two different representations are showed in Figure 3.2.

In order to be able to calculate 3-dimensional descriptors, amino acids have to be characterised by three Cartesian coordinates. Being the α-carbon common to every amino acid the Cartesian coordinates of that atom have been chosen as the coordinates of the whole amino acid. In this way the amino acid structure can be coded as a sequence of α-carbon, each of them described by some properties of the corresponding isolated amino acid.

The amino acid properties chosen to characterise the isolated amino acids are presented in the next sections.

3.4 Amino acid characterisation

3.4.1 Introduction

The variety and specificity of protein 3-dimensional structures and biological functions are due to the combination of the 20 different amino acids as specified by the genetic code. The amino acids are the building blocks of proteins and peptides each having different characteristics in terms of the shape, the volume, and the chemical reactivity among others. Due to the fact that the molecular descriptors described in chapter 4 chosen to characterise proteins and peptides can be calculated in an unweighted way, that is considering every amino acid equal to the others, or weighting every amino acid by a descriptive property it has been necessary to chose some properties in order to characterise the 20 natural amino acids. One of the most comprehensive resource of amino acid properties freely available on line is the amino acid index database (AAindex http://www.genome.ad.jp/dbget/aaindex.html).

3.4.2 The amino acid index database (AAindex)

Amino acid index (AAindex) is a database of numerical indices representing various physicochemical, biochemical and statistical properties of amino acids and pairs of amino acids [Nakai *et al.* (1986), Tomii and Kanehisa (1996), Kawashima *et al.* (1999), Kawashima and Kanehisa (2000)]. AAindex database has been made publicly available by the Japanese GenomeNet database service.

AAindex consists of three sections:

1. AAindex1 for the amino acid index of 20 numerical values;

2. AAindex2 for the amino acid mutation matrix;

3. AAindex3 for the statistical protein contact potentials.

All data are derived from published literature.

The first section, AAindex1, has been considered as a possible resource in order to identify relevant properties of the 20 natural amino acids. This section (AAindex ver. 9.1) contains a list of 544 amino acid indices. Each entry consists of an accession number, a short description on the index, the reference information, and the numerical values for the property of 20 amino acids. In some instances the values are not reported for all 20 amino acids. The properties collected in the AAindex database have been divided in six major classes:

1. α and turn propensities;
2. β propensity;
3. amino acid composition;
4. hydrophobicity;
5. physicochemical properties;
6. other properties.

The first three classes can be considered as statistical properties of the amino acids, while the fourth and the fifth classes include physicochemical properties. Both, statistical and physicochemical properties have been considered in order to build different weighting schemes for the twenty natural amino acids.

3.5 Choice of the characterising amino acid properties

3.5.1 Introduction

In order to be able to evaluate how different properties highlight different kind of information, three weighting schemes have been defined. A weighting scheme is a collection of properties of the 20 natural amino acids. Molecular descriptors have been calculated using the weighting schemes separately. In the next sections the proposed weighting schemes are presented.

The first two weighting schemes are made of five different properties each. The first one collects five physicochemical properties while the second one is an array of statistical properties of the 20 natural amino acids. Finally a third weighting scheme is proposed. The last weighting scheme has been introduced calculating three different WHIM global descriptors [Todeschini *et al.* (1994, 1995, 1996b,a), Todeschini and Gramatica (1997c,a,b), Todeschini *et al.* (1997)] calculated on the isolated structure of the 20 natural amino acids.

3.5.2 Physicochemical weights

Depending on the polarity of the side chain, amino acids vary in their hydrophilic or hydrophobic character. These properties are important in protein structure and protein-protein interactions. The importance of the physical properties of the side chains comes from the influence this has on the amino acid residues' interactions with other structures, both within a single protein and between proteins.

Table 3.1: Weighting scheme values for the 20 AAs. mw (molecular Weight, scaled), p (polarity, scaled), hyb (hydrophobicity, scaled), ras (residue accessible surface area in folded protein, scaled) and hyl (hydrophilicity scale).

Amino acid	1-letter	mw	p	hyb	ras	hyl
Ala	A	0.651	0.973	0.614	0.57	0.78
Arg	R	1.272	1.261	0.6	2.052	1.58
Asn	N	0.965	1.393	0.063	1.437	1.2
Asp	D	0.972	1.562	0.466	1.14	1.35
Cys	C	0.885	0.661	1.072	0.433	0.55
Glu	E	1.068	1.261	0	1.619	1.19
Gln	Q	1.075	1.477	0.473	1.117	1.45
Gly	G	0.548	1.081	0.071	0.525	0.68
His	H	1.133	1.249	0.614	0.981	0.99
Ile	I	0.958	0.625	2.222	0.41	0.47
Leu	L	0.958	0.589	1.531	0.525	0.56
Lys	K	1.068	1.357	1.157	2.212	1.1
Met	M	1.09	0.685	1.178	0.707	0.66
Phe	F	1.207	0.625	2.025	0.547	0.47
Pro	P	0.841	0.961	1.954	1.14	0.69
Ser	S	0.768	1.105	0.049	1.003	1
Thr	T	0.87	1.033	0.049	1.072	1.05
Trp	W	1.492	0.649	2.66	0.73	0.7
Tyr	Y	1.324	0.745	1.884	1.368	1
Val	V	0.856	0.709	1.319	0.41	0.51

The distribution of hydrophilic and hydrophobic amino acids determines the tertiary structure of the protein, and their physical location on the outside structure of the proteins influences their quaternary structure. For example, soluble proteins have surfaces rich with polar amino acids like serine and threonine, while integral membrane proteins tend to have outer ring of hydrophobic amino acids that anchors them into the lipid bilayer, and proteins anchored to the membrane have a hydrophobic end that locks into the membrane. Similarly, proteins that have to bind to positively-charged molecules have surfaces rich with negatively charged amino acids like glutamate and aspartate, while proteins binding to negatively-charged molecules have surfaces rich with positively charged chains like lysine and arginine.

Moreover it is generally accepted that in distantly related proteins, structure is more conserved than sequence. Proteins that have diverged beyond detectable sequence similarity still retain the architecture and topology of their ancestral fold, in the known protein structures there are several families within which the molecules maintain the same basic folding pattern over ranges of sequence homology from near-identity down to below 20%. This means that structural details are not maintained, it is function that is maintained.

In both closely and distantly related proteins the general response to mutation is conformational change. Variations in conformation in families of homologous proteins that retain a common function reveal how the structures accommodate changes in amino acid sequence. Residues active in function are resistant to mutation because changing them would interfere, explicitly and directly, with function. It is the ability of protein structures to accommodate mutations in non-functional residues that permits a large amount of apparently no adaptive change to occur. Surface residues not involved in function are usually free to mutate. Loops on the surface can often accommodate changes by local refolding, provided that they are not involved directly in function This behaviour is due to the fact that protein may well have similar structures and functions due to physicochemical reasons.

Starting from the assumption that the physicochemical properties of the amino acids are responsible for the 3D structure and the functionality of the protein the first weighting scheme has been defined collecting five different physicochemical properties from the amino acid index database.

The selected indices are:

1. molecular weight [Fasman (1976)] (FASG760101);
2. polarity [Grantham (1974)] (GRAR740102);
3. hydrophobicity [Jones (1975)] (JOND750101);
4. residue accessible surface area in folded protein [Chothia (1976)] (CHOC76010);
5. hydrophilicity scale [Kuhn *et al.* (1995)] (KUHL950101).

Once selected, the five indices have been separately scaled in order to obtain values with mean equal to one. Scaled index values are showed in Table 3.1. Hydrophilicity has not been scaled due to the fact that this property is already scaled.

The physicochemical weighting scheme has been used to calculate molecular descriptors in order to perform a sensitivity analysis of the proposed methodology, see chapter 7, they have been applied also on the cluster analysis of two different protein folds in chapter 8 and in a regression analysis in order to predict two biological properties in chapter 9.

3.5.3 Statistical weights

The structure and the sequence of many proteins is currently known. This information can be used in order to define some statistical properties related to the occurrence of the 20 natural amino acids in different kind of proteins or different secondary structure elements (SSEs), namely, α helices, β strands, structural turns, and loops [Kabsch and Sander (1983)].

Secondary structure conservation has been studied in structural alignments of protein families and SSE substitution matrices have been created [Mizuguchi and Blundell (2000)]. The conservation of SSEs has been also studied in some specific protein families [Cygler *et al.* (1993)]; protein loops and their flanking regions have been found to be conserved to the same extent in an analysis of a large set of proteins [Liu *et al.* (2002)].

Protein structures can be divided into four major structural classes, according to their secondary structure content and arrangement (SCOP [Murzin *et al.* (1995), LoConte *et al.* (2002), Andreeva *et al.* (2004)]). There are two homogeneous classes and two heterogeneous classes. The homogenous classes consists of structures containing mainly α helices (termed all alpha) or containing mainly β strands (all beta). The two heterogeneous classes comprise both α helices and β strands. The alpha/beta class consists of mainly parallel β sheets (beta-alpha-beta units), and the alpha+beta class that consists of mainly antiparallel β sheets (segregated α and β regions) [Murzin *et al.* (1995), LoConte *et al.* (2002), Andreeva *et al.* (2004)]. Each class differs in its secondary structure content.

Proteins with similar sequences adopt similar structure [Chothia and Lesk (1986), Doolittle (1981)]. However, similar structures can have less than 12% sequence identity [Murzin *et al.* (1995), LoConte *et al.* (2002), Andreeva *et al.*

Table 3.2: Statistical weighting scheme values for the 20 AAs. rf_bs (relative frequency in beta-sheet), rfo (relative frequency of occurrence), rm (relative mutability), rf_ah (relative frequency in alpha-helix) and rf_rt (relative frequency in reverse-turn).

Amino acid	1-letter	**rf_bs**	**rfo**	**rm**	**rf_ah**	**rf_rt**
Ala	A	0.3059	0.8182	0.8152	0.8105	0.2566
Arg	R	0.4118	0.4805	0.6304	0.4632	0.3224
Asn	N	0.1412	0.3766	0.8587	0.4000	0.5855
Asp	D	0.0941	0.4935	0.6630	0.5474	0.6711
Cys	C	0.1176	0.0779	0.2065	0.6211	0.2697
Glu	E	0.1882	0.3506	0.6413	0.7895	0.3816
Gln	Q	0.1294	0.6234	0.5652	0.9684	0.4013
Gly	G	0.3294	0.7792	0.2717	0.0421	0.8224
His	H	0.5176	0.1169	0.7174	0.7368	0.1974
Ile	I	0.9529	0.5065	0.8478	0.4737	0.0789
Leu	L	0.4471	1.0000	0.3152	0.8211	0.1316
Lys	K	0.1529	0.5844	0.5109	0.7474	0.3750
Met	M	0.3882	0.1299	0.7391	1.0000	0.0000
Phe	F	0.8000	0.3377	0.2826	0.5789	0.1250
Pro	P	0.0000	0.4805	0.3587	0.0000	1.0000
Ser	S	0.3647	0.7143	1.0000	0.3158	0.6184
Thr	T	0.6706	0.5844	0.8913	0.3158	0.4211
Trp	W	0.5882	0.0000	0.0000	0.4947	0.2368
Tyr	Y	0.7176	0.2338	0.2717	0.2105	0.4342
Val	V	1.0000	0.6753	0.7935	0.4105	0.0526

(2004), Holm and Sander (1996), Brenner *et al.* (1996), Rost (1997)]. Most amino acids within a protein can thus be changed without affecting its structure, including the secondary structure [Rost (1999)]. Previous experiments have shown that both helices and strands can undergo numerous mutations and still keep their secondary structure - either β-strand or α helix - and also maintain structural stability [He *et al.* (2004), Heinz *et al.* (1994), Blaber *et al.* (1995)].

Due to these considerations the second weighting scheme studied during this PhD thesis is not based upon physicochemical properties but collects five different indices taken from the AAindex database classes: α and turn propensities, β propensity and amino acid composition.

The weighting scheme proposed in this section has been called the statistical weighting scheme, the five selected indices are:

1. relative frequency in beta-sheet [Prabhakaran (1990)] (PRAM900103);

2. relative frequency of occurrence [Jones *et al.* (1992)] (JOND920101);

3. relative mutability [Jones *et al.* (1992)] (JOND920102);

4. relative frequency in alpha-helix [Prabhakaran (1990)] (PRAM900102);

5. relative frequency in reverse-turn [Prabhakaran (1990)] (PRAM900104).

Index values are showed in Table 3.2.

Due to the considerations that statistical weights strongly depend on protein related information a preliminary evaluation of this weighting scheme has been performed on the cluster analysis of two different protein folds, a detailed description of this application can be found in chapter 8.

3.5.4 WHIM weights

Aside from the twenty standard amino acids, there is a vast number of "nonstandard amino acids". Two of these can be encoded in the genetic code, but are rather rare in proteins. Selenocysteine is incorporated into some proteins at a UGA codon, which is normally a stop codon [Driscoll and Copeland (2003)]. Pyrrolysine is used by some methanogenic archaea in enzymes that they use to produce methane. It is coded for with the codon UAG [Krzycki (2005)].

Nonstandard amino acids often occur as intermediates in the metabolic pathways for standard amino acids - for example ornithine and citrulline occur in the urea cycle, part of amino acid catabolism [Curis *et al.* (2005)].

Nonstandard amino acids are usually formed through modifications to standard amino acids. For example, homocysteine is formed through the transsulfuration pathway or by the demethylation of methionine via the intermediate metabolite *S*-adenosyl methionine [Brosnan and Brosnan (2006)], while dopamine is synthesized from l-DOPA, and hydroxyproline is made by a post-translational modification of proline [Kivirikko and Pihlajaniemi (1998)].

In order to be able to characterise not only the twenty natural amino acids but also the nonstandard amino acids it has been necessary to identify a weighting scheme not depending from the amino acid index database.

The adopted weighting scheme has been obtained calculating three different global Weighted Holistic Invariant Molecular descriptors (WHIM) descriptors [Todeschini *et al.* (1994, 1995, 1996b,a), Todeschini and Gramatica (1997c,a,b), Todeschini *et al.* (1997), Todeschini and Gramatica (1998)] from the molecular

Table 3.3: Weighting scheme values for the 20 AAs. Am (WHIM global dimension descriptor / weighted by atomic masses, scaled), Km (WHIM shape descriptor / weighted by atomic masses). Dm (WHIM global density descriptor / weighted by atomic masses).

Amino acid	**1-letter**	**Am**	**Km**	**Dm**
Ala	A	0.3634	0.4430	0.2330
Arg	R	1.9266	0.7980	0.3130
Asn	N	0.9274	0.4970	0.2960
Asp	D	0.8575	0.4290	0.3700
Cys	C	0.6683	0.4990	0.2530
Glu	E	0.9970	0.5840	0.3810
Gln	Q	1.1128	0.4040	0.3260
Gly	G	0.2343	0.5420	0.3220
His	H	1.0631	0.7590	0.2740
Ile	I	0.9845	0.5820	0.2660
Leu	L	1.1486	0.5210	0.3370
Lys	K	1.5369	0.8120	0.3340
Met	M	1.0385	0.4610	0.2940
Phe	F	1.3731	0.5790	0.2710
Pro	P	0.5536	0.4870	0.2910
Ser	S	0.4656	0.3910	0.2810
Thr	T	0.6918	0.4850	0.3070
Trp	W	2.3415	0.6410	0.2970
Tyr	Y	1.6385	0.6620	0.2840
Val	V	0.7066	0.5100	0.2980

structure of the isolated amino acids. Three WHIM descriptors (*Am - global dimension descriptor, Km - global shape descriptor, Dm - global density descriptor*) have been calculated using the classical atom based approach describing every atoms belonging to the amino acids using the atomic mass. WHIM descriptors are built in such a way as to capture relevant molecular 3D information regarding molecular size, shape, symmetry and atom distribution with respect to invariant reference frames.

They are divided into two main classes: directional WHIM descriptors and global WHIM descriptors.

Directional WHIM descriptors are calculated as some univariate statistical indices on the projections of the atoms along each individual principal axis, while the global WHIMs are directly calculated as a combination of the former, thus simultaneously accounting for the variation of molecular properties along the three principal directions in the molecule. In this case, any information individually related to each principal axis disappears and the description is related only to a global view of the molecule.

Within the WHIM approach, a molecule is seen as a configuration of points (the atoms) in the three-dimensional space defined by the Cartesian axes (x, y, z). In order to obtain a unique reference frame, principal axes of the molecule are calculated. Then, projections of the atoms along each of the principal axes are performed and their dispersion and distribution around the geometric centre are evaluated.

Indeed, the algorithm consists of calculating the eigenvalues and eigenvectors of a weighted covariance matrix of the centred Cartesian coordinates of a molecule, obtained from different weighting schemes w for the atoms:

$$s_{qq'} = \frac{\sum_{i=1}^{nAT} w_i(q_i - \overline{q})(q'_i - \overline{q'})}{\sum_{i=1}^{nAT} w_i} \tag{3.2}$$

where $s_{qq'}$ is the weighted covariance between the atomic coordinates q and q' (q, $q' = [x,\ y,\ z]$), nAT is the number of atoms, w_i the atomic property (that is the atomic mass in our case), q_i and q'_i represent the coordinates of the i-th atom, and the corresponding average value.

A fundamental role in the WHIM descriptor calculation is played by the eigenvalues λ_1, λ_2 and λ_3 of the weighted covariance matrix of the molecule atomic coordinates. Each eigenvalue represents a dispersion measure (i.e., the weighted variance) of the projected atoms along the considered principal axis, thus accounting for the molecular size along that principal direction. Relationships among the eigenvalues are used to describe the molecular shape. For example, for an ideal straight molecule both λ_2 and λ_3 are equal to zero and the global shape Kw is equal to 1 (maximum value); for an ideal spherical molecule all three eigenvalues are equal to 1/3 and Kw is 0.

Exploiting the new coordinates t_k of the atoms along the principal axes, the atom distribution and density around the molecule centre are evaluated by an inverse function of the kurtosis k ($\eta = 1/k$). Low values of the kurtosis are obtained when the atom projections assume opposite values with respect to the centre. When an increasing number of atom projections are within the extreme projections along a principal axis, the kurtosis value increases (i.e., kurtosis equal to 1.8 for a uniform distribution of points, to 3.0 for a normal distribution). When the kurtosis value tends to infinity the corresponding η value tends to zero.

The three WHIM descriptors chosen to describe the amino acids are a global dimension descriptor Am, calculated as:

$$Am = \lambda_1 + \lambda_2 + \lambda_3 \tag{3.3}$$

The second one (Km) is a global shape descriptor calculated as:

$$Km = \frac{3}{4}\sum_{k=1}^{3} | \frac{\lambda_k}{\sum_{k=1}^{3} \lambda_k} - \frac{1}{3} | \tag{3.4}$$

The last one (Dm) is a global density descriptor calculated as:

$$Dm = \eta_1 + \eta_2 + \eta_3 \tag{3.5}$$

The values of *Am, Km, Dm* for the twenty natural amino acids is reported in table Table 3.3. The WHIM weighting scheme has been applied to regression analysis in chapter 9 in order to predict two biological properties.

chapter 4

Molecular Descriptors

"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." [Todeschini and Consonni (2000)].

4.1 Introduction

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, food sciences, health research and quality control, being obtained when molecules are transformed into a molecular representation allowing some mathematical treatment. Many molecular descriptors have been proposed derived from different theories and approaches with the aim of predicting biological and physicochemical properties of molecules [Todeschini and Consonni (2000)].

The information content of a molecular descriptor depends on the kind of molecular representation that is used and on the defined algorithm for its calculation (Figure 4.1). There are simple molecular descriptors derived by counting some atom-types or structural fragments in the molecule, other derived from algorithms applied to a topological representation (molecular graph) and usually called topological or 2D-descriptors, and there are molecular descriptors derived

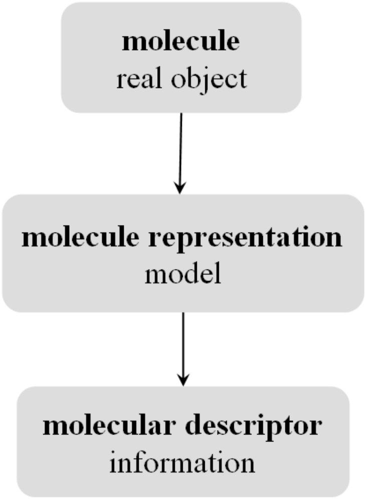


Figure 4.1: Molecular descriptors are numbers able to extract small pieces of chemical information from the different molecule representations

from a geometrical representation that are called geometrical or 3D-descriptors.

All the molecular descriptors must contain, to varying extents, chemical information, must satisfy some basic invariant properties and general requirements, and must be derived from well-established procedures which enable molecular descriptors to be calculated for any set of molecules. It is obvious almost trivial - that a single descriptor or a small number of descriptors cannot wholly represent the molecular complexity or model all the physico-chemical responses and biological interactions. As a consequence, although we must get used to living with approximate models, we have to keep in mind that “approximate”is not a synonym of “useless”.

The field of molecular descriptors is strongly interdisciplinary and involves a mass of different theories. For the definition of molecular descriptors, a knowledge of algebra, graph theory, information theory, computational chemistry, theories of organic reactivity and physical chemistry is usually required, although at different levels. For the use of the molecular descriptors, a knowledge of statistics, chemometrics, and the principles of the QSAR/QSPR approaches is necessary in addition to the specific knowledge of the problem (Figure 4.2).

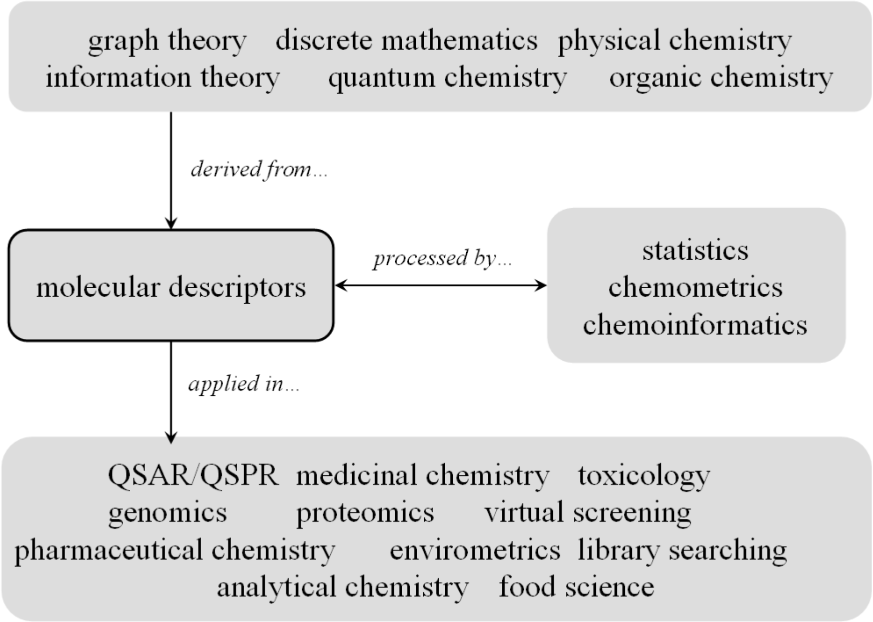


Figure 4.2: Key role of molecular descriptors in scientific research.

4.2 Constitutional descriptors

Constitutional descriptors are the most simple and commonly used descriptors, reflecting the molecular composition of a compound without any information about its molecular geometry.

The constitutional descriptors proposed have been introduced during this PhD thesis. The number of molecular descriptors calculated depends on the number of amino acid properties w used to weight the amino acids. The calculated constitutional descriptors are:

- the number of amino acids belonging to the peptide or protein sequence (one calculated descriptor for each sequence);

- the sum of the k-th property wk along the peptide sequence; i.e. the sum of molecular weight, polarity, hydrophobicity, residue accessible surface area in folded protein and hydrophilicity if we are considering the physico-chemical weighting scheme.

$$Wk_{sum} = \sum_{i=1}^{N} wk_i \tag{4.1}$$

where k is the considered property, N is the number of amino acids belonging to the considered sequence and wk_i is the considered property for the i-th amino acid. The number of calculated descriptors for each sequence is K, where K is the number of properties in the weighting scheme;

- the average sum of the k-th property wk along the peptide sequence; i.e. the average sum of molecular weight, polarity, hydrophobicity, residue accessible surface area in folded protein and hydrophilicity if we are considering the physico-chemical weighting scheme.

$$Wk_{asum} = \frac{\sum_{i=1}^{N} wk_i}{N} \tag{4.2}$$

where k is the considered property, N is the number of amino acids belonging to the considered sequence and wk_i is the considered property for the i-th amino acid. The number of calculated descriptors for each sequence is K, where K is the number of properties in the weighting scheme;

- the absolute frequency of the 20 different amino acids in each sequence i.e. absolute frequency of alanines, arginine, asparagine, aspartic acid, cysteine, glutami acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine and valine (20 calculated descriptors for each sequence);

- the relative frequency of the 20 different amino acids in each sequence i.e. relative frequency of alanines, arginine, asparagine, aspartic acid, cysteine, glutami acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine and valine (20 calculated descriptors for each sequence).

Considering the three weighting schemes proposed in chapter 3, 51 constitutional descriptors have been calculated while using the physico-chemical weights or the statistical weights, both weighting schemes are constituted by 5 different amino acid properties. Considering the WHIM weighting scheme 47 constitutional descriptors have been calculated.

Although their simplicity there are several cases where constitutional descriptors can be successfully used. In fact some peptide properties result from the proportion of only one or a few amino acids and it is possible to model properties from the number of the relevant amino acids in the protein [Siebert (2001, 2003)]. One of the situations where the proportion of amino acids of different types rather than a precise sequence is thought to impact protein properties is in what are called by food chemists functional properties. Various functional properties have been listed by different authors and include solubility, wettability, gelation, fat binding, water binding, emulsifying capacity, and foam, film, and glass formation [Pomeranz (1991), Phillips *et al.* (1994), Damodaran (1995), Hettiarachchy and Ziegler (1994), Fligner and Mangino (1991)]. A number of physicochemical properties (hydrophobicity, melting point, etc.) have been related to the proportions of individual amino acids or particular classes of amino acids (e.g. acidic, basic, hydrophilic, hydrophobic, aromatic, etc.) in a protein [Phillips *et al.* (1994)]. A number of the functional properties have in turn been related to protein physicochemical properties [Phillips *et al.* (1994), MacRitchie (1992)]. For example, hydrophobicity, either in a domain or of an entire protein, is associated with foaming, gel formation, and binding of nonpolar flavor compounds [Phillips *et al.* (1994), Nakai *et al.* (1986)].

4.3 Topological descriptors

Topological descriptors, as the name suggests, consider the topology of a molecule. These descriptors are numerical quantifiers of molecular topology [Todeschini and Consonni (2000)] that are mathematically derived in a direct and unambiguous manner from the structural graph of a molecule, usually an H-depleted molecular graph. Topological descriptors characterise structural features of the molecule such as size, shape, symmetry, branching and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. Since topological descriptors consider the molecule as a mathematical graph, a number of these descriptors are simply various graph invariants or other functions of the molecular graph.

4.3.1 2D autocorrelation descriptors

2D autocorrelation descriptors are molecular descriptors which describe how a considered property is distributed along a topological molecular structure.

This set includes:

- autocorrelations ATS (i.e. Autocorrelation of a Topological Structure) proposed by Moreau and Broto [Broto *et al.* (1984a,b,c)];
- autocorrelations MATS calculated by the Moran coefficient [Moran (1950)];
- autocorrelations GATS calculated by the Geary coefficient [Geary (1954)].

Autocorrelation descriptors combine chemical information given by property values in specified molecule regions and structural information. These are based on a conceptual dissection of the molecular structure and the application of an autocorrelation function to molecular properties measured in different molecular regions.

The Broto-Moreau autocorrelation $ATSkw$, w being the amino acid property used to weight the peptide sequence and k the lag, is evaluated by considering separately all the contributions of each different path length (lag) in the peptide sequence, as collected in the topological distance matrix. In other words, the total spatial autocorrelation at lag k $ATSkw$ is obtained by summing all the products w_ixw_j of all the pairs of amino acid i and j, for which the topological distance equals the lag as:

$$ATSkw = \sum_{i=1}^{nSK-1} \sum_{j=i+1}^{nSK} w_i w_j \delta_{ij} \tag{4.3}$$

where nSK is the number of amino acids belonging to the peptide and δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ if $\delta_{ij} =$ k, zero otherwise, δ_{ij} being the topological distance between two considered amino acid).

The Moran autocorrelation $MATSkw$, w being the amino acid property used to weight the peptide sequence and k the lag, is calculated by applying the Moran coefficient to the amino acid sequence:

$$MATSkw = \frac{\frac{1}{\Delta}\sum_{i=1}^{nSK} \sum_{j=1}^{nSK} \delta_{ij}(w_i - \overline{w})(w_j - \overline{w})}{\frac{1}{nSK}\sum_{i=1}^{nSK}(w_i - \overline{w})^2} \tag{4.4}$$

where w_i is any amino acid property, $\overline{w}$ is its average value on the peptide, nSK is the number of amino acids, δ_{ij} is the Kronecker delta (δ_{ij} = 1 if δ_{ij} = k, zero otherwise, δ_{ij} being the topological distance between two considered amino acids). Δ is the sum of the Kronecker deltas, i.e. the number of amino acid pairs at distance equal to k. Moran coefficient usually takes value in the interval [-1,+1]. Positive spatial autocorrelation corresponds to positive values of the coefficient whereas negative spatial autocorrelation produces negative values.

The Geary autocorrelation $GATSkw$, w being the atomic property used to weight the peptide sequence and k the lag, is calculated by applying the Geary coefficient to the amino acid sequence:

$$GATSkw = \frac{\frac{1}{2\Delta}\sum_{i=1}^{nSK}\sum_{j=1}^{nSK}\delta_{ij}(w_i - w_j)^2}{\frac{1}{nSK-1}\sum_{i=1}^{nSK}(w_i - \overline{w})^2} \tag{4.5}$$

where w_i is any amino acid property, $\overline{w}$ is its average value on the peptide, nSK is the number of amino acids, δ_{ij} is the Kronecker delta (δ_{ij} = 1 if δ_{ij} = k, zero otherwise, δ_{ij} being the topological distance between two considered amino acids). Δ is the sum of the Kronecker deltas, i.e. the number of amino acid pairs at distance equal to k. Geary coefficient is a distance-type function varying from zero to infinite. Strong spatial autocorrelation produces low values of this index; moreover, positive autocorrelation translates in values between 0 and 1 whereas negative autocorrelation produces values larger than 1; therefore, the reference ”no correlation” is coefficient value equal to 1.

To obtain uniform-length descriptors for a set of peptides, 2D-autocorrelation descriptors are calculated using a lag from 1 to 8. Autocorrelations at lag 0 are not provided being a simple sum of the squares of amino acid properties. Finally, to avoid too large numbers, a logarithmic transformation of the Moreau-Broto autocorrelation values (ATS) is performed as *ln(1 + value)*.

4.4 Geometrical descriptors

Geometrical descriptors characterise the shape and extent of the molecule in terms of its 3-dimensional Cartesian coordinates. As a result accurate coordinates are required and so the structure must be geometry optimised before these descriptors can be calculated [Todeschini and Gramatica (1998)].

4.4.1 Weighted Holistic Invariant Molecular (WHIM) descriptors

WHIM descriptors (Weighted Holistic Invariant Molecular descriptors) are geometrical descriptors based on statistical indices calculated on the projections of the atoms along principal axes [Todeschini *et al.* (1994, 1995, 1996b,a), Todeschini and Gramatica (1997c,a,b), Todeschini *et al.* (1997), Todeschini and Gramatica (1998)].

They are built in such a way as to capture relevant molecular 3D information regarding molecular size, shape, symmetry and atom distribution with respect to invariant reference frames. The algorithm consists in performing a Principal Component Analysis (PCA) on the centred cartesian coordinates of a molecule by using a weighted covariance matrix obtained from different weighting schemes for the amino acids. Depending on the kind of weighting scheme, different covariance matrices and different principal axes are obtained. Thus, the WHIM approach can be viewed as a generalisation searching for the principal axes with respect to a defined amino acid property, the weighting scheme.

For each weighting scheme, a set of statistical indices is calculated on the atoms projected onto each principal component. The invariance to translation of the calculated parameters is due to the centering of the atomic coordinates and the invariance to rotation is due to the uniqueness of the PCA solution.

WHIM descriptors are divided into two main classes:

1. directional WHIM descriptors;
2. global WHIM descriptors.

Directional WHIM descriptors are calculated as some univariate statistical indices on the projections of the amino acids along each individual principal axis. Directional WHIM descriptors can be divided in four groups:

1. directional WHIM size indices;
2. directional WHIM shape indices;
3. directional WHIM simmetry indices;
4. directional WHIM density indices.

The first group of descriptors consists of the directional WHIM size indices defined as the eigenvalues λ_1, λ_2 and λ_3 of the weighted covariance matrix of the molecule amino acid α-carbon coordinates; they account for the molecular size along each principal direction. The second group consists of the directional WHIM shape descriptors ϑ_1, ϑ_2 and ϑ_3, calculated as eigenvalues ratios and related to molecular shape. The third group of descriptors consists of the directional WHIM symmetry indices γ_1, γ_2 and γ_3 calculated as mean information content [Klir and Folger (1988)] on the symmetry along each component with respect to the centre of the scores. Finally, the fourth group of descriptors consists of the directional WHIM density indices which are related to the amino acids distribution and density around the origin and along the principal axes.

The global WHIMs are directly calculated as a combination of the directional WHIM descriptors, thus simultaneously accounting for the variation of molecular properties along the three principal directions in the molecule. In this case, any information individually related to each principal axis disappears and the description is related only to a global view of the molecule.

WHIM descriptors are invariant to translation due to the centring of the atomic coordinates and invariant to rotation due to the uniqueness of the principal axes, thus resulting free from any prior alignment of molecules. In many cases, size descriptors can play, in modelling, a significant role independently of the measured directions, allowing simpler models. Thus, in view of the importance of this quantity, a group of descriptors of the global dimension of a molecule is considered in three different ways based on size, shape, symmetry and density analogously to the directional WHIM descriptors.

4.4.2 GEometry, Topology, and Atom-Weights AssemblY (GETAWAY) descriptors

The GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors [Consonni and Todeschini (2001), Consonni *et al.* (2002a,b)] have been proposed as chemical structure descriptors derived from a representation of molecular structure by the Molecular Influence Matrix (MIM), denoted by $\mathbf{H}$ and defined as the following:

$$\mathbf{H} = \mathbf{M}\left(\mathbf{M}^{\mathrm{T}}\mathbf{M}\right)^{-1}\mathbf{M}^{\mathrm{T}} \tag{4.6}$$

where $\mathbf{M}$ is the molecular matrix consisting of the centred Cartesian coordinates x, y, z of the molecule atoms or amino acids in a chosen conformation, and the superscript T refers to the transposed matrix. Atomic coordinates are assumed to be calculated with respect to the geometrical centre of the molecule in order to obtain translation invariance. The molecular information matrix is a symmetric matrix and shows rotational invariance with respect to the molecule coordinates, thus resulting independent of molecule alignment.

The diagonal elements h_{ii} of the molecular influence matrix, called leverages, range from 0 to 1 and encode atomic information related to the "influence"of each molecule atom in determining the whole shape of the molecule; in effect, mantle atoms always have higher h_{ii} values than atoms near the molecule centre. Moreover, the magnitude of the maximum leverage in a molecule depends on the size and shape of the molecule. As derived from the geometry of the molecule, leverage values are effectively sensitive to significant conformational changes and to the bond lengths that account for amino acids types. Each off-diagonal element h_{ij} represents the degree of accessibility of the j-th amino acid to interactions with the i-th amino acid, or, in other words, the attitude of the two considered amino acids to interact with each other. A negative sign for the off-diagonal elements means that the two amino acids occupy opposite molecular regions with respect to the centre, hence the degree of their mutual accessibility should be low.

The influence/distance matrix **R** has been derived from the molecular influence matrix H as the following:

$$[R_{ij}] = [\frac{\sqrt{h_{ii}h_{jj}}}{r_{ij}}]_{ij} \qquad i \neq j \tag{4.7}$$

where h_{ii} and h_{jj} are the leverages of the two considered amino acids, and r_{ij} is their distance. The diagonal elements of the matrix **R** are zero. The squared root product of the leverages of two amino acids is divided by their distance in order to make less significant contributions from pairs of amino acids far apart, according to the basic idea that interactions between amino acids in the molecule decreases as their distance increases.

A first set of the GETAWAY descriptors has been derived by applying some traditional matrix operators and concepts of information theory both to the molecular influence matrix **H** and the influence/distance matrix **R**. Most of these descriptors are simply calculated only by the leverages used as the amino acid weighting.

Another set of GETAWAY descriptors is based on the spatial autocorrelation formulas, weighting the molecule amino acids by properties w together with 3D information encoded by the elements of the molecular influence matrix $\mathbf{H}$ and influence/distance matrix $\mathbf{R}$.

chapter 5

Chemometric Methods

Analytical chemical systems can be described by means of tables (matrices) in which each row corresponds to a sample and each column to a variable describing the system. This is the typical input for chemometric methods, which consider all the variables at the same time and extract information in a global way. The chemometric methods used during the applications of this thesis are collected and briefly explained in the present chapter.

5.1 Data structure

Traditionally, in chemometrics, $\mathbf{X}$ denotes the data matrix, while the number of rows (samples) and columns (variables) is usually indicated by n and p respectively. Each entry of this matrix, x_{ij}, represents the value of the j-th variable for the i-th sample. Other qualitative information regarding the samples can be added to the data matrix, in order to make the results more readable, but only the data matrix $\mathbf{X}$ is considered during the algorithms.

Depending on the applied chemometric method, some other information can be needed in order to develop a multivariate model: when classification or regression techniques are used, a response vector (or matrix) $\mathbf{Y}$ is used during the calculations. This vector (matrix) contains the qualitative or quantitative responses to be modelled and has usually dimensions n times 1, i.e. each entry y_i

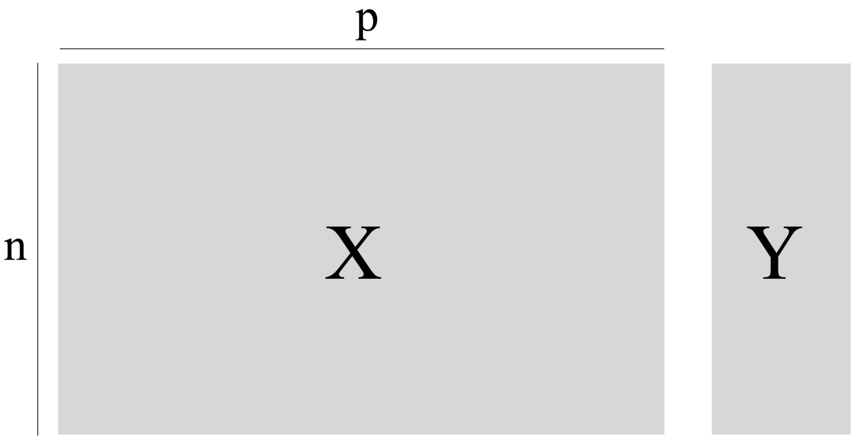


Figure 5.1: Typical representation of multivariate data. Where **X** is the data matrix, the descriptor values, and **Y** the response matrix.

of the vector represents the value of the response for the i-th sample. If more responses are considered in the same model, **Y** has dimensions n times r, where r is the number of considered responses. In Figure 5.1 a schematic representation of a multivariate data structure is shown.

In all the applications of this thesis, each row of **X** represent a molecule (the sample), that can be a peptide or a protein represented by the molecular descriptor values (the variables), and each column of **X** represent a single calculated molecular descriptor, describing the molecules.

5.2 Principal Component Analysis

Principal Component Analysis (PCA) is the most common method used to display the structure of the multivariate data [Wold *et al.* (1987), Kvalheim (1987)]. PCA is a well-known chemometric technique, which projects the data in a reduced hyperspace, defined by the principal components. These are linear combinations of the original variables, with the first principal component having the largest variance, the second principal component having the second-largest variance, and so on. In this way it is possible to retain a number of components lower than the number of original variables, i.e. it is possible to reduce the data dimension:

the number of components to be retained can be chosen on the basis of different parameters, linked to the variance explained by each principal component.

5.3 Sensitivity analysis

Sensitivity analysis is the study of how the variation in the output of a model (numerical or otherwise) can be apportioned, qualitatively or quantitatively, to different sources of variation. A sensitivity analysis has been performed in chapter 7 in order to evaluate the capability of different descriptor blocks to discriminate among different mutations on the same peptide sequence.

5.4 Variable selection techniques

Variable selection techniques can be used in order to improve chemometric models: these techniques are in fact able to retain and preserve only the variables, which contain significant information for a specific task. Moreover, the increase of dimensions and complexity of datasets and the decrease in time-consumption in algorithms support approaches based on variable selection techniques.

5.4.1 All subset selection

The all subset selection method is the most simple variable selection method: this technique searches all the possible models by using all the available combinations of variables. Usually, an exhaustive search of all the possible solutions is not feasible: in fact, if there is a total number of p variables, the number N of all the possible combinations is:

$$N = \frac{p!}{(p-c)!c!} \tag{5.1}$$

where c is the number of considered variables for each combination. This means that considering 50 variables ($p = 50$) and selecting just 5 variables ($c = 5$), the total number of combinations is N = 50!/((50-5)!5!) = 2118760, i.e. a huge number of models should be computed. Consequently, this selection method is applicable only with a low number of variables.

5.4.2 Forward selection

Forward Variable Selection is a simple selection technique, which starts with no variables and adds one variable at a time to the chemometric model: the inclusion of a variable is based on the optimisation of a chosen parameter [Jennrich (1977)] that depends on the selection task, e.g. a classification quality parameter, such as the number of errors, or a regression parameter, such as the response residuals.

Forward Selection can depend on the first selected variables, since all the others are added to the model when these variables are still present and consequently the new variables can just contribute to solve marginal modelling fittings. On the other hand this method is usually faster and less time-consuming than other classification techniques, such as Genetic Algorithms, which explore in a more complete way the available information and searches for the best solution with an higher number of possibilities, but, as a consequence, are more time-consuming.

5.4.3 Genetic Algorithms

Genetic Algorithms (GAs) select subsets of variables that maximise the predictive power of multivariate models and perform this selection by considering populations of models generated with an evolution process and optimised according to an objective function [Goldberg (1989), Leardi *et al.* (1992), Leardi (1994, 2001), Todeschini *et al.* (2003)].

Genetic algorithms have been created as an optimisation strategy to be used especially when complex response surfaces do not allow the use of better-known methods (simplex, experimental design techniques, etc.). These algorithms, conveniently modified, can also be a valuable tool in solving the feature selection problem. The subsets of variables selected by genetic algorithms are generally more efficient than those obtained by classical methods of feature selection, since they can produce a better result by using a lower number of features. This is due to the fact that the performance of the model is sensitive to the choice of the features used to construct the model.

Exhaustive evaluation of possible feature subsets is usually unfeasible in practise because of the large amount of computational effort required, like in the molecular descriptor approach, due to the availability of hundreds of features/descriptors that can be calculated from a single protein. Genetic algorithms, which belong to a class of randomised heuristic search techniques, offer an attractive approach to obtain near-optimal solutions to such optimisation problems.

In the following, the studied genetic algorithm strategy for variable subset selection is presented. The feature selection is based on the evolution of a population of models, i.e. a set of ranked models according to some objective function. In genetic algorithm terminology, each population individual is called chromosome and is a binary vector , where each position (a gene) corresponds to a variable (1 if included in the model, 0 otherwise). Each chromosome represents a model given by a subset of variables. Once the objective function to optimise is defined the genetic algorithm evolution starts, based on three main steps:

1. **Random initialisation of the population.** The model population is built initially by random models with a number of variables between 1 and L, where L is the maximum number of variables allowed in a model. The value of the selected objective function of each model is calculated in a process called evaluation. The models are then ordered with respect to the selected objective function - model quality - (the best model is in first place in the population, the worst at position P, where P is the model population size);

2. **Crossover.** From the actual population, pairs of models are selected (randomly or with a probability function of their quality). Then, from each pair of selected models (parents), a new model is generated, preserving the common characteristics of the parents (i.e. variables excluded in both models remain excluded, variables included in both models remain included) and mixing the opposite characteristics according to the crossover probability. If the generated son coincides with one of the individuals already present in the actual population, it is rejected; otherwise, it is evaluated. If the objective function value is better than the worst value in the population, the model is included in the population, in the place corresponding to its rank; otherwise, it is no longer considered. This procedure is repeated for several pairs.

3. **Mutation.** After a number of crossover iterations, the population proceeds through the mutation process. This means that for each individual of the population every gene is randomly changed into its opposite or left unchanged. Mutated individuals are evaluated and included in the population if their quality is acceptable. This process is controlled by mutation probability which is commonly set at low values, thus allowing only a few

mutations and new individuals not too far away from the generating individual.

Unlike the classical genetic algorithm, in the studied approach [Todeschini *et al.* (2003), Mob (2007)] crossover and mutation steps are kept disjoint. Population crossover and mutation are alternatively repeated until a stop condition is encountered (e.g., a user-defined maximum number of iterations) or the process is ended arbitrarily.

An important characteristic of the GA-VSS method is that it provides not a single model but a population of acceptable models; this characteristic, sometimes considered a disadvantage, makes the evaluation of variable relationships with response from different points of view possible. The studied approach extends the genetic strategy based on the evolution of a single population of models to a more complex genetic strategy based on the evolution of more than one population. These populations evolve independently from each other and, after a number of iterations, they can be combined according to different criteria, thus obtaining a new population with different evolutionary capabilities.

Models can be optimised by different statistical parameters to measure their quality. Moreover, the genetic parameters that control the population evolution can be changed during the model searching. Mutation and crossover probabilities are tailored by this strategy. Finally, once the best models from one or more optimised populations are obtained, bootstrap and y-scrambling techniques can be used for further validation.

5.5 Model validation

Since the main aim of chemometric models (both regression and classification models) is the application of the models to unknown samples, great attention has been focused on their predictive capabilities.

In fact, if we consider a simple regression model, it is demonstrated that the fitting performances of the model always increase if new variable are added, even if these new variables are random variables or do not contain useful information. On the other hand, the predictive performances of the model increase only when informative variables are added to the model, otherwise they decrease. This simple case clarifies why the prediction capabilities of a model have to be always tested. All the models carried out in this thesis have been validated using different procedures, depending on the case in analysis.

5.5.1 Leave-one-out (LOO) and leave-more-out (LMO)

The leave-one-out (LOO) procedure is one of the most used validation techniques: it removes each sample from the data set, one at a time, then the model is rebuilt and the response of the removed sample is predicted by using the obtained model. All the samples are sequentially removed and predicted. Finally the mean of the predicted responses obtained on all the samples is calculated. Since LOO can provide over optimistic results [Golbraikh and Tropsha (2002)], also a more robust validation technique (leave-more-out, LMO) has been applied [Burden *et al.* (1997), Baumann and Stiefl (2004)]. In the LMO procedure, a percentage s of samples is randomly removed from the data set; then, the model is rebuilt without these objects and the responses of the removed sample are predicted by the obtained model. This procedure is repeated r times, always with a random selection of s samples. Finally the mean of the predicted responses is calculated. As explained, the LMO procedure is more robust then LOO, but also more time-consuming. Furthermore, LOO gives always the same result, i.e. it is perfectly reproducible, while LMO is based on a initial random selection of the samples to be predicted and can provide different results each time it is applied. The advantages (robustness) and disadvantages (time-consuming, not perfectly reproducible) of LMO can be avoided by means of another validation technique: the samples are divided in different groups (cross-validation groups) and one group at a time is removed from the training set, the model is rebuilt without the left out objects and the responses of the removed sample are predicted. This procedure is repeated for each cross-validation group, and finally the mean of the predicted responses is calculated.

5.5.2 Bootstrap technique

Bootstrapping is a modern, computer-intensive, general purpose approach to statistical inference, falling within a broader class of resampling methods. By bootstrap validation technique [Efron (1979, 1982, 1987)], the original size of the data set (n) is preserved for the training set, by the selection of n objects with repetition; in this way the training set usually consists of repeated objects and the evaluation set of the objects left out. The model is calculated on the training set and responses are predicted on the evaluation set. All the squared differences between the true response and the predicted response of the objects of the evaluation set are collected in *PRESS (predictive residual sum of squares)*. This procedure of building training sets and evaluation sets is repeated thousands of time, *PRESS* are summed and the average predictive power is calculated.

5.5.3 Y-scrambling

Y-scrambling validation technique is adopted to check models with chance correlation, i.e. models where the independent variables are randomly correlated to the response variables. The test is performed by calculating the quality of the model (usually R^2 or, better, Q^2) randomly modifying the sequence of the response vector $\mathbf{y}$, i.e. by assigning to each object a response randomly selected from the true responses [Lindgren *et al.* (1996), Eriksson *et al.* (1997)]. If the original model has no chance correlation, there is a significant difference in the quality of the original model and that associated with a model obtained with random responses. For a model to be valid, the desirable intercept limits should be R^2<*0.3* and Q^2<*0.05*. If both limits are exceeded, the model should be treated with caution. The procedure is repeated several hundreds of times.

part II

Applications

chapter 6

List of applications

Introduction

In the second part of this thesis three different applications of the studied methodology are presented.

The first application described in chapter 7 is a sensitivity analysis performed on an artificial data set. A portion of thirteen amino acids is extracted from the streptavidin protein. On this sequence amino acids having different position have been mutated obtaining different data sets of the mutated sequence corresponding to the position of the mutated amino acid. This artificial data set have been used in order to evaluate the capability of two different descriptor blocks, constitutional and auto-correlation descriptors, to be able to discriminate among different mutated amino acids.

The second application described in chapter 8 has been studied in order to evaluate the capability of the proposed approach to be able to discriminate among proteins belonging to the same fold. Two protein folds are collected from the SCOP database [Murzin *et al.* (1995), LoConte *et al.* (2002), Andreeva *et al.* (2004)] and then all the four descriptor blocks proposed are calculated and principal component analysis were performed. This application was used also to study the capability of different weighting scheme to be able to highlight different kind of information depending also on the calculated molecular descriptors.

The third application described in chapter 9 is the conclusive test performed in order to get the evidence that this methodology is valuable. A peptide data set collected from the literature [Andersson *et al.* (1998)] has been analysed. Two biological responses have been modelled using the constitutional and the autocorrelation descriptors and the models have been obtained by the genetic algorithm variable subset selection technique explained in section 5.4. The obtained models have been validated by bootstrap and y-scrambling analysis described in section 5.5.

Finally in chapter 10 a web-based application developed in order to get publicly available the algorithms described in this PhD thesis is described. In this final chapter some screen shots of this application and the instructions in order to use it are presented.

chapter 7

Sensitivity Analysis of the proposed methodology: the Streptavidin Dataset

7.1 Introduction

The first application described in this part of the thesis has been performed in order to evaluate the capability of the molecular descriptor based approach described in the theory (see part I) to be able to characterise and discriminate among different peptide sequences.

Starting from a protein sequence constituted of 120 amino acids and 1743 atoms, the streptavidin protein (see Figure 7.1), a short sequence of 13 amino acids has been extracted. On this sequence a series of mutations have been performed. Consequently different data sets have been obtained, each data set corresponding to different mutations; on these data sets principal component analysis have been performed in order to evaluate how two different descriptor blocks are able to discriminate among the different mutations. Only the 2-dimensional structure of these peptide sequences has been used, thus only the constitutional and the autocorrelation descriptors have been calculated.

In the next sections the diverse capabilities of these two different descriptor blocks to discriminate among different mutations will be shown.

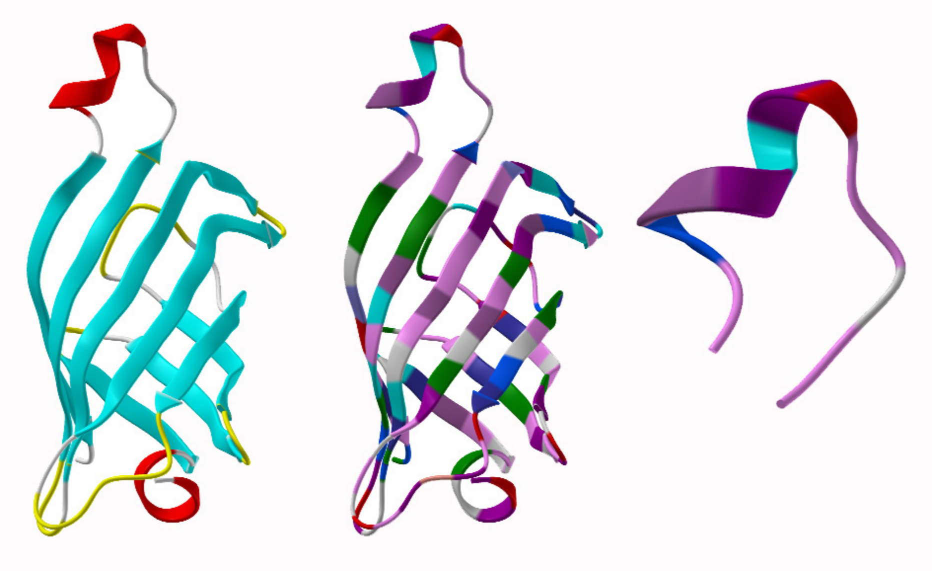


Figure 7.1: Schematic representation of the streptavidin protein. On the left schematic representation of the tertiary structure of the whole protein. In the middle schematic representation of the whole protein coloured by residue type. On the right schematic representation of the amino acid sequence extracted by the streptavidin protein (amino acids from 111 to 123) coloured by residue type.

7.2 Streptavidin dataset

Streptavidin is a 60,000 dalton tetrameric protein purified from the bacterium *Streptomyces avidinii*. It finds wide use in molecular biology through its extraordinarily strong affinity for the vitamin biotin; the dissociation constant (Kd) of the biotin-streptavidin complex is on the order of $\sim$ 10-15 mol/L, ranking among one of the strongest known non-covalent interactions [Wilchek and Bayer (1989), Miyamoto and Kollman (1993), Zimmermann and Cox (1994)].

As introduced in the first section of this chapter a sequence of 13 amino acids has been extracted from the streptavidin protein. The sequence extracted comes from the amino acid in position 111 to the amino acid in position 123. The streptavidin protein and the extracted sequence is shown in Figure 7.1. Once extracted this sequence has been virtually mutated, this means that the amino

Table 7.1: The peptide sequence of 13 amino acids extracted from the streptavidin protein.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
T	S	G	T	T	E	A	N	A	W	K	S	T

acid in position 2 of the extracted sequence (a serine - S) has been changed in order to obtain 19 more sequences. Each of them containing in position 2 one of the others natural amino acids as represented in Table 7.2.

Table 7.2: The 20 peptide sequences of 13 amino acids extracted from the streptavidin protein with a single mutation in position 2. Amino acid serine (S) in the original sequence (first row in the table) has been substituted with all the other 19 natural amino acids in an iterative way.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
T	S	G	T	T	E	A	N	A	W	K	S	T
T	A	G	T	T	E	A	N	A	W	K	S	T
T	C	G	T	T	E	A	N	A	W	K	S	T
T	D	G	T	T	E	A	N	A	W	K	S	T
T	E	G	T	T	E	A	N	A	W	K	S	T
T	F	G	T	T	E	A	N	A	W	K	S	T
T	G	G	T	T	E	A	N	A	W	K	S	T
T	H	G	T	T	E	A	N	A	W	K	S	T
T	I	G	T	T	E	A	N	A	W	K	S	T
T	K	G	T	T	E	A	N	A	W	K	S	T
T	L	G	T	T	E	A	N	A	W	K	S	T
T	M	G	T	T	E	A	N	A	W	K	S	T
T	N	G	T	T	E	A	N	A	W	K	S	T
T	P	G	T	T	E	A	N	A	W	K	S	T
T	Q	G	T	T	E	A	N	A	W	K	S	T
T	R	G	T	T	E	A	N	A	W	K	S	T
T	T	G	T	T	E	A	N	A	W	K	S	T
T	V	G	T	T	E	A	N	A	W	K	S	T
T	W	G	T	T	E	A	N	A	W	K	S	T
T	Y	G	T	T	E	A	N	A	W	K	S	T

This data set has been used as starting point in order to build four new data sets each of them collecting 400 peptide sequences. The four new data sets has been obtained mutating in position 5, 7, 9 and 11 the starting data set of twenty sequences. Each new data set has been built considering each sequence of the starting data set and mutating a single amino acid in a single position.

In this way from every starting sequence 20 new mutated sequences have been obtained. Considering that the starting data set is constituted by 20 sequences every mutated data set is constituted by 400 different amino acid sequences.

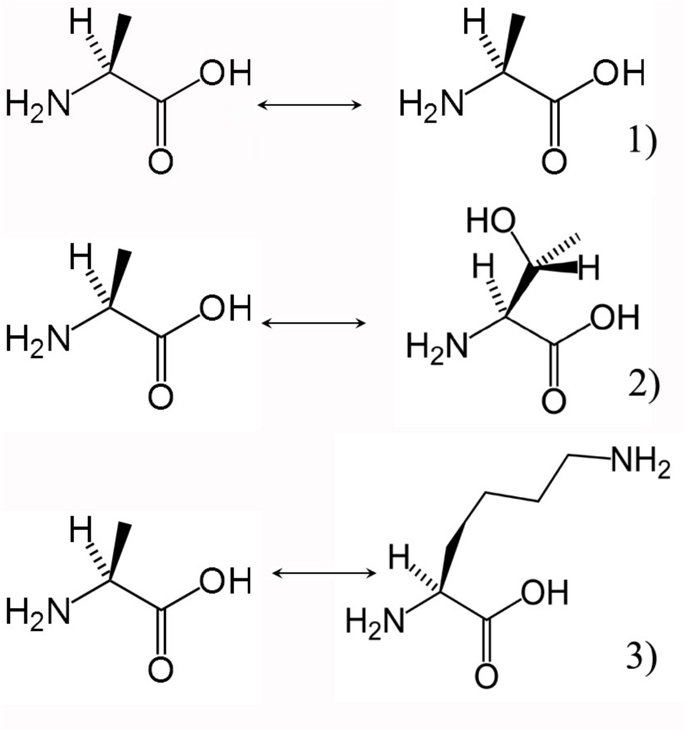


Figure 7.2: Illustration of the different compared data sets using the mutated amino acids. On the left the mutation in position 7, where the original amino acid is an alanine, on the right the compared mutations. In position 9 the amino acid in the original sequence is an alanine (label 1), in position 5 the amino acid in the original sequence is a threonine (label 2) and finally in position 11 the amino acid in the original sequence is a lysine (label 3).

7.3 Results and Discussions

Once the four data sets have been constituted the constitutional and the autocorrelation molecular descriptor blocks have been calculated on all the four data sets of 400 peptide sequences.

The molecular descriptors have been calculated using the physicochemical weighting scheme described in section 3.5.2. This weighting scheme is constituted of five different physicochemical properties collected from the amino acid index database. The selected indices are:

1. molecular weight [Fasman (1976)] (FASG760101);
2. polarity [Grantham (1974)] (GRAR740102);
3. hydrophobicity [Jones (1975)] (JOND750101);
4. residue accessible surface area in folded protein [Chothia (1976)] (CHOC76010);
5. hydrophilicity scale [Kuhn *et al.* (1995)] (KUHL950101).

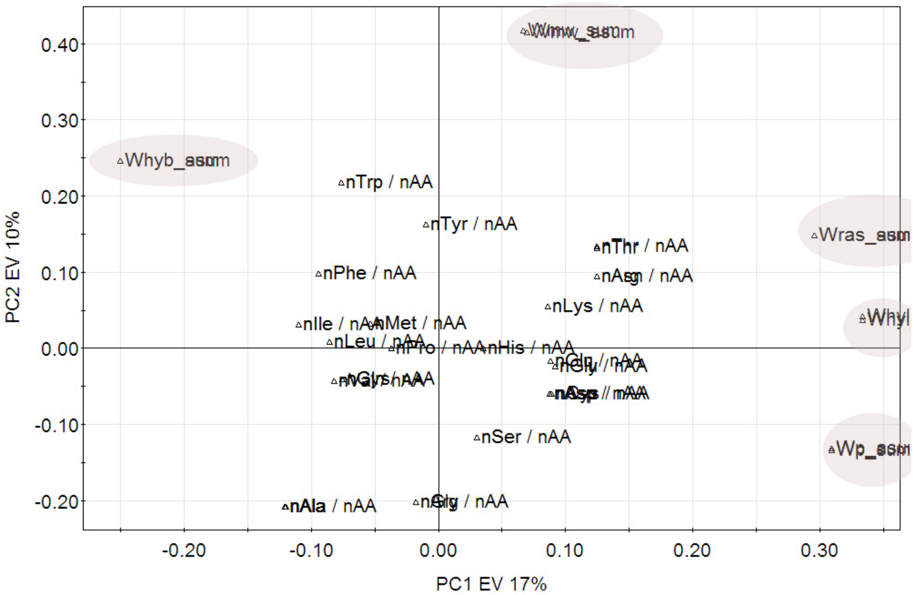


Figure 7.3: Loading plot of the first two principal components calculated on the constitutional descriptors for the data sets with mutation in position 7 and 9. The global descriptors *Wk_sum* and *Wk_asum*, where *k* is the considered weight, are highlighted.

Once computed the descriptor values a sensitivity analysis has been performed in order to evaluate how the calculated molecular descriptors are able to discriminate among the four different mutated data sets. The data set collecting all the

400 sequences with mutation in position 7 has been compared with the other three data sets, describing the mutation in position 5, 9 and 11.

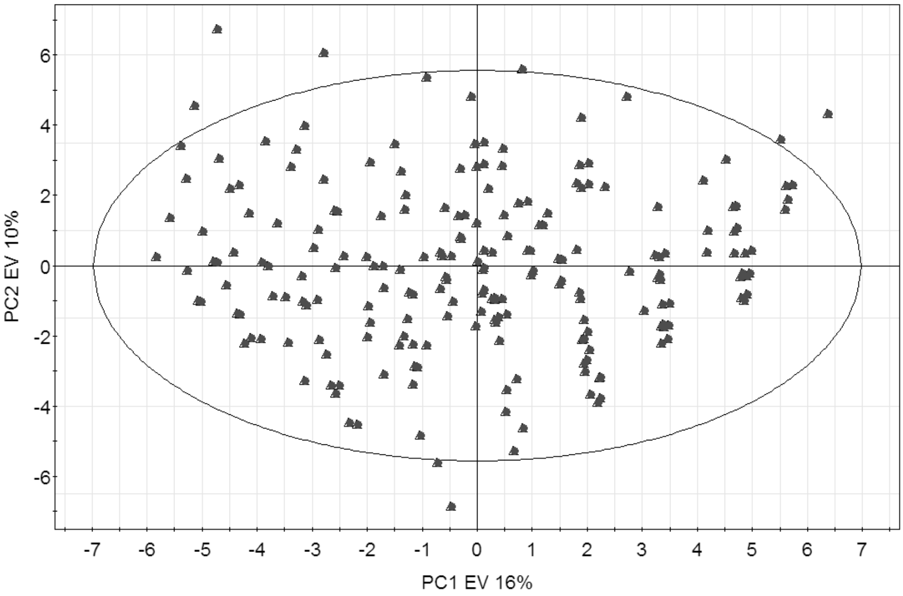


Figure 7.4: Score plot of the first two principal components for the data sets with mutation in position 7 (empty triangle) and 9 (filled circles). PCA performed on constitutional descriptors.

In order to evaluate if and how different molecular descriptor blocks are able to discriminate among different mutations the four data sets have been compared. Iteratively the data set describing the mutation in position 7 has been compared to the other three data sets. The descriptor values calculated for the mutation in position 7 have been added to the descriptor values of the other three data sets obtaining four data sets, each of one describing two different mutations, comprising 800 peptide sequences each. On these data sets principal component analysis have been performed, results are showed in Figure 7.4, Figure 7.5, Figure 7.6, Figure 7.7, Figure 7.8 and Figure 7.9.

In Figure 7.3 is displayed the loading plot of the constitutional descriptors for the first two principal components for the data set with peptide sequences mutated in position 7 and 5. The explained variance (EV) of the first component is

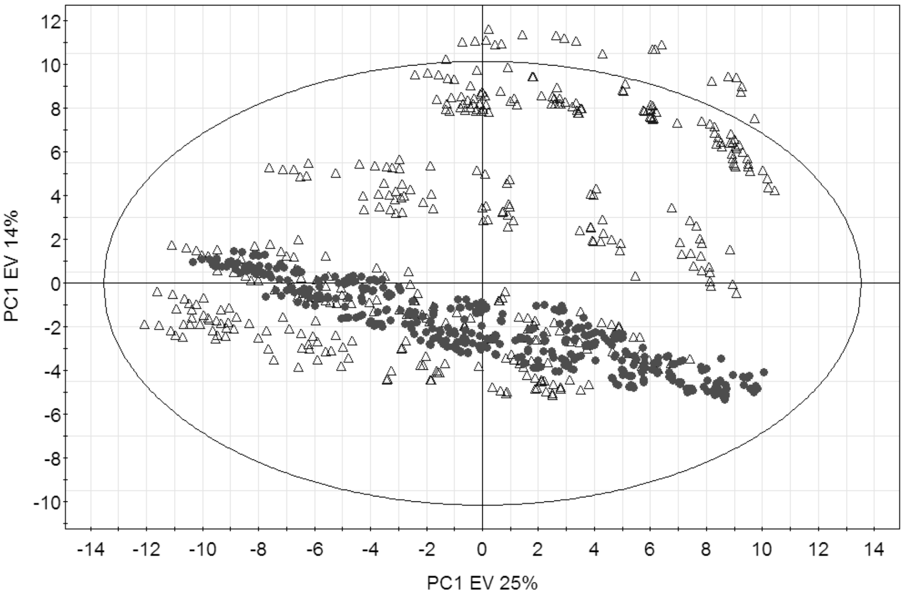


Figure 7.5: Score plot of the first two principal components for the data sets with mutation in position 7 (empty triangle) and 9 (filled circles). PCA performed on autocorrelation descriptors.

17% while the explained variance of the second principal component is 10%. The molecular descriptors with higher loadings in the first two principal components are the descriptors correlated to the global dimension of the peptide sequences. These descriptors are the sum and the average sum of the different physicochemical properties used to characterise the amino acids, respectively the symbols of these descriptors are *Wk_sum* and *Wk_asum*, where *k* is the considered weight (for a complete list of descriptor symbols and descriptions see chapter III).

Accordingly to the loading plot showed in Figure 7.3 peptide sequences with high values for *Wmw_sum*, e.g. peptide sequences with high molecular weight, will have high values for the second principal component; peptide sequences with high values for *Wras_sum*, *Whyl_sum* and *Wp_sum* will have high values for the first principal component and peptide sequences with low values for *Whyl_sum* will have low values on the first principal component.

Principal component analysis of the mutation in position 7 and 9 is showed in Figure 7.4 and Figure 7.5. Peptide sequences mutated in position 7 are represented in the figure with empty triangle while peptide sequences mutated in position 5 are represented by filled circle. Principal component analysis on the constitutional descriptors (Figure 7.4) show that this descriptor block is not able to discriminate among mutations in position 9 and mutations in position 7. This result is due to the fact that in the original sequences both in position 9 and in position 7 there is an alanine. Constitutional descriptors, avoiding the information related to the topology and the connectivity, are not able to discriminate among mutation if the final composition of the peptide sequences is the same. Instead principal component analysis on the autocorrelation descriptors (Figure 7.5) show that introducing the information about the connectivity and the relationships among amino acids belonging to the same peptide sequences it is possible to discriminate among different mutations.

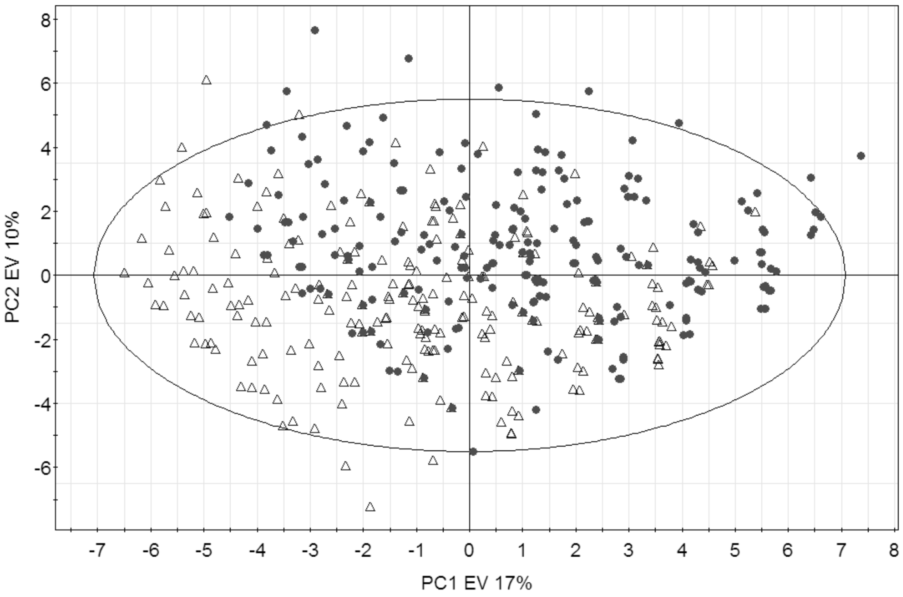


Figure 7.6: Score plot of the first two principal components for the data sets with mutation in position 7 (empty triangle) and 5 (filled circles). PCA performed on constitutional descriptors.

In Figure 7.6 and Figure 7.7 principal component analysis performed on the

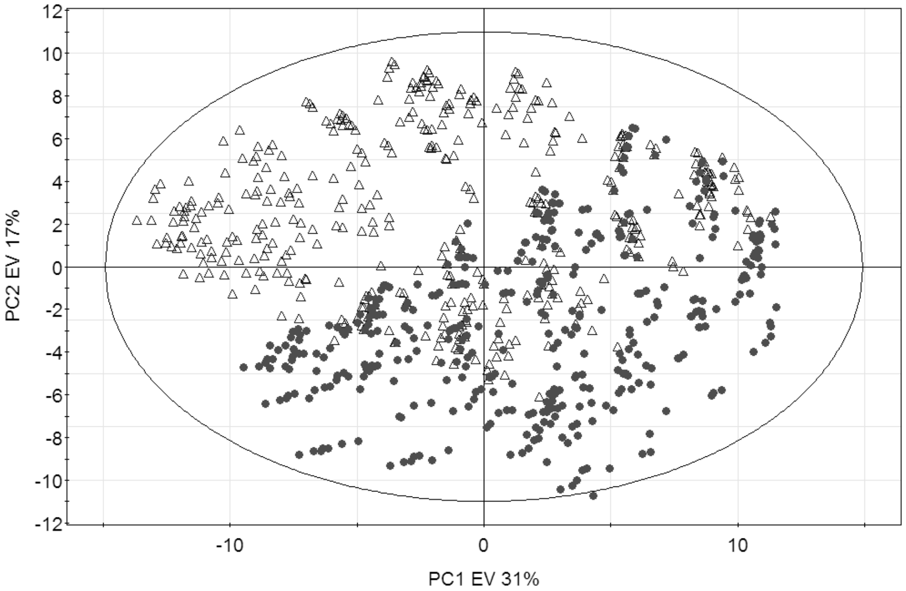


Figure 7.7: Score plot of the first two principal components for the data sets with mutation in position 7 (empty triangle) and 5 (filled circles). PCA performed on autocorrelation descriptors.

data set collecting the mutations occurred in position 7 and in position 5 is showed. The principal component analysis performed on the constitutional descriptors (Figure 7.6) show that this descriptor block is quite able to discriminate between mutation in position 7 and mutation in position 5. This change of behaviour for the constitutional descriptors is due to the fact that in the original sequences in position 7 there is an alanine while in position 5 there is a threonine. These two amino acids have different physicochemical properties and constitutional descriptors are able to highlight these differences. Analogously, but in a clearer way, also the autocorrelation descriptors are able to discriminate between this two mutations (Figure 7.7).

Finally in Figure 7.8 and Figure 7.9 principal component analysis performed on the data set collecting the mutations occurred in position 7 and in position 11 is presented. In the original sequences in position 7 there is an alanine while in position 11 there is a lysine, these two amino acids have very different physic-

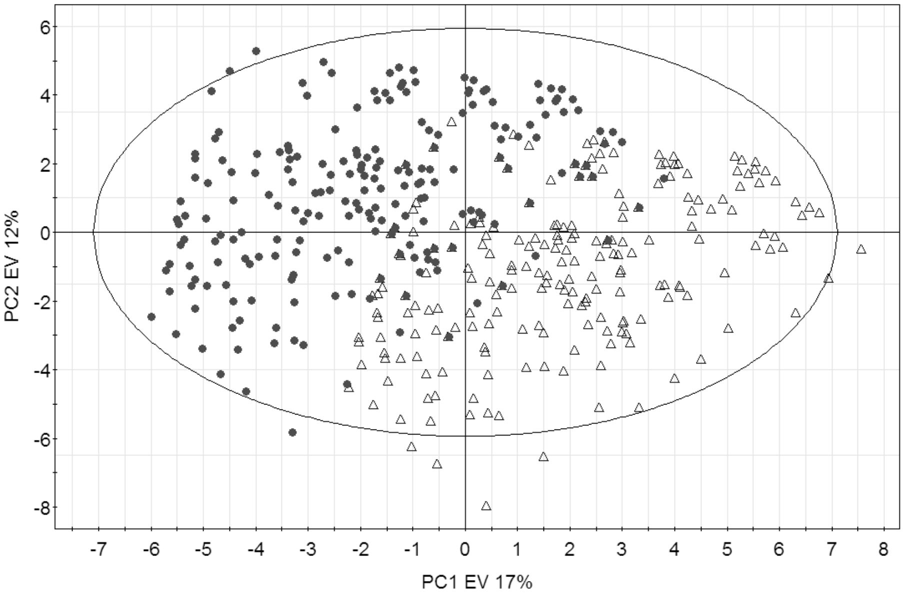


Figure 7.8: Score plot of the first two principal components for the data sets with mutation in position 7 (empty triangle) and 11 (filled circles). PCA performed on constitutional descriptors.

ochemical properties. These differences are highlighted both by the principal component analysis performed on the constitutional (Figure 7.8) and on the autocorrelation descriptors (Figure 7.9).

7.4 Conclusions

The preliminary analysis described in this chapter show that the molecular descriptor based approach presented in this PhD thesis is able to discriminate among different peptide sequences. Constitutional and autocorrelation descriptor blocks are able to discriminate among homogeneous data sets where a single mutation position is changed.

The capability of the two considered descriptor blocks to discriminate among different mutations are diverse. Autocorrelation descriptors are able to discriminate among different mutation position also when the final composition of the

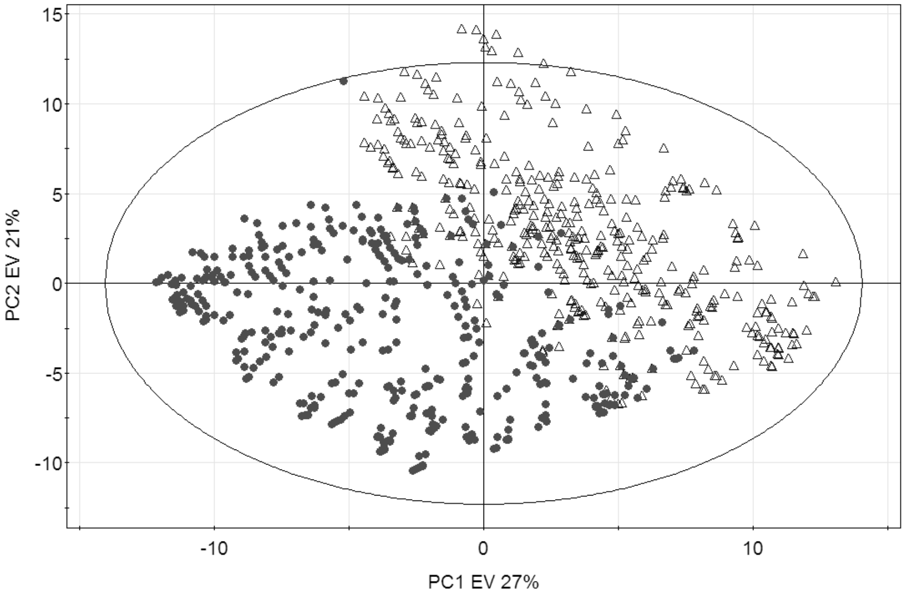


Figure 7.9: Score plot of the first two principal components for the data sets with mutation in position 7 (empty triangle) and 11 (filled circles). PCA performed on autocorrelation descriptors.

two data sets is the same, like for mutation in 7 and 9. Anyway constitutional descriptors show results similar to autocorrelation descriptors when the differences among the final composition of the mutated amino acid sequences increases.

In the next chapter also the 3-dimensional descriptors have been used and the capability of different weighting scheme to highlight different sources of information is described.

chapter 8

Cluster analysis of two different protein folds

8.1 Introduction

In this chapter has been analysed the capability of the proposed molecular descriptor based approach to discriminate among superfamilies and families in two selected folds according to the Structural Classification of Proteins (SCOP) representation [Murzin *et al.* (1995), LoConte *et al.* (2002), Andreeva *et al.* (2004)].

This investigation has been performed using the Principal Component Analysis - PCA (see section 5.2). Four descriptor blocks have been calculated for both protein folds using two different weighting schemes, physicochemical and statistical weighting schemes. Principal component analysis has been applied in order to evaluate the capabilities of the two different weighting schemes to highlight different kind of information.

8.2 The SCOP database

Structural Classification of Proteins (SCOP) [Murzin *et al.* (1995), LoConte *et al.* (2002), Andreeva *et al.* (2004)] is the most cited resource for classifying proteins, it provides a detailed and comprehensive description of the structural and evo-

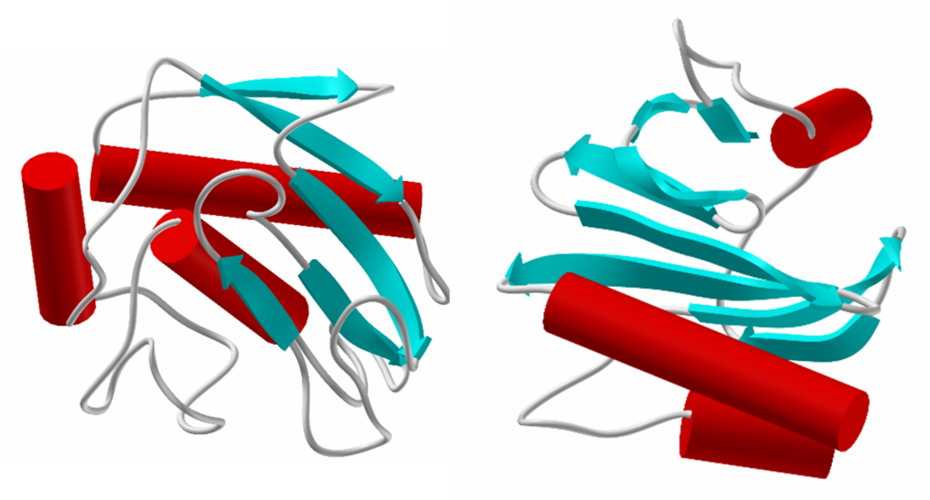


Figure 8.1: Schematic representation of two different protein folds. On the left d1mmq__ protein belonging to zincin-like fold (D.92), on the right d3nul__ protein belonging to profilin-like fold (D.110).

lutionary relationships of proteins whose three-dimensional structures have been determined.

The SCOP database is organized on a number of hierarchical levels that embody the evolutionary and structural relationships, these levels are: family, superfamily, fold and class. *Families* contain protein domains that share a clear common evolutionary origin, as evidenced by sequence identity, of 30% and greater, or extremely similar structure and function. *Superfamilies* consist of families whose proteins have low sequence identities but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable. *Folds* consist of one or more superfamilies that have same major secondary structures in same arrangement with the same topological connections.

The different folds are grouped into five structural classes on the basis of the secondary structures of which they are composed. All alpha (for proteins whose structure is essentially formed by α-helices). All beta (for those whose structure is essentially formed by β-sheets). Alpha and beta (for proteins with α-helices and β-strands that are largely interspersed). Alpha plus beta (for those in which α-helices and β-strands are largely segregated) and multi-domain (for those with domains of different fold and for which no homologues are known at present).

8.3 Choice of two different protein folds

Proteins to be investigated have been searched in the SCOP database [Murzin *et al.* (1995), LoConte *et al.* (2002), Andreeva *et al.* (2004)] from which have been excluded the domains with sequential similarity higher than 95% in order to avoid the use of not meaningful folds. In this reduced database some constraints to reduce the number of useful folds have been introduced.

Selected folds have been chosen from folds belonging to domains with a mixture of helix and sheet structures, α/β (alpha and beta), and $\alpha + \beta$ (alpha plus beta). The α/β domains principally consist of a single β-sheet, with α-helices joining the C-terminus of one strand to the N-terminus of the next. Domains that have the α and β units largely separated in sequence fall into the $\alpha + \beta$ class.

Other constraints applied to the fold selection are:

1. Superfamily: between 2 and 7;

2. Family: between 6 and 18;

3. Domain: at least 30 for the selected fold;

4. Family/Superfamily: higher than 2.

After this preliminary selection carried out to obtain folds with a representative number of families and superfamilies, two different folds have been selected (see Figure 8.1):

1. Zincin-like fold (D.92);

2. Profilin-like fold (D.110).

The first one (D.92) has been selected due to its representative composition in superfamilies and families. It is constituted by 2 superfamilies (metalloproteases and beta-N-acetylhexosaminidase), the first one clustered in 15 and the second one in 2 families. The Zincin-like fold comprises 56 domains, it belongs to $\alpha + \beta$ class and contains mixed β-sheet with connection over free side of the sheet.

Zincin-like fold proteins are mainly represented by the zinc metalloprotease enzyme family, a very well studied enzyme class involved in very diverse processes ranging from embryonic development to cancer and classified into distinct families exhibiting shared zinc binding motifs [Hooper (1994)]. One such family are the matrix metalloproteases (MMPs). Misregulation of MMPs is believed to contribute to pathological conditions such as cancer [Kleiner and Stetler-Stevenson (1999)], angiogenesis [Lohmander *et al.* (1993)], osteoarthritis, rheumatoid arthritis [Murphy and Hembry (1992)], remodelling in Alzheimer disease [Peress *et al.* (1995)] and pulmonary emphysema [Skiles *et al.* (2001, 2004)].

The Profilin-like fold has been selected due to its environmental significance; in effect, the PAS (Per-ARNT-Sim) superfamily of proteins belongs to this fold, which is a widely studied collection of single- and multidomain proteins involved in inducing and regulating some of the basic adaptive mechanisms of the cell [Taylor and Zhulin (1999), Gu *et al.* (2000), Repik *et al.* (2000), Kewley *et al.* (2004), Pandini and Bonati (2005)]. This fold is constituted by 7 superfamilies, 16 families and 36 domains. Also profilin-like fold belongs to $\alpha + \beta$ class that comprises proteins with mainly antiparallel beta sheets (segregated alpha and beta regions). Its structure core is constituted by two α-helices and five stranded anti parallel sheets.

8.4 Results and Discussions

Two analyses have been separately performed, one on the D.92 data set and the other on the D.110 data set. The former consists of 56 domains, the latter consists of 36; both data sets have been described by four molecular descriptor blocks. The four calculated descriptor blocks have been weighted using two different weighting schemes: the physicochemical and the statistical weighting scheme. The properties used to characterise the amino acids for the physicochemical weighting schemes are:

1. molecular weight [Fasman (1976)] (FASG760101);
2. polarity [Grantham (1974)] (GRAR740102);
3. hydrophobicity [Jones (1975)] (JOND750101);
4. residue accessible surface area in folded protein [Chothia (1976)] (CHOC76010);
5. hydrophilicity scale [Kuhn *et al.* (1995)] (KUHL950101).

While the properties considered in the statistical weighting scheme are:

1. relative frequency in beta-sheet [Prabhakaran (1990)] (PRAM900103);

2. relative frequency of occurrence [Jones *et al.* (1992)] (JOND920101);

3. relative mutability [Jones *et al.* (1992)] (JOND920102)

4. relative frequency in alpha-helix [Prabhakaran (1990)] (PRAM900102)

5. relative frequency in reverse-turn [Prabhakaran (1990)] (PRAM900104).

These two descriptor schemes are described in section 3.5.

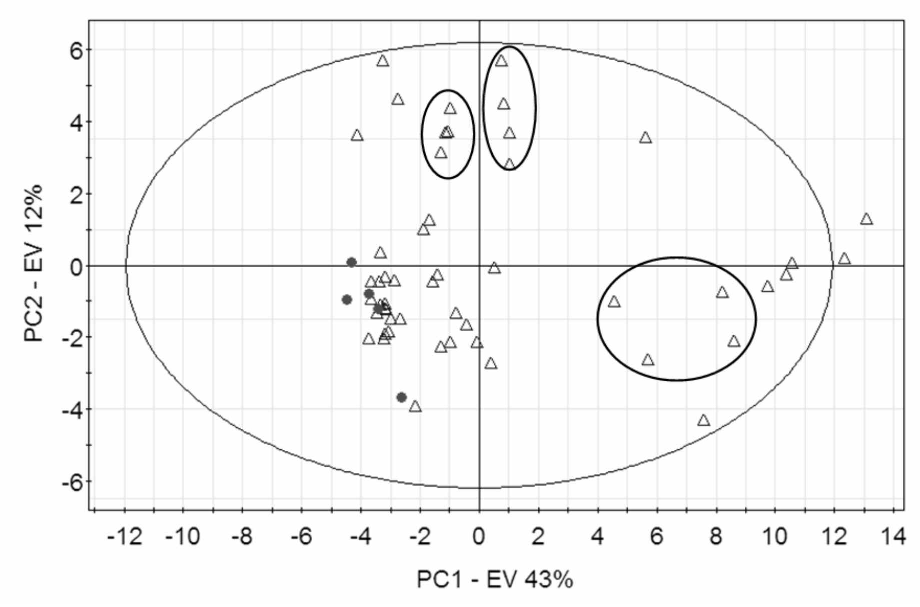


Figure 8.2: Score plot of the first two principal components for the data set D.92. PCA performed on constitutional descriptors weighted by physcico-chemical weights. Metalloproteases superfamily is represented by filled circle, beta-N-acetylhexosaminidase is represented by empty triangles. Families clustered correctly are highlighted by ovals.

Both protein folds have been analysed using the four descriptor blocks independently, every fold has been analysed considering constitutional descriptors and their principal components, 2D autocorrelation descriptors and their principal components, then WHIM descriptors and their PCs and finally GETAWAY descriptors and their PCs.

Both considered weighting schemes collects five different properties, so the number of calculated molecular descriptors is equal for the two weighting schemes.

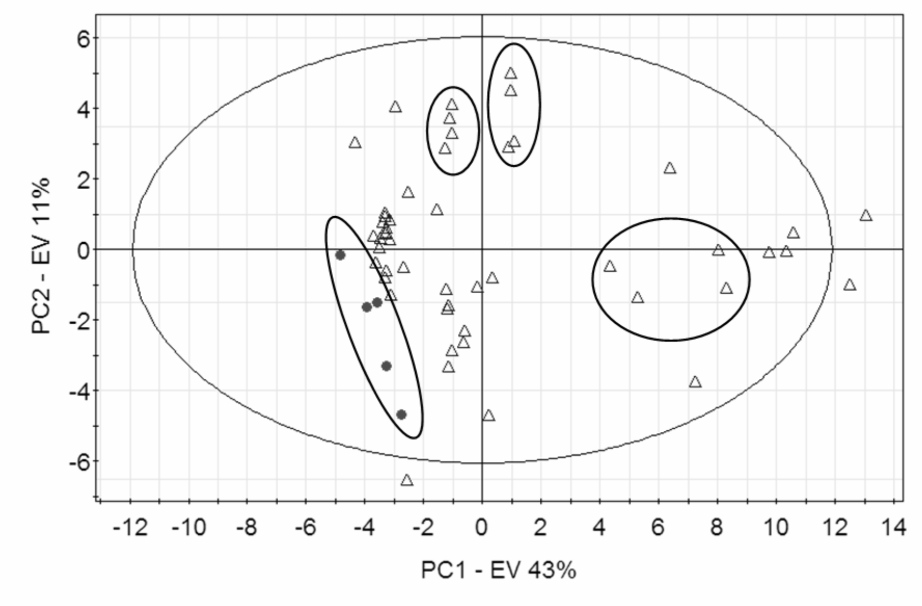


Figure 8.3: Score plot of the first two principal components for the data set D.92. PCA performed on constitutional descriptors weighted by statistical weights. Metalloproteases superfamily is represented by filled circle, beta-N-acetylhexosaminidase is represented by empty triangles. Families clustered correctly are highlighted by ovals.

Constitutional descriptor block consists of 51 descriptors, 2D autocorrelation descriptors consist of 120 molecular descriptors, WHIM block of 99 descriptors and the GETAWAY block of 235 descriptors. After the calculation all the constant descriptors have been excluded; all the descriptors with a pair correlation, with another descriptor, higher than 0.99 have been also excluded.

The obtained score plots allow an introductory view of the ability of the selected molecular descriptors to adequately represent the clusters in which the different domains are grouped. For both data sets the constitutional and the GETAWAY descriptors are deeply analysed in order to evaluate the capability of the two different weighting schemes to highlight different kind of information.

8.4.1 Zincin-like fold (D.92)

Zincin-like fold (D.92) consists of 2 superfamilies, 17 families and 56 domains. The D.92 superfamilies are the metalloproteases superfamily that collects 2 fam-

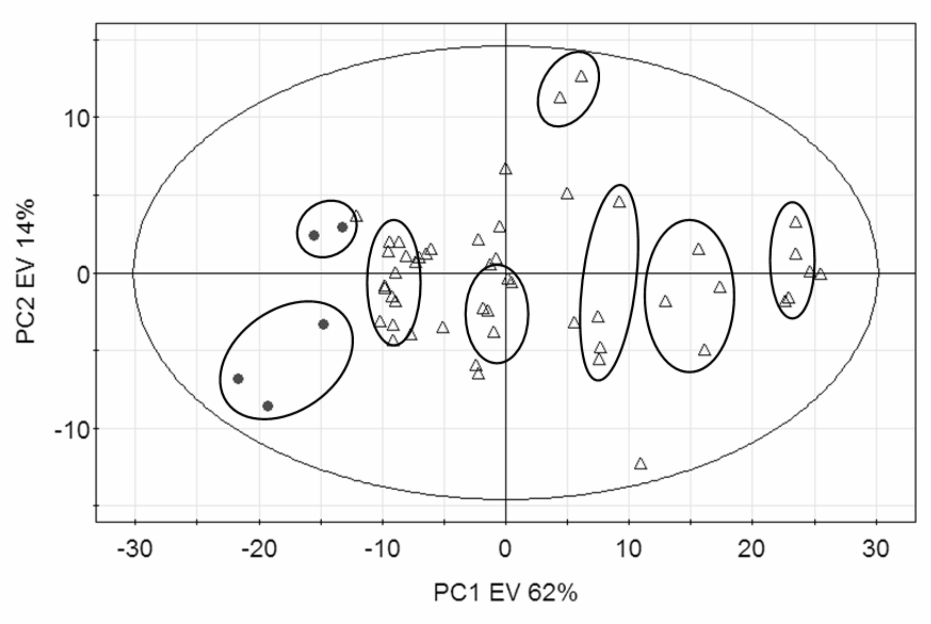


Figure 8.4: Score plot of the first two principal components for the data set D.92. PCA performed on GETAWAY descriptors weighted by physcicochemical weights. Metalloproteases superfamily is represented by filled circle, beta-N-acetylhexosaminidase is represented by empty triangles. Families clustered correctly are highlighted by ovals.

ilies, one with 2 and one with 3 domains; and the beta-N-acetylhexosaminidase superfamily that collects 15 families, 7 of them described by only one domain. The metalloproteases superfamily collects domains with a number of amino acids between 131 and 147 while the beta-N-acetylhexosaminidase superfamily comprise domains with a number of amino acids between 132 and 696.

In Figure 8.2 is represented the scatter plot of the first two principal components calculated on the constitutional descriptors weighted by the physicochemical properties. The first two PCs explain more than 55% of the total variance and more than 90% of variance is explained by the first 10 PCs. Analogously the scatter plot of the first two principal components calculated on the constitutional descriptors weighted by the statistical properties is shown in Figure 8.3. The explained variance of the first two PCs is 44% and the first 10 PCs explain more than 90% of the total variance of the data set.

All the descriptors of these two blocks have the same values, except for the

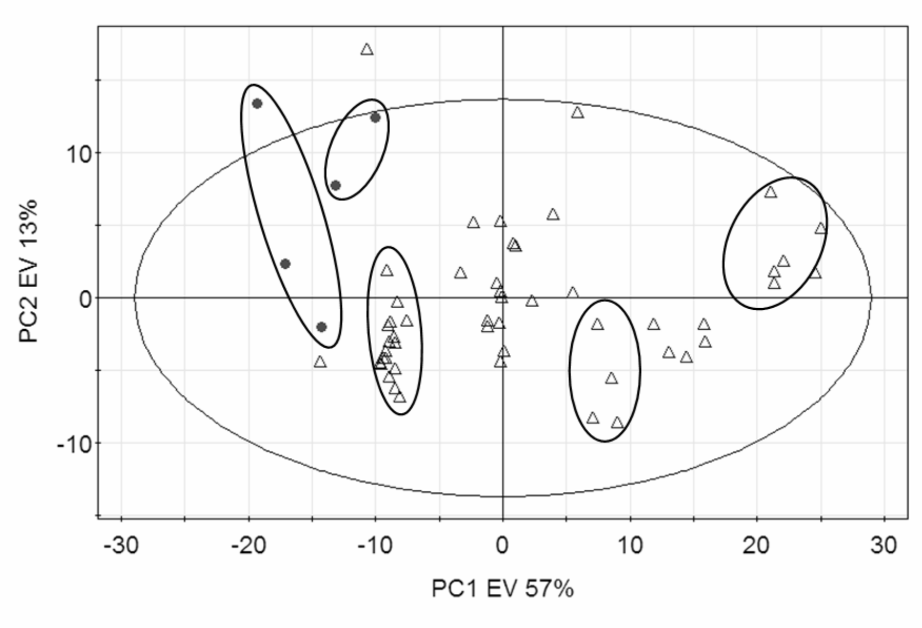


Figure 8.5: Score plot of the first two principal components for the data set D.92. PCA performed on GETAWAY descriptors weighted by statistical weights. Metalloproteases superfamily is represented by filled circle, beta-N-acetylhexosaminidase is represented by empty triangles. Families clustered correctly are highlighted by ovals.

Wk_sum and *Wk_asum* that are the only descriptors depending on the weighting scheme among the constitutional descriptor block.

Considering the constitutional descriptors weighted by the physicochemical properties, the significant variables are *Wmw_sum*, *Wras_sum* and *Whyb_sum* that have high relevance on the first PC, while on the second PC the relevant descriptors are the average sum of molecular weight and hydrophobicity (*Wmw_asum* and *Whyb_asum*). Conversely all the weighted descriptors are relevant on the first two principal components calculated on the constitutional descriptors weighted by the statistical properties. The sum of the statistical properties *Wk_sum*, where *k* is the considered weight, have high relevance on the first PC while the average sum of the weights (*Wk_asum*) are significant on the second PC.

The scatter plot in Figure 8.3 shows a clear separation between the two superfamilies, while the physicochemical weighting scheme (Figure 8.2) is not able to

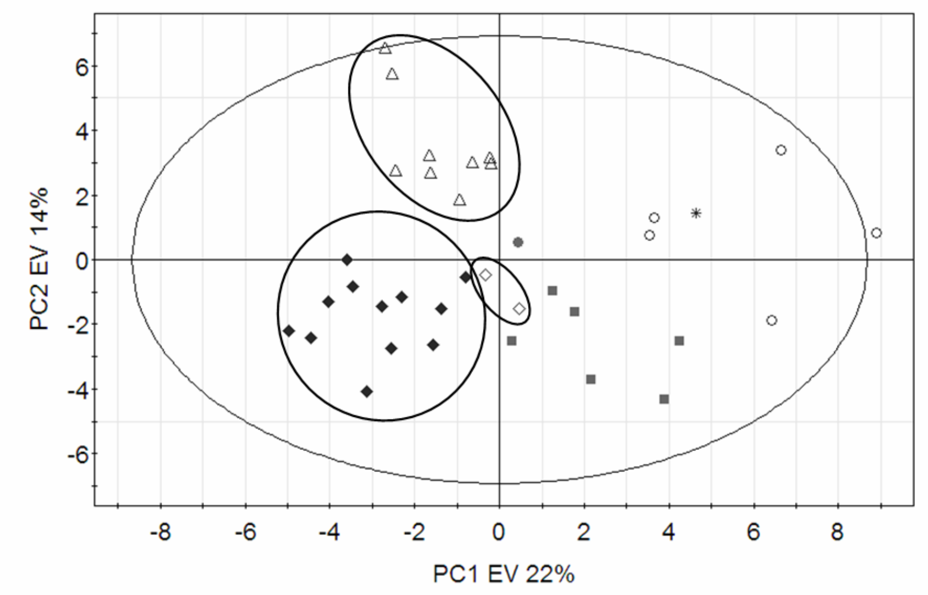


Figure 8.6: Score plot of the first two principal components for the data set D.110. PCA performed on constitutional descriptors weighted by physicochemical weights. Superfamilies clustered correctly are highlighted by ovals.

discriminate between metalloproteases superfamily and beta-N-acetylhexosaminidase superfamily. The differences between the two scatter plots is due to the different information collected by the statistical weights.

In the two figures the families correctly separated by the others are highlighted by ovals, three families are properly clustered using the physicochemical properties and four using the statistical weights.

In Figure 8.4 and Figure 8.5 are showed the first two principal components calculated on the GETAWAY descriptors using the physicochemical weighting scheme and the statistical weighting scheme respectively. Looking deeply at the scatter plots obtained from the GETAWAY descriptors a first evidence is that the physicochemical weights seem more capable to discriminate among superfamilies and families than the statistical weights. In Figure 8.4 eight different families are recognised while only five families are properly segregated in Figure 8.5.

The observed behaviour of the two different weighting scheme can be related to the information collected by the two different types of selected properties. Statistical properties work better with constitutional descriptors while physicochemical properties work better with 3-dimensional descriptors, like the GET-

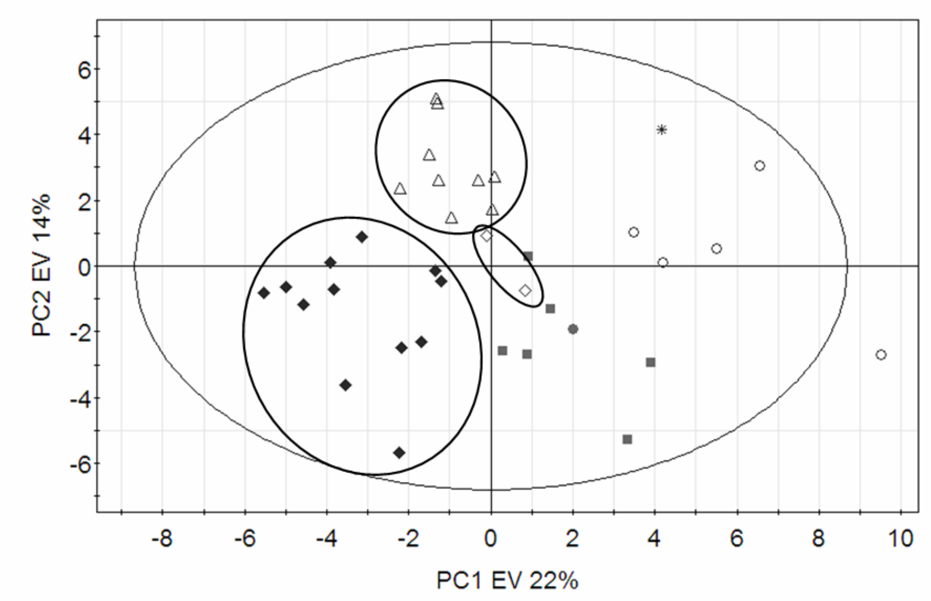


Figure 8.7: Score plot of the first two principal components for the data set D.110. PCA performed on constitutional descriptors weighted by statistical weights. Superfamilies clustered correctly are highlighted by ovals.

AWAY descriptors. Statistical properties are more related to the composition of a protein while physicochemical properties are more related to the 3-dimensional structure. Anyway 3-dimensional descriptors are more capable to discriminate among families than constitutional descriptors.

8.4.2 Profilin-like fold (D.110)

Profilin-like fold (D.110) is more complex than D.92 fold, it is composed by 7 superfamilies, 16 families and only 36 domains. It means that some superfamilies and families have modest descriptive information and it could be more difficult to discriminate them. Profilin-like fold consists of domains with a number of amino acids between 100 and 186.

The Sensor kinase superfamily collects 2 domains; the Roadblock/LC7 and the LuxR superfamilies collects only 1 domain each. These three superfamilies are the worst characterized but in any case they are not confused too much with the other superfamilies.

In Figure 8.6 is represented the scatter plot of the first two principal compo-

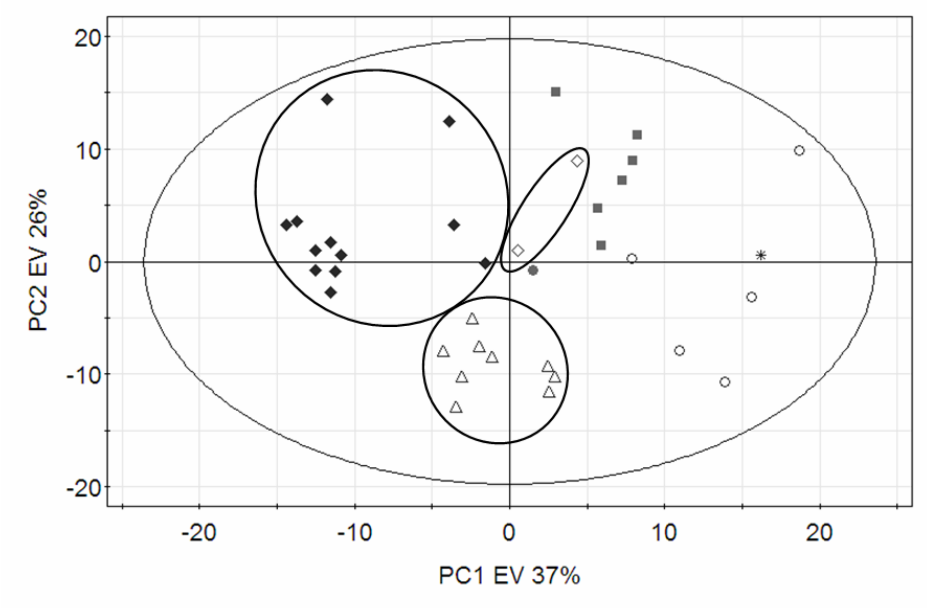


Figure 8.8: Score plot of the first two principal components for the data set D.110. PCA performed on GETAWAY descriptors weighted by physicochemical weights. Superfamilies clustered correctly are highlighted by ovals.

nents calculated on the constitutional descriptors weighted by the physicochemical properties. Analogously the scatter plot of the first two principal components calculated on the constitutional descriptors weighted by the statistical properties is shown in Figure 8.7. For both data sets the first two principal components collects only the 36% of the total variability in the original data.

Like for D.92 fold all the descriptors of these two blocks have the same values, except for the *Wk_sum* and *Wk_asum* that are the only descriptors depending on the weighting schemes among the constitutional descriptor block. Considering the constitutional descriptors weighted by the physicochemical properties the significant variables are *Wp_sum*, *Whyl_sum* and *Whyb_sum* that have high relevance on the first PC, while on the second PC the relevant descriptors are the average sum of all the physicochemical properties except the polarity (*Wmw_asum*, *Whyl_asum*, *Wras_asum* and *Whyb_asum*).

The scatter plot in Figure 8.7 shows a comparable separation than in Figure 8.6 among the seven superfamilies. The weighted descriptors calculated in the constitutional descriptor block show the same behaviour observed for the

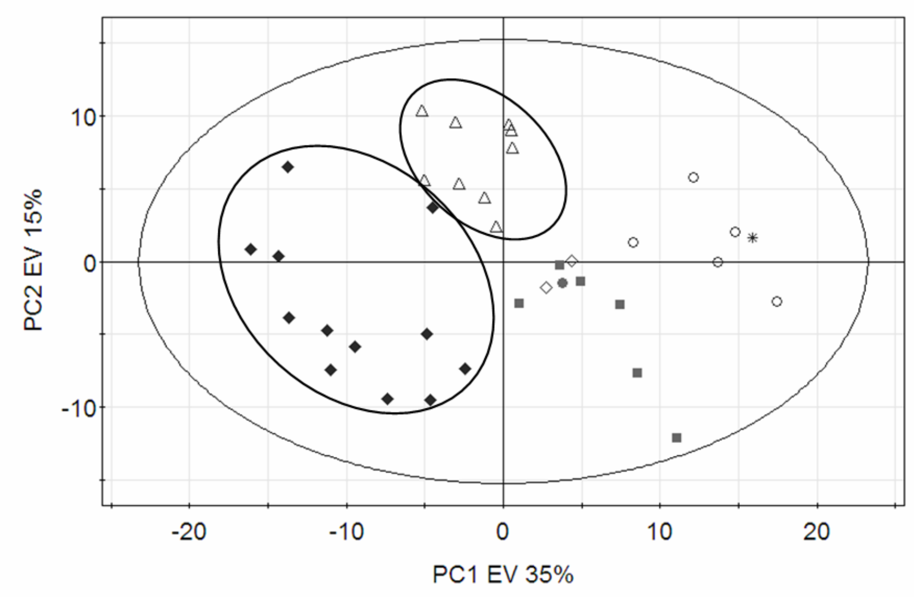


Figure 8.9: Score plot of the first two principal components for the data set D.110. PCA performed on GETAWAY descriptors weighted by statistical weights. Superfamilies clustered correctly are highlighted by ovals.

D.92 fold. The sum of the statistical properties *Wk_sum*, where *k* is the considered weight, have high relevance on the first PC while the average sum of the weights (*Wk_asum*) are significant on the second PC.

In Figure 8.8 and Figure 8.9 are showed the first two principal components calculated on the GETAWAY descriptors using the physicochemical weighting scheme and the statistical weighting scheme respectively. For both weighting schemes more than 90% of variance is explained by the first 11 PCs. The first two principal components in Figure 8.8 explain 66% of the total variability in the original data, while 50% of the total variability in the original data is explained by the first two PCs in Figure 8.9.

The greater complexity of D.110 data set compared to D.92 is due to the high number of superfamilies and families together with a low number of domains. Conversely to the results obtained with the D.92 data set, for the D.110 data set GETAWAY descriptors are not able to discriminate better than constitutional descriptors among different superfamilies and families. For the D.110 data set both descriptor blocks bring an analogous kind of information.

8.5 Conclusions

The application presented in this chapter, diversely from the sensitivity analysis performed in chapter 7, has been applied on proteins whose structure is known. The availability of the 3-dimensional information enables the calculation of both 2-dimensional descriptors and geometrical descriptors. The analysis has been conducted in order to evaluate the capabilities of different weighting schemes to highlight different sources of information.

The attention has been focused on the diverse results obtained calculating descriptors characterising the amino acids with two different weighting schemes. A physicochemical and a statistical weighting scheme. Analysing the results obtained applying the principal component analysis emerges that the statistical weighting scheme appear more informative if it is used together with constitutional descriptors while the physicochemical properties seems more useful if linked together 3-dimensional descriptors.

This behaviour can be conducted to the fact that statistical properties are more related to the composition of proteins while the physicochemical properties of the amino acids are responsible for the 3-dimensional structure of the proteins.

chapter 9

Prediction of two biological properties

9.1 Introduction

The last application developed during this PhD thesis using the proposed molecular descriptor based approach is a regression analysis performed on a peptide data set. Twenty peptide sequences taken from literature [Andersson *et al.* (1998)] have been described using molecular descriptors. Two different weighting schemes have been adopted, the physicochemical weighting scheme described in section 3.5.2 and the WHIM weighting scheme described in section 3.5.4. Due to the absence of the three-dimensional structure of the peptides only constitutional and auto-correlation descriptors have been calculated.

The analysis has been conducted independently between descriptors calculated with physicochemical and WHIM weighting scheme in order to evaluate the capabilities of the two different representation to highlight the relevant information needed to model the two biological responses.

Table 9.1: The twenty modelled peptide sequences and their values for the APTT response, both original and *log*-transformed.

Peptide	Sequence	APTT	Log(1 + APTT)
1	PKPRPDR	5.52	0.81
2	SWKHYW	0.58	0.2
3	SWKYYW	0.79	0.25
4	SWVDAW	1.56	0.41
5	RQGRYWL	1.5	0.4
6	PPGEMD	2.66	0.56
7	EGEGGM	1.58	0.41
8	RHWNIEGRPWWS	0.66	0.22
9	SEWAIEGRPHGW	1.21	0.34
10	FLRGEV	2.32	0.52
11	FMHLST	2.26	0.51
12	FMRPQM	4.14	0.71
13	FGWGQN	4.87	0.77
14	CWPMTRGC	1.09	0.32
15	KPRWWMWK	0.05	0.02
16	KSWQVWVK	0.8	0.26
17	KSWKYYWK	0.04	0.02
18	SWKYYWK	0.03	0.01
19	KSWKYYW	0.03	0.01
20	KMMSWKGK	0.7	0.23

9.2 Andersson Dataset

The twenty sequences evaluated in this application have been collected from the literature [Andersson *et al.* (1998)], these sequences belong to a peptide library of 190 hits from Pharmacia & Upjohn. The twenty considered peptides have different lengths, from 6 to 12 amino acids. All the twenty peptides showed activity with respect to the two biological responses modelled using molecular descriptors, the two responses are:

1. activated partial thromboplastin time (APTT);
2. thromboplastin time (TBPL).

The partial thromboplastin time (PTT) or activated partial thromboplastin time (aPTT or APTT) is a performance indicator measuring the efficacy of both the “intrinsic”(now referred to as the contact activation pathway) and the common coagulation pathways. Apart from detecting abnormalities in blood clotting,

Table 9.2: The twenty modelled peptide sequences and their values for the TBPL response, both original and *log*-transformed.

Peptide	Sequence	TBPL	Log(1 + TBPL)
1	PKPRPDR	17.4	1.26
2	SWKHYW	2.17	0.5
3	SWKYYW	2.34	0.52
4	SWVDAW	1.26	0.35
5	RQGRYWL	6.06	0.85
6	PPGEMD	3.04	0.61
7	EGEGGM	1.2	0.34
8	RHWNIEGRPWWS	0.71	0.23
9	SEWAIEGRPHGW	0.58	0.2
10	FLRGEV	1.94	0.47
11	FMHLST	3.5	0.65
12	FMRPQM	54	1.74
13	FGWGQN	14.64	1.19
14	CWPMTRGC	0.77	0.25
15	KPRWWMWK	0.13	0.05
16	KSWQVWVK	1.1	0.32
17	KSWKYYWK	0.75	0.24
18	SWKYYWK	1.5	0.4
19	KSWKYYW	0.71	0.23
20	KMMSWKGK	0.49	0.17

it is also used to monitor the treatment effects with heparin, a major anticoagulant.

The biological activities are expressed as 50% inhibition concentration (IC_{50}) in μM, since the biological activities ranged from 0.03 to 5.52 for APTT and from 0.13 to 54 for TBPL, a *log* transformation have been performed prior to modelling. The twenty peptide sequences, both original and *log*-transformed values are showed in Table 9.1 for APTT and Table 9.2 for TBPL.

9.3 Results and Discussions

In order to model the two biological responses considered in this application two different descriptors blocks have been calculated, constitutional and 2-D autocorrelation descriptors. Three dimensional descriptors have not been calculated due to the missing information related to the 3-dimensional structure of the considered peptides.

Table 9.3: A summary of the final models obtained for APTT response using the physicochemical weighting scheme.

Size	**Variables**	**R2**	**Q2**	**Q2boot**
4	nPro nTrp nAla/nAAs nAsp/nAAs	88.3	83.41	81.02
4	ATS6mw ATS1hyl MATS4hyl GATS4p	88.84	81.29	76.34
4	nTrp nLys/nAAs nMet/nAAs GATS1hyb	89.21	80.46	75.25
3	ATS6mw ATS1hyl MATS4hyl	84.84	75.43	74.6
3	nTrp nLys/nAAs GATS1hyb	84.09	75.34	73.34
3	Whyb_sum nPhe nPro	84.79	73.94	72.27
2	nTrp nAsp/nAAs	73.08	65.8	65.1
2	nPro/nAAs ATS3hyb	72.73	60.49	56.73
2	ATS5hyb GATS5mw	65.78	55.34	56.1

Table 9.4: A summary of the final models obtained for TBPL response using the physicochemical weighting scheme.

Size	**Descriptors**	**R2**	**Q2**	**Q2boot**
4	nPhe nGlu/nAAs nPro/nAAs MATS7ras	88.21	73.54	67.12
4	nGln nTrp nArg/nAAs nGlu/nAAs	85.14	73.52	56.25
4	ATS3ras ATS5ras MATS4ras GATS2ras	85.36	73.48	68.37
3	nGlu/nAAs nPro/nAAs MATS7ras	83.96	68.04	62.8
3	nGln nAsp/nAAs nGlu/nAAs	65.83	58.1	47.62
3	ATS1ras ATS5ras GATS5mw	73.31	54.8	51.36
2	nGlu/nAAs MATS7ras	67.14	51.68	51.53
2	nTrp nGlu/nAAs	65.06	46.01	43.72
2	ATS3mw ATS1ras	53.28	36.71	34.37

The two calculated descriptor blocks have been weighted using two different weighting schemes: the physicochemical and the WHIM weighting scheme. The molecular descriptors calculated with the physicochemical and the WHIM weighting scheme have been considered separately but following the same approach. Once calculated the molecular descriptors models have been built using Genetic Algorithms (GAs) [Goldberg (1989), Leardi *et al.* (1992), Leardi (1994, 2001), Todeschini *et al.* (2003)] as implemented in the MobyDigs package [Todeschini *et al.* (2003), Mob (2007)] in order to select subsets of variables that maximise the predictive power of the multivariate models.

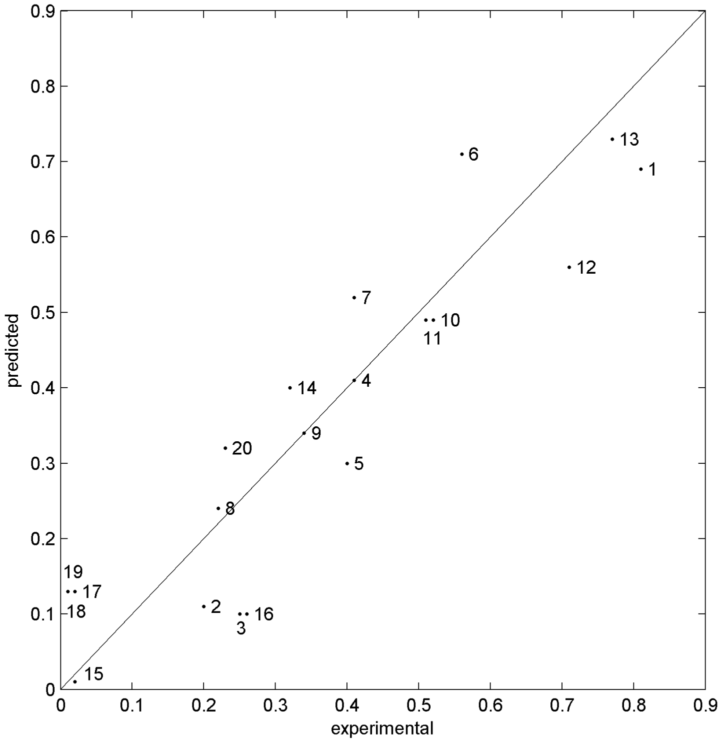


Figure 9.1: Experimental vs. predicted values of log(1 + APTT) for the best model with 4 variables (nPro, nTrp, nAla/nAAs, nAsp/nAAs) using physicochemical weights. (Q^2 = 83.41).

9.3.1 Physicochemical weighting scheme

The properties used to characterise the amino acids with the physicochemical weighting scheme are:

1. molecular weight [Fasman (1976)] (FASG760101);
2. polarity [Grantham (1974)] (GRAR740102);
3. hydrophobicity [Jones (1975)] (JOND750101);
4. residue accessible surface area in folded protein [Chothia (1976)] (CHOC76010);
5. hydrophilicity scale [Kuhn *et al.* (1995)] (KUHL950101).

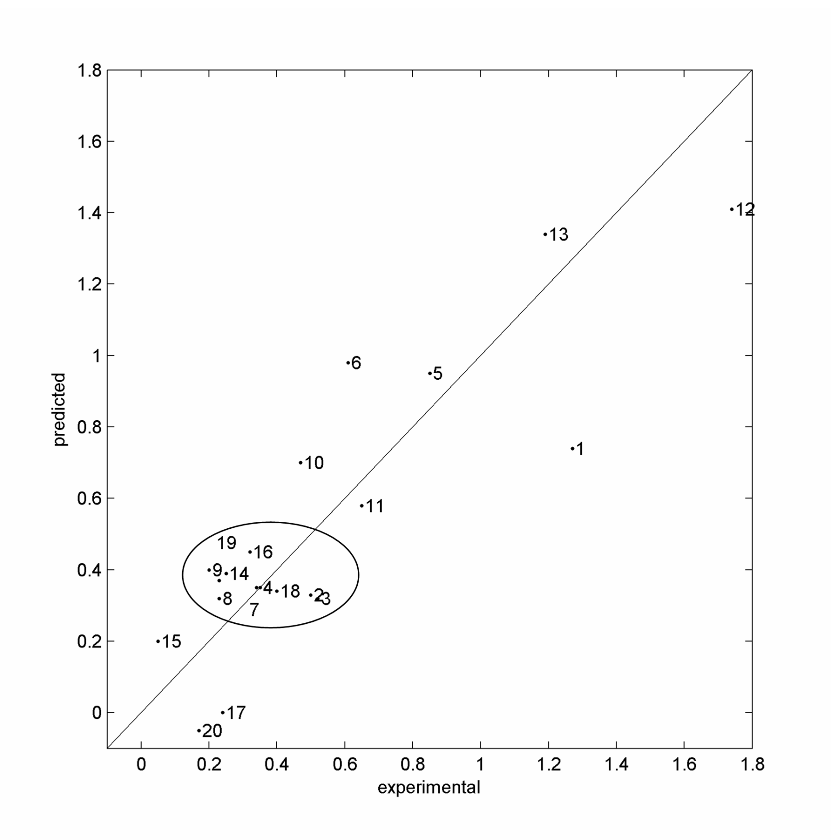


Figure 9.2: Experimental vs. predicted values of log(1 + TBPL) for the best model with 4 variables (nPhe, nGlu/nAAs, nPro/nAAs, MATS7ras) using physicochemical weights. (Q^2 = 73.54).

The physicochemical weighting scheme is described in section 3.5.2.

Physicochemical weighting scheme collects five different properties; constitutional descriptor block consists of 51 descriptors and 2D autocorrelation descriptors consist of 120 molecular descriptors.

The two different responses (APTT and TBPL) have been modelled separately using in both cases genetic algorithms in order to perform the variable subset selection (see section 5.4). For both biological responses the following steps have been performed:

1. two different variable populations have been created, the first one collecting all the constitutional descriptors (51 descriptors) and the second one

collecting all the autocorrelation descriptors (120 descriptors);

2. the selected fitness function was Q^2 leave-one-out (see section 5.5.1);
3. a preliminary all subset model approach has been performed looking for the best models with two variables;
4. the maximum number of variables for each model has been set to three variables;
5. once the two populations have been stabilised one new population has been created merging the constitutional and the autocorrelation populations collecting all 171 variables and preserving the best models;
6. the maximum number of variables for each model has been then increased to four variables;
7. the best five models have been retained from each population;
8. the stability of the selected models has been tested with bootstrap and y-scrambling analysis.

Three different model populations have been obtained for each response. A model population collecting constitutional descriptors, a model population collecting autocorrelation descriptors and a model population collecting both descriptor blocks.

In Table 9.3 and Table 9.4 the best models for dimensions between 2 and 4 are listed for APTT and TBPL responses respectively. Models with two descriptors are obtained using the all subset selection (see section 5.4.1) while models with three and four variables have been obtained using the genetic algorithms. Models with same dimension obtained using constitutional, autocorrelation or both descriptor blocks had similar predictive power. The APTT response is modelled better than the TBPL response. Models being constituted by four variables have a Q^2 ranging between 80.46 to 83.41 for APTT response while models with four variables obtained for TBPL had a Q^2 ranging between 73.48 to 73.54.

Looking deeply at the best models emerge that APTT response is modelled using different constitutional descriptors, the most frequent are number of prolines (*nPro*) and number of tryptophan (*nTrp*), one of these two descriptors occurs in every model containing at least one constitutional descriptor. Autocorrelation descriptors mostly used are weighted by molecular weight (*mw* suffix), hydrophobicity (*hyb*) and hydrophilicity scale (*hyl*). Only one model among the best ones include also autocorrelation descriptors weighted by polarity (*p*). No models include autocorrelation descriptors weighted by residue accessible surface area (*ras*). The best model obtained for APTT response using four molecular descriptors is represented in Figure 9.1.

TBPL response conversely is better modelled by autocorrelation descriptors weighted by residue accessible surface area (*ras*). The best 4-dimensional autocorrelation descriptors model is constituted only by descriptors weighted by residue accessible surface area. Only two models include a molecular descriptor weighted by molecular weight. No models for TBPL include autocorrelation descriptors weighted by hydrophobicity, hydrophilicity or polarity. The most frequent constitutional descriptors is the relative frequency of glutamic acid ($nGlu/nAAs$) in a single peptide, this descriptor is selected in all models containing at least one constitutional descriptor. The best model obtained for TBPL response using four molecular descriptors is represented in Figure 9.2.

The final models have been further validated by bootstrap [Efron (1979, 1982, 1987)] and response permutations [Lindgren *et al.* (1996), Eriksson *et al.* (1997)] (see sections 5.5.2, 5.5.3). In order to calculate the average predictive power (Q^2_{BOOT}), bootstrap procedure is repeated 5000 of time, Q^2_{BOOT} values are reported in Table 9.3 and Table 9.4. Y-scrambling procedure is repeated 300 of times. Once the model validation has been performed the Y-scrambling parameters ($a(R^2)$ and $a(Q^2)$) are calculated, final values of $a(R^2)$ and $a(Q^2)$ for the models reported in table Table 9.3 and Table 9.4 are included in the expected limits.

9.3.2 WHIM weighting scheme

The same data set studied weighting the molecular descriptors by physicochemical properties has been studied also weighting the calculated descriptors by the WHIM indices. The properties considered in the WHIM weighting scheme are:

1. Am (WHIM global dimension descriptor / weighted by atomic masses, scaled);
2. Km (WHIM shape descriptor / weighted by atomic masses);
3. Dm (WHIM global density descriptor / weighted by atomic masses).

The WHIM weighting scheme is described in section 3.5.4.

Table 9.5: A summary of the final models obtained for APTT response using the WHIM weighting scheme.

Size	Descriptors	R2	Q2	Q2boot
4	nPro nArg/nAAs ATS4Am GATS5Dm	95.76	92.86	89.52
4	ATS3Km ATS1Dm ATS2Dm GATS1Dm	92.14	84.37	81.8
3	nPro ATS4Am GATS5Dm	92.98	88.78	87.9
3	WAm_sum nAsn/nAAs nAsp/nAAs	82.47	73.6	69.44
3	ATS6Km ATS1Dm MATS2Dm	80.15	70.86	68.64
2	GATS2Dm GATS6Dm	72.47	60.51	60.76

Table 9.6: A summary of the final models obtained for TBPL response using the WHIM weighting scheme.

Size	Models	R2	Q2	Q2boot
4	nPro nGlu / nAAs ATS2Km ATS4Km	86.07	71.25	64.7
4	ATS1Dm ATS2Dm GATS3Km GATS1Dm	78.48	60.77	55.94
3	nGlu / nAAs nPro / nAAs MATS7Km	80.32	61.93	45.19
3	ATS1Dm ATS2Dm GATS1Dm	72.8	57.9	55.18
2	nGlu / nAAs MATS7Km	64.21	48.62	39.67
2	ATS1Dm ATS2Dm	55.5	35.08	33.15

WHIM weighting scheme comprise three different properties and the resulting calculated descriptors are 47 constitutional and 72 autocorrelation descriptors.

Exactly as for the data set obtained by molecular descriptors calculated using the physicochemical weighting scheme the two different responses (APTT and TBPL) have been modelled separately using in both cases genetic algorithms in order to perform the variable subset selection. The following steps have been performed:

1. two different variable populations have been created, the first one collecting all the constitutional descriptors (47 descriptors) and the second one collecting all the autocorrelation descriptors (72 descriptors);

2. the selected fitness function has been Q^2 leave-one-out (see section 5.5.1);

3. a preliminary all subset model approach has been performed looking for the best models with two variables;

4. the maximum number of variables for each model has been set to three variables;

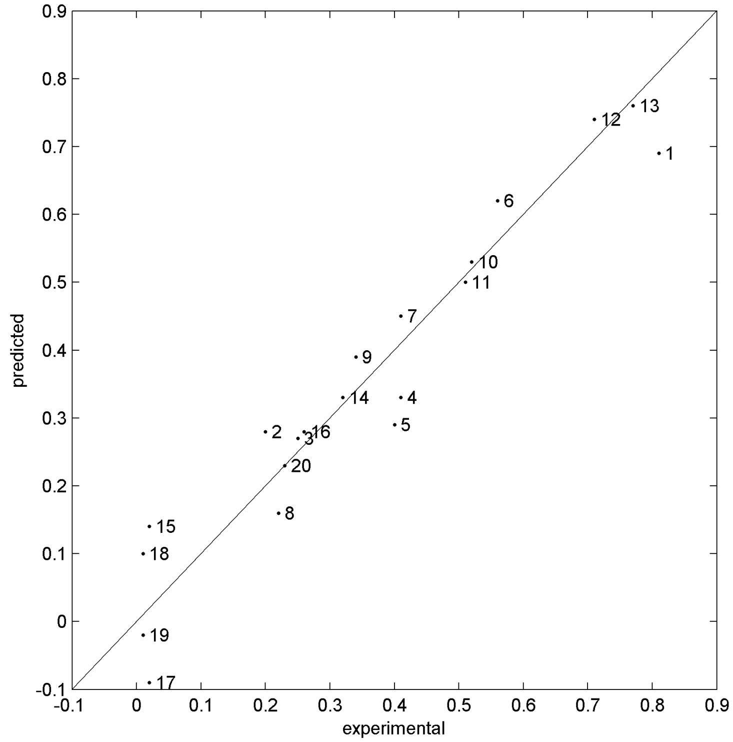


Figure 9.3: Experimental vs. predicted values of log(1 + APTT) for the best model with 4 variables (nPro, nArg/nAAs, ATS4Am, GATS5Dm) using WHIM weights (Q^2 = 92.86).

5. once the two populations have been stabilised one new population has been created merging the constitutional and the autocorrelation populations collecting all 119 variables and preserving the best models;

6. the maximum number of variables for each model has been increased to four variables;

7. the best five models have been retained from each population;

8. the stability of the selected models has been tested with bootstrap and y-scrambling analysis.

Three different model populations have been obtained for each response. The first one comprising constitutional descriptors, the second one the 2D autocorrelation descriptors and the last one collecting both descriptor blocks.

The best models for dimensions between 2 and 4 are listed in Table 9.5 and Table 9.6 for APTT and TBPL responses respectively. Models with two descriptors are obtained using the all subset selection while models with three and four variables have been obtained using the genetic algorithms. Only one model being constituted only by constitutional descriptors is reported in Table 9.5, it includes the average sum of the WHIM global dimension index (*WAm_sum*). Anyway, considering APTT response the models are significantly better than the models obtained using the physicochemical descriptors. The best model, being constituted by two constitutional and two autocorrelation descriptors, has a Q^2 equal to 92.86% that is more than ten points higher than the best model with four variables obtained using the physicochemical descriptors. The best model with four variables being constituted only by autocorrelation descriptors has a Q^2 equal to 84.37%.

Mostly used autocorrelation descriptors are weighted by WHIM global density index (*Dm* suffix), these descriptors appear in all the models containing at least one autocorrelation descriptor. Descriptors calculated using the WHIM global shape index (*Km*) and WHIM global dimension index (*Am*) occurs in two of the five models containing autocorrelation descriptors. Considering the constitutional descriptors number of prolines (*nPro*) occurs in two different models, always beside *ATS4Am* and *GATS5Dm* autocorrelation descriptors. The best model obtained for APTT response using four molecular descriptors weighted by WHIM indices is represented in Figure 9.3.

Models obtained for TBPL response using the WHIM weighting scheme have a lower predictive power compared to those obtained using physicochemical descriptors. Models constituted only by constitutional descriptors are omitted in Table 9.6 due to the fact that are the same models reported in Table 9.4 because no models for TBPL include weighted constitutional descriptors.

Like for the previous section the final models have been further validated by bootstrap and response permutations. In order to calculate the average predictive power (Q^2_{BOOT}), bootstrap procedure is repeated 5000 of time, Q^2_{BOOT} values are reported in Table 9.5 and Table 9.6. Y-scrambling procedure is repeated 300 of times. Once the model validation has been performed the Y-scrambling parameters ($a(R^2)$ and $a(Q^2)$) are calculated, final values of $a(R^2)$ and $a(Q^2)$ for the models reported in table Table 9.5 and Table 9.6 are included in the expected limits.

9.4 Conclusions

This final application confirm the capability of the proposed methodology to model responses of a considered data set of peptide of different lengths. The models obtained using the proposed methodology are significantly better than the models taken from literature [Andersson *et al.* (1998)], both using the physicochemical and the WHIM weighting scheme.

APTT response is better modelled than TBPL response, the reason is probably due to the not homogeneous distribution of the response values for TBPL. In Figure 9.2 a cluster 10 peptides among 20 with response values between 0.2 and 0.6 is highlighted by an oval. This kind of distribution, where a small portion of the response space is deeply described and a lot of the response space is not well represented, is usually an obstacle to build good models.

The results obtained using the physicochemical weighting scheme confirm the capability of the presented simplified representation of the peptide structure to describe a peptidic data set. The capability of the WHIM weighting scheme to improve the predictive power of the molecular descriptor models can be conducted to the 3-dimensional information contained by the WHIM global dimension indices used as weighting scheme. WHIM global dimension indices are calculated on the 3-dimensional structure of the isolated amino acids.

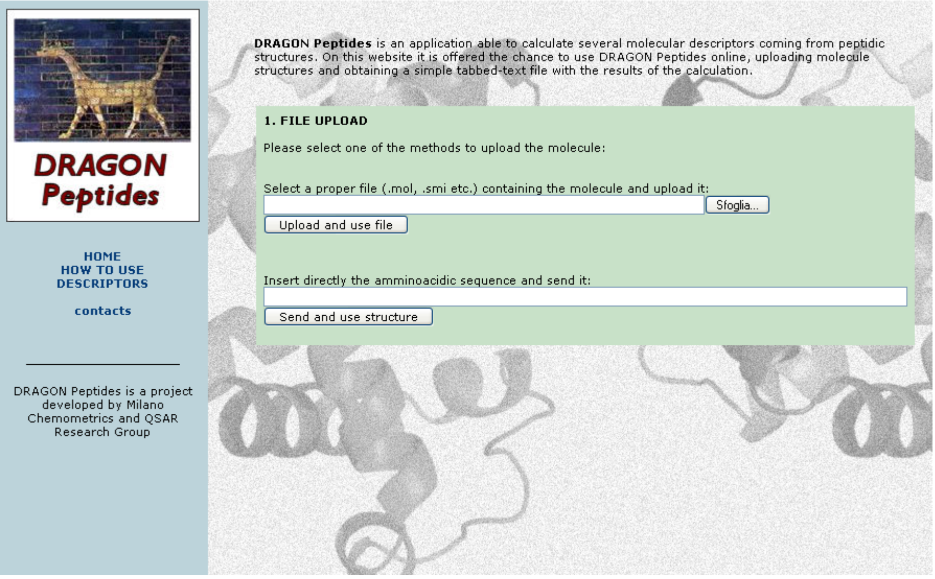
chapter 10

A Web-based Application

10.1 Introduction

Some resources in order to model peptide sequences are available on the web. One of them is the amino acid index database[Nakai *et al.* (1986), Tomii and Kanehisa (1996), Kawashima *et al.* (1999), Kawashima and Kanehisa (2000)], described in section 3.4.2. Another database freely available online is the JenPep database [Blythe *et al.* (2002)], the database is available via the Internet. An HTML interface allowing searching of the database can be found at the following address: http://www.jenner.ac.uk/JenPep. JenPep is a family of relational databases, it contains quantitative data on peptide binding to Major Histocompatibility Complexes (MHCs) [Burden and Winkler (2005), Doytchinova and Flower (2007b), Doytchinova *et al.* (2004)] and to Transmembrane Peptide Transporter (TAP), as well as an annotated list of T-cell epitopes [Doytchinova and Flower (2001), Doytchinova *et al.* (2006), Doytchinova and Flower (2006a)]. AntiJen is the successor of JenPep, it is a database system focused on the integration of kinetic, thermodynamic, functional, and cellular data within the context of immunology and vaccinology [Toseland *et al.* (2005)]. Another resource is MHCpred and its last version MHCpred 2.0 that are Perl implementation of partial least squares-based, multivariate statistical method for the quantitative prediction of peptide binding to major histocompatibility complexes (MHCs) [Guan *et al.* (2003b,a, 2006), Hattotuwagama *et al.* (2004a,b, 2006)].

During the PhD thesis, a web-based application have been developed. This application allow the calculation of the molecular descriptors described in this thesis. The calculation can be performed both on peptides and proteins using the physicochemical weighting scheme. This resource, called DragonP, is available on line at: http://www.michem.unimib.it/dragonP/



DRAGON Peptides is a project developed by Milano Chemometrics and QSAR Research Group

DRAGON Peptides is an application able to calculate several molecular descriptors coming from peptidic structures. On this website it is offered the chance to use DRAGON Peptides online, uploading molecule structures and obtaining a simple tabbed-text file with the results of the calculation.

1. FILE UPLOAD

Please select one of the methods to upload the molecule:

Select a proper file (.mol, .smi etc.) containing the molecule and upload it:

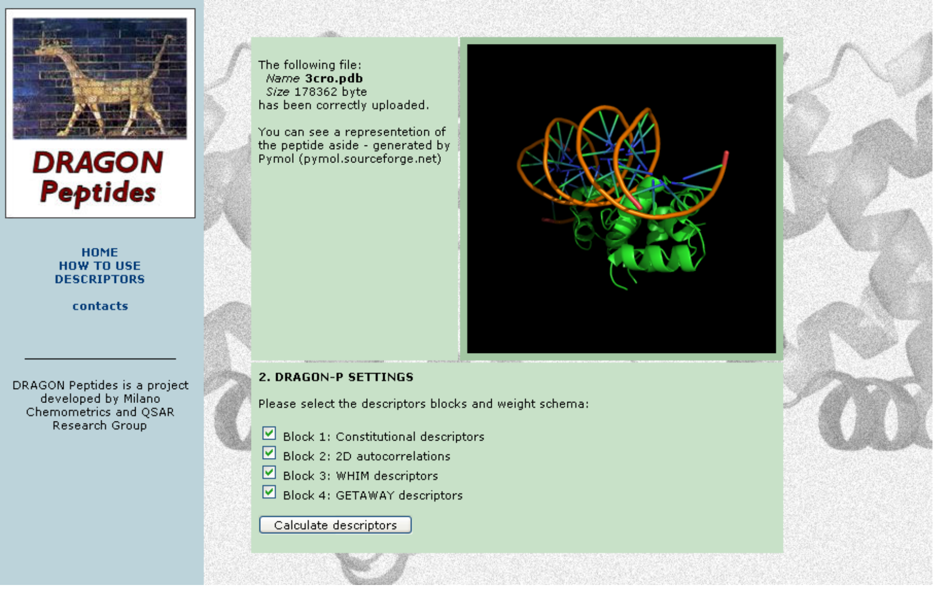
Insert directly the amminoacidic sequence and send it:

Figure 10.1: Screenshot of DragonP web application. File upload, peptides can loaded uploading a file or directly typing the peptide sequence.

10.2 Description

The DragonP Web Application is a dedicated web site that allows the execution of the DragonP application in batch mode, after uploading a molecule. This project has been implemented on a dedicated server located at the Milano Chemometrics and QSAR Research Group, on which is also installed the beta version of DragonP for Linux; the batch execution of the software is made possible thanks to PHP technology.

In the homepage of the project, some information about DragonP can be found, such as the explanation of the four blocks of descriptors that can be calculated, together with a short how-to for the website.



DRAGON Peptides is a project developed by Milano Chemometrics and QSAR Research Group

The following file:
Name **3cro.pdb**
Size 178362 byte
has been correctly uploaded.

You can see a representetion of the peptide aside - generated by Pymol (pymol.sourceforge.net)

2. DRAGON-P SETTINGS

Please select the descriptors blocks and weight schema:

☑ Block 1: Constitutional descriptors
☑ Block 2: 2D autocorrelations
☑ Block 3: WHIM descriptors
☑ Block 4: GETAWAY descriptors

Figure 10.2: DragonP web application. DragonP settings, choice of the molecular descriptor blocks and view of the protein tertiary structure visualised by PyMol.

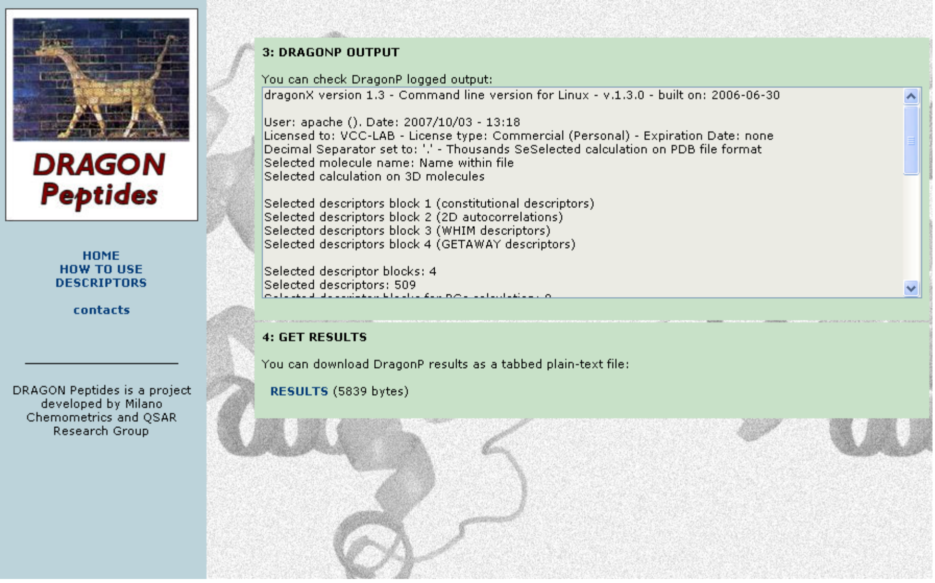
10.2.1 File upload

In the first step it is necessary to upload a proper peptide. It is possible to choose two ways, as DragonP can be executed on a molecule file containing a peptide, or just by inserting directly an amino acid sequence.

Molecule files shall be of a known format; DragonP can handle the following file format: HYPERCHEM files (*.hin), Tripos files (*.mol) or MDL files. Amino acid sequences can be inserted in 1-letter format (such as AMTMA) or in 3-letters format (such as AlaMetThrMetAla). The screenshot of the file upload window is shown in Figure 10.1.

10.2.2 DragonP settings

Once the target peptide is given, some information about the file uploaded are shown on the website, first of all if the file has been correctly uploaded. On the right panel, the given molecule is shown, using the tertiary structure representation; this is made possible using the PyMol Molecular Graphic System (http://www.pymol.org) which runs on the webserver. In the lower panel,the



DRAGON Peptides is a project developed by Milano Chemometrics and QSAR Research Group

3: DRAGONP OUTPUT

You can check DragonP logged output:

```
dragonX version 1.3 - Command line version for Linux - v.1.3.0 - built on: 2006-06-30

User: apache (). Date: 2007/10/03 - 13:18
Licensed to: VCC-LAB - License type: Commercial (Personal) - Expiration Date: none
Decimal Separator set to: '.' - Thousands SeSelected calculation on PDB file format
Selected molecule name: Name within file
Selected calculation on 3D molecules

Selected descriptors block 1 (constitutional descriptors)
Selected descriptors block 2 (2D autocorrelations)
Selected descriptors block 3 (WHIM descriptors)
Selected descriptors block 4 (GETAWAY descriptors)

Selected descriptor blocks: 4
Selected descriptors: 509
```

4: GET RESULTS

You can download DragonP results as a tabbed plain-text file:

RESULTS (5839 bytes)

Figure 10.3: DragonP web application. DragonP output, view of the log file and button to download the calculated descriptors in a tabbed text-file.

user can choose which of the four descriptors block has to be included in the calculation. The screenshot of the DragonP settings window is shown in Figure 10.2.

10.2.3 DragonP output

The output log of the calculation is shown in a box, so that the user can check if the application have been run correctly. The log file produced by DragonP includes some information about the calculation or rejection of the molecule, the input file format, the selected descriptors and the calculation time.

The final results of the calculation are stored in a plain-text, tab-separated file. It can be downloaded by clicking on the given link. The text format is easily imported and manipulated by most of the software, as it simply reports a table with the descriptors values on each column, and the molecules on each row. The screenshot of the DragonP output window is shown in Figure 10.3.

chapter 11

Conclusions and Perspectives

This PhD thesis presents a methodology for the characterisation of protein and peptide sequences and structures by means of a molecular descriptor based approach. In the first part of the thesis the state of the art related to protein and peptide characterisation using chemometric methods is presented together with the theory that support the proposed approach.

In the last years several methodology and applications have been proposed in the literature. Actually most of the applications are related to the description of short peptides in order to predict chemical and biological properties using the z-scores approach [Hellberg *et al.* (1987), Sjstrm *et al.* (1995), Sandberg *et al.* (1998), Andersson *et al.* (1998), Edman *et al.* (1999), Nystrm *et al.* (2000), Doytchinova *et al.* (2002), Doytchinova and Flower (2003), Guan *et al.* (2005), Doytchinova and Flower (2005, 2006b,a, 2007b,a)].

On the contrary the proposed methodology is related to an holistic representation of the molecular structure using molecular descriptors and can be applied both on short peptides and on big proteins being constituted on a huge amount of amino acids. Nowadays the traditional molecular descriptor based approach is inapplicable on big molecules such as polypeptides and proteins, since an atom based representation impede the calculation of molecular descriptors on complex molecules represented by thousands of atoms. The amino acid based representation studied and presented in this thesis avoids the problems related to the calculation of molecular descriptors on big molecules; moreover it prevents the problems related to information redundancy correlated to the common structural features shared by all amino acids, including an α-carbon to which an amino group, a carboxyl group, and a variable side chain are bonded.

The present study demonstrates that the proposed approach is able to provide valuable information on the characterisation of peptides and proteins.

The presented methodology has been deeply evaluated in the second part of this thesis where three different applications of the methodology are described. In chapter 7 a sensitivity analysis has been performed, an artificial data set have been used in order to evaluate the capability of two different descriptor blocks (constitutional and auto-correlation descriptors), to be able to discriminate among different mutated amino acids. The conducted analysis showed that also small changes on a peptide sequence are highlighted using the proposed descriptor based approach suggesting that molecular descriptors are able to discriminate among peptide sequences that differ only on a small portion of the amino acid sequence.

The second application presented in chapter 8 has been developed on two different protein folds evaluating how different weighting schemes (i.e. different amino acid representations), affect the information collected by diverse blocks of molecular descriptors. Particularly, it has been showed that, constitutional descriptors are more informative if calculated using the statistical weighting scheme especially on the Zincin-like fold, while 3-dimensional descriptors showed a clearer separation of the different families and superfamilies if calculated using the physicochemical weighting scheme. The proposed approach differs from the traditional comparison methodology due to its alignment independent comparison.

Finally a practical application is described in chapter 9, where a peptide data set taken from literature has been described using the proposed approach. Molecular descriptors have been calculated on the peptide sequences considering separately two different weighting schemes, the physicochemical and the WHIM weighting scheme. Both amino acid characterisations, combined with genetic algorithms for variable subset selection, produced models that are considerably better than models taken from the literature.

In conclusion, the proposed approach has given encouraging results, both on peptides and on proteins characterisation. Anyway, the studied methodology could be more improved and studied. A lot of molecular descriptors can be evaluated in order to be applied on peptide and protein characterisation. The evaluation of different weighting scheme can be deeply analysed due to the fact that different amino acid characterisations can highlight different kind of information suggesting that a correct choice of the weighting scheme and molecular descriptor types are related to the information content.

Bibliography

(2007). Talete srl - mobydigs for windows (software for the calculation of regression models using genetic algorithms for variable selection), version 1.0. [citations in 5.4.3 and 9.3]

(2007). Talete srl, dragon for linux - software for molecular descriptors calculation. [citation in 1.2]

Andersson, P. M., Sjstrm, M., and Lundstedt, T. (1998). Preprocessing peptide sequences for multivariate sequence-property analysis. *Chemometrics and Intelligent Laboratory Systems*, **42**, 41–50. [citations in 1.1, 6, 9.1, 9.2, 9.4, and 11]

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2004). Scop database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res*, **32**, 226–229. [citations in 2, 3.5.3, 6, 8.1, 8.2, and 8.3]

Balaban, A. T. (1976). *Chemical Applications of Graph Theory*. Academic Press. [citation in 3]

Baumann, K. and Stiefl, N. (2004). Validation tools for variable subset regression. *Journal of Computer-Aided Molecular Design*, **18**, 549–562. [citation in 5.5.1]

Blaber, M., Baase, W. A., Gassner, N., and Matthews, B. W. (1995). Alanine scanning mutagenesis of the alpha-helix 115-123 of phage t4 lysozyme: effects on structure, stability and the binding of solvent. *Journal of Molecular Biology*, **246**, 317–330. [citation in 3.5.3]

Blythe, M. J., Doytchinova, I. A., and Flower, D. R. (2002). Jenpep: a database of quantitative functional peptide data for immunology. *Bioinformatics*, **18**(3), 434–439. [citation in 10.1]

Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, and R.; Zaliani, A. (1997). Ms-whim, new 3d theoretical descriptors derived from molecular surface properties: A comparative 3d qsar study in a series of steroids. *J. Comput.-Aided Mol. Des.*, **11**, 79–92. [citation in 3.2]

Brenner, S. E., Chothia, C., Hubbard, T. J., and Murzin, A. G. (1996). Understanding protein structure: using scop for fold interpretation. *Methods Enzymol*, **266**, 635–643. [citation in 3.5.3]

Brosnan, J. and Brosnan, M. (2006). The sulfur-containing amino acids: an overview. *J Nutr*, **136**, 1636–1640. [citation in 3.5.4]

Broto, P., Moreau, G., and Vandicke, C. (1984a). Molecular structures: Perception, autocorrelation descriptor and sar studies. *Eur.J.Med.Chem.*, **19**, 79–84. [citation in 4.3.1]

Broto, P., Moreau, G., and Vandicke, C. (1984b). Molecular structures: perception, autocorrelation descriptor and sar studies. *Eur.J.Med.Chem.*, **19**, 71–78. [citation in 4.3.1]

Broto, P., Moreau, G., and Vandicke, C. (1984c). Molecular structures: Perception, autocorrelation descriptor and sar studies. *Eur.J.Med.Chem.*, **19**, 66–70. [citation in 4.3.1]

Burden, F. R. and Winkler, D. A. (2005). Predictive bayesian neural network models of mhc class ii peptide binding. *J Mol Graph Model*, **23**(6), 481–489. [citation in 10.1]

Burden, F. R., Brereton, R. G., and Walsh, P. T. (1997). Cross-validatory selection of test and validation sets in multivariate calibrationand neural networks as applied to spectroscopy. *Analyst*, **122**, 1015–1022. [citation in 5.5.1]

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.*, **105**, 1–14. [citations in 4, 4, 4, 4, III, and 11.1]

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826. [citation in 3.5.3]

Cocchi, M. and Johansson, E. (1993). Amino acids characterization by grid and multivariate data analysis. *Quant. Struct.-Act. Relat.*, **12**, 1–8. [citation in 3.2]

Collantes, E. R. and Dunn III, W. J. (1995). Amino acids side chain descriptors for quantitative structure-activity relationship studies of peptide analogues. *J. Med. Chem.*, **38**, 2705–2713. [citation in 3.2]

Consonni, V. and Todeschini, R. (2001). *Rational Approaches to Drug Design*, pages 235–240. Prous Science, Barcelona (Spain). [citation in 4.4.2]

Consonni, V., Todeschini, R., and Pavan, M. (2002a). Structure/response correlations and similarity/diversity analysis by getaway descriptors. 1. theory of the novel 3d molecular descriptors. *J. Chem. Inf. Comp. Sci.*, **42**, 682–692. [citation in 4.4.2]

Consonni, V., Todeschini, R., Pavan, M., and Gramatica, P. (2002b). Structure/response correlations and similarity/diversity analysis by getaway descriptors. 2. application of the novel 3d molecular descriptors to qsar/qspr studies. *J.Chem.Inf.Comp.Sci.*, **42**, 693–705. [citation in 4.4.2]

Curis, E., Nicolis, I., Moinard, C., Osowska, S., Zerrouk, N., Bnazeth, S., and Cynober, L. (2005). Almost all about citrulline in mammals. *Amino Acids*, **29**, 177–205. [citation in 3.5.4]

Cygler, M., Schrag, J. D., Sussman, J. L., Harel, M., Silman, I., K., G. M., and Doctor, B. P. (1993). Relationship between sequence conservation and threedimensional structure in a large family of esterases, lipases, and related proteins. *Protein Sci*, **2**, 366–382. [citation in 3.5.3]

Damodaran, S. (1995). *Structure-function relationship of food proteins*, volume In Protein Functionality in Food Systems, pages 1–37. Marcel Dekker, New York. [citation in 4.2]

Doolittle, R. F. (1981). Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159. [citation in 3.5.3]

Doytchinova, I. A. and Flower, D. R. (2001). Toward the quantitative prediction of t-cell epitopes: comfa and comsia studies of peptides with affinity for the class i mhc molecule hla-a*0201. *J Med Chem*, **44**(22), 3572–3581. [citation in 10.1]

Doytchinova, I. A. and Flower, D. R. (2003). Towards the in silico identification of class ii restricted t-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics*, **19**(17), 2263–2270. [citations in 1.1 and 11]

Doytchinova, I. A. and Flower, D. R. (2005). In silico identification of supertypes for class ii mhcs. *J Immunol*, **174**(11), 7085–7095. [citations in 1.1 and 11]

Doytchinova, I. A. and Flower, D. R. (2006a). Class i t-cell epitope prediction: improvements using a combination of proteasome cleavage, tap affinity, and mhc binding. *Mol Immunol*, **43**(13), 2037–2044. [citations in 1.1, 10.1, and 11]

Doytchinova, I. A. and Flower, D. R. (2006b). Modeling the peptide-t cell receptor interaction by the comparative molecular similarity indices analysis-soft independent modeling of class analogy technique. *J Med Chem*, **49**(7), 2193–2199. [citations in 1.1 and 11]

Doytchinova, I. A. and Flower, D. R. (2007a). Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. *Vaccine*, **25**(5), 856–866. [citations in 1.1 and 11]

Doytchinova, I. A. and Flower, D. R. (2007b). Predicting class i major histocompatibility complex (mhc) binders using multivariate statistics: comparison of discriminant analysis and multiple linear regression. *J Chem Inf Model*, **47**(1), 234–238. [citations in 1.1, 10.1, and 11]

Doytchinova, I. A., Blythe, M. J., and Flower, D. R. (2002). Additive method for the prediction of protein-peptide binding affinity. application to the mhc class i molecule hla-a*0201. *J Proteome Res*, **1**(3), 263–272. [citations in 1.1 and 11]

Doytchinova, I. A., Walshe, V. A., Jones, N. A., Gloster, S. E., Borrow, P., and Flower, D. R. (2004). Coupling in silico and in vitro analysis of peptide-mhc binding: a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. *J Immunol*, **172**(12), 7495–7502. [citation in 10.1]

Doytchinova, I. A., Guan, P., and Flower, D. R. (2006). Epijen: a server for multistep t cell epitope prediction. *BMC Bioinformatics*, **7**, 131. [citation in 10.1]

Driscoll, D. and Copeland, P. (2003). Mechanism and regulation of selenoprotein synthesis. *Annu Rev Nutr*, **23**, 17–40. [citation in 3.5.4]

Edman, M., Jarhede, T., Sjstrm, M., and Wieslander, A. (1999). Different sequence patterns in signal peptides from mycoplasmas, other gram-positive bacteria, and escherichia coli: a multivariate data analysis. *Proteins*, **35**(2), 195–205. [citations in 1.1 and 11]

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26. [citations in 5.5.2 and 9.3.1]

Efron, B. (1982). *The Jackknife, the Boostrap and Other Resampling Methods*. Society for Industrial and Applied mathematics, Philadelphia, PA. [citations in 5.5.2 and 9.3.1]

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of American Statistical Association*, **82**, 171–200. [citations in 5.5.2 and 9.3.1]

Eriksson, L., Johansson, E., and Wold, S. (1997). *Quantitative Structure-Activity Relationship Model Validation. In Quantitative Structure-Activity Relationships in Environmental Sciences*, volume VII, page 381397. SETAC Press, Pensacola, FL. [citations in 5.5.3 and 9.3.1]

Fasman, G., editor (1976). *Handbook of Biochemistry and Molecular Biology*, volume 1. CRC Press, Cleveland. [citations in 1, 1, 1, 1, and 11.1]

Fauchere, J. and Pliska, V. (1983). Hydrophobic parameters of amino acid side chain from the partitioning of n-acetyl-amino-acid amides. *Eur.J. Med. Chem.*, **4**, 369–375. [citation in 3.2]

Fligner, K. L. and Mangino, M. E. (1991). *Relationship of composition to protein functionality*, volume Interactions of Food Proteins. American Chemical Society, Washington DC. [citation in 4.2]

Free, S. M. J. and Wilson, J. W. (1964). A mathematical contribution to structure activity studies. *J. Med. Chem.*, **7**, 395–399. [citation in 2.2]

Gallop, A., Barrett, R., Dower, W., Fodor, S., and Gordon, E. (1994). Applications of combinatorial technologies to drug discovery. 1. background and peptide combinatorial libraries. *J. Med. Chem.*, **37**, 1233–1251. [citation in 1.1]

Gancia, E., Bravi, G., Mascagni, P., and Zaliani, A. (2000). Global 3d-qsar methods: Ms-whim and autocorrelation. *J Comput Aided Mol Des*, **14**(3), 293–306. [citation in 3.2]

Geary, R. (1954). The contiguity ratio and statistical mapping. *Incorp. Statist.*, **5**, 115–145. [citation in 4.3.1]

Golbraikh, A. and Tropsha, A. (2002). Beware of q2! *Journal of Molecular Graphics and Modelling*, **20**, 269–276. [citation in 5.5.1]

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley. [citations in 5.4.3 and 9.3]

Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849–857. [citation in 3.2]

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864. [citations in 2, 2, 2, 2, and 11.1]

Gu, Y. Z., Hogenesch, J. B., and Bradfield, C. A. (2000). The pas superfamily: sensors of environmental and developmental signals. *Annu Rev Pharmacol Toxicol*, **40**, 519–561. [citation in 8.3]

Guan, P., Doytchinova, I. A., Zygouri, C., and Flower, D. R. (2003a). Mhcpred: A server for quantitative prediction of peptide-mhc binding. *Nucleic Acids Res*, **31**(13), 3621–3624. [citation in 10.1]

Guan, P., Doytchinova, I. A., Zygouri, C., and Flower, D. R. (2003b). Mhcpred: bringing a quantitative dimension to the online prediction of mhc binding. *Appl Bioinformatics*, **2**(1), 63–66. [citation in 10.1]

Guan, P., Doytchinova, I. A., Walshe, V. A., Borrow, P., and Flower, D. R. (2005). Analysis of peptide-protein binding using amino acid descriptors: prediction and experimental verification for human histocompatibility complex hla-a0201. *J Med Chem*, **48**(23), 7418–7425. [citations in 1.1 and 11]

Guan, P., Hattotuwagama, C. K., Doytchinova, I. A., and Flower, D. R. (2006). Mhcpred 2.0: an updated quantitative t-cell epitope prediction server. *Appl Bioinformatics*, **5**(1), 55–61. [citation in 10.1]

Hansch, C. and Fujita, T. (1964). Analysis. a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, **86**, 16161626. [citation in 2.2]

Hansch, C. A. (1969). Quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.*, **2**, 232–239. [citation in 2.2]

Harary, F. (1971). Graph theory. In *Addison Wesley*. [citation in 3]

Hattotuwagama, C. K., Guan, P., Doytchinova, I. A., and Flower, D. R. (2004a). New horizons in mouse immunoinformatics: reliable in silico prediction of mouse class i histocompatibility major complex peptide binding affinity. *Org Biomol Chem*, **2**(22), 3274–3283. [citation in 10.1]

Hattotuwagama, C. K., Guan, P., Doytchinova, I. A., Zygouri, C., and Flower, D. R. (2004b). Quantitative online prediction of peptide binding to the major histocompatibility complex. *J Mol Graph Model*, **22**(3), 195–207. [citation in 10.1]

Hattotuwagama, C. K., Toseland, C. P., Guan, P., Taylor, D. J., Hemsley, S. L., Doytchinova, I. A., and Flower, D. R. (2006). Toward prediction of class ii mouse major histocompatibility complex peptide binding affinity: in silico bioinformatic evaluation using partial least squares, a robust multivariate statistical technique. *J Chem Inf Model*, **46**(3), 1491–1502. [citation in 10.1]

He, M. M., Wood, Z. A., Baase, W. A., Xiao, H., and Matthews, B. W. (2004). Alanine-scanning mutagenesis of the beta-sheet region of phage t4 lysozyme suggests that tertiary context has a dominant effect on betasheet formation. *Protein Sci*, **13**, 2716–2724. [citation in 3.5.3]

Heinz, D. W., Baase, W. A., Zhang, X. J., Blaber, M., Dahlquist, F. W., and Matthews, B. W. (1994). Accommodation of amino acid insertions in an alpha-helix of t4 lysozyme. structural and thermodynamic analysis. *Journal of Molecular Biology*, **236**, 869–886. [citation in 3.5.3]

Hellberg, S., Sjstrm, M., and Wold, S. (1986). The prediction of bradykinin potentiating potency of pentapeptides. an example of a peptide quantitative structure-activity relationship. *Acta Chem Scand B*, **40**(2), 135–140. [citation in 3.2]

Hellberg, S., Sjstrm, M., Skagerberg, B., and Wold, S. (1987). Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem*, **30**(7), 1126–1135. [citations in 1.1, 3.2, 3.2, and 11]

Hellberg, S., Eriksson, L., Jonsson, J., Lindgren, F., Sjstrm, M., Skagerberg, B., Wold, S., and Andrews, P. (1991). Minimum analogue peptide sets (maps) for quantitative structure-activity relationships. *Int J Pept Protein Res*, **37**(5), 414–424. [citation in 3.2]

Hettiarachchy, N. S. and Ziegler, G. R., editors (1994). *Protein Functionality in Food Systems*. Marcel Dekker, New York. [citation in 4.2]

Holm, L. and Sander, C. (1996). The fssp database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Research*, **24**, 206–209. [citation in 3.5.3]

Hooper, N. M. (1994). Families of zinc metalloproteases. *FEBS Lett*, **354**(1), 1–6. [citation in 8.3]

Jennrich, R. J. (1977). *Stepwise discriminant analysis*. Wiley, NewYork (USA). [citation in 5.4.2]

Jones, D. (1975). Amino acid properties and side-chain orientation in proteins: A cross correlation approach. *J. Theor. Biol.*, **50**, 167–183. [citations in 3, 3, 3, 3, and 11.1]

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, **8**(3), 275–282. [citations in 2, 3, 2, 3, and 11.2]

Jonsson, J., Eriksson, L., Hellberg, S., Sjstrm, M., and Wold, S. (1989). Multivariate parametrization of 55 coded and noncoded amino acids. *Quant. Struct.-Act. Relat.*, **8**, 204–209. [citations in 3.2 and 3.2]

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637. [citation in 3.5.3]

Kawashima, S. and Kanehisa, M. (2000). Aaindex: amino acid index database. *Nucleic Acids Res*, **28**(1), 374. [citations in 3.4.2 and 10.1]

Kawashima, S., Ogata, H., and Kanehisa, M. (1999). Aaindex: Amino acid index database. *Nucleic Acids Res*, **27**(1), 368–369. [citations in 3.4.2 and 10.1]

Kewley, R. J., Whitelaw, M. L., and Chapman-Smith, A. (2004). The mammalian basic helix-loop-helix/pas family of transcriptional regulators. *Int J Biochem Cell Biol*, **36**(2), 189–204. [citation in 8.3]

Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occuring amino acids. *J. Protein Chem.*, **4**, 23–55. [citation in 3.2]

Kivirikko, K. and Pihlajaniemi, T. (1998). Collagen hydroxylases and the protein disulfide isomerase subunit of prolyl 4-hydroxylases. *Adv Enzymol Relat Areas Mol Biol*, **72**, 325–398. [citation in 3.5.4]

Kleiner, D. E. and Stetler-Stevenson, W. G. (1999). Matrix metalloproteinases and metastasis. *Cancer Chemother Pharmacol*, **43 Suppl**, S42–S51. [citation in 8.3]

Klir, G. J. and Folger, T. A. (1988). *Fuzzy Sets, Uncertainty and Information*. Prentice-Hall, Englewood Cliffs (NJ). [citation in 4.4.1]

Krzycki, J. (2005). The direct genetic encoding of pyrrolysine. *Curr Opin Microbiol*, **6**, 706–712. [citation in 3.5.4]

Kuhn, L. A., Swanson, C. A., Pique, M. E., Tainer, J. A., and Getzoff, E. D. (1995). Atomic and residue hydrophilicity in the context of folded protein structures. *J Proteins*, **23**, 536–547. [citations in 5, 5, 5, 5, and 11.1]

Kvalheim, O. M. (1987). Latent-structure decompositions (projections) of multivariate data. *Chemometrics and Intelligent Laboratory Systems*, **2**, 283–290. [citation in 5.2]

Leardi, R. (1994). Application of a genetic algorithm to feature selection under full validationconditions and to outlier detection. *Journal of Chemometrics*, **8**, 65–79. [citations in 5.4.3 and 9.3]

Leardi, R. (2001). Genetic algorithms in chemometrics and chemistry: a review. *Journal of Chemometrics*, **15**, 559–569. [citations in 5.4.3 and 9.3]

Leardi, R., Boggia, R., and Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, **6**, 267–281. [citations in 5.4.3 and 9.3]

Lindgren, F., Hansen, B., Karcher, W., Sjstrm, M., and Eriksson, L. (1996). Model validation by permutation tests: applications to variable selection. *J. Chemom.*, **10**, 521532. [citations in 5.5.3 and 9.3.1]

Liu, J., Tan, H., and Rost, B. (2002). Loopy proteins appear conserved in evolution. *J Mol Biol*, **322**, 53–64. [citation in 3.5.3]

LoConte, L., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2002). Scop database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.*, **30**, 264–267. [citations in 2, 3.5.3, 6, 8.1, 8.2, and 8.3]

Lohmander, L. S., Hoerrner, L. A., and Lark, M. W. (1993). Metalloproteinases, tissue inhibitor, and proteoglycan fragments in knee synovial fluid in human osteoarthritis. *Arthritis Rheum*, **36**(2), 181–189. [citation in 8.3]

MacRitchie, F. (1992). Physicochemical properties of wheat proteins in relation to functionality. *AdV. Food Nutr. Res.*, **36**, 1–87. [citation in 4.2]

Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., de Jong, S., Lewi, P. J., and Smeyers Verbeke, J. (1997). *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier, Amsterdam. [citation in 2.1]

Mauri, A., Consonni, V., Pavan, M., and Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. **56**, 237–248. [citation in 1.2]

Miyamoto, S. and Kollman, P. A. (1993). Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches. *Proteins*, **16**(3), 226–245. [citation in 7.2]

Mizuguchi, K. and Blundell, T. (2000). Analysis of conservation and substitutions of secondary structure elements within protein superfamilies. *Bioinformatics*, **16**, 1111–1119. [citation in 3.5.3]

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23. [citation in 4.3.1]

Murphy, G. and Hembry, R. M. (1992). Proteinases in rheumatoid arthritis. *J Rheumatol Suppl*, **32**, 61–64. [citation in 8.3]

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536540. [citations in 2, 3.5.3, 6, 8.1, 8.2, and 8.3]

Nakai, S., Li Chan, E., and Hayakawa, S. (1986). Contribution of protein hydrophobicity to its functionality. *Nahrung*, **30**, 327–336. [citations in 3.4.2, 4.2, and 10.1]

Nelson, D. L. and Cox, M. M. (2005). *Lehninger's Principles of Biochemistry*. W. H. Freeman and Company, New York. [citation in 2.3]

Nystrm, ., Andersson, P., and Lundstedt, T. (2000). Multivariate data analysis of topographically modified alpha-melanotropin analogues using auto and cross auto covariances (acc). *Quant Struct-Act Relat.*, **19**, 264–269. [citations in 1.1 and 11]

Pandini, A. and Bonati, L. (2005). Conservation and specialization in pas domain dynamics. *Protein Eng Des Sel*, **18**(3), 127–137. [citation in 8.3]

Peress, N., Perillo, E., and Zucker, S. (1995). Localization of tissue inhibitor of matrix metalloproteinases in alzheimer's disease and normal brain. *J Neuropathol Exp Neurol*, **54**(1), 16–22. [citation in 8.3]

Phillips, L. G., Whitehead, D. M., and Kinsella, J. E. (1994). *Structure-Function Properties of Food Proteins*. Academic Press, New York. [citation in 4.2]

Pomeranz, Y. (1991). *Functional Properties of Food Components*. Academic Press, New York. [citation in 4.2]

Prabhakaran, M. (1990). The distribution of physical, chemical and conformational properties in signal and nascent peptides. *Biochem J*, **269**(3), 691–696. [citations in 1, 4, 5, 1, 4, 5, and 11.2]

Repik, A., Rebbapragada, A., Johnson, M. S., Haznedar, J. O., Zhulin, I. B., and Taylor, B. L. (2000). Pas domain residues involved in signal transduction by the aer redox sensor of escherichia coli. *Mol Microbiol*, **36**(4), 806–816. [citation in 8.3]

Rost, B. (1997). Protein structures sustain evolutionary drift. *Fold Des 1997*, **2**, 19–24. [citation in 3.5.3]

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng*, **12**, 85–94. [citation in 3.5.3]

Sandberg, M., Eriksson, L., Jonsson, J., Sjstrm, M., and Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids. *J Med Chem*, **41**(14), 2481–2491. [citations in 1.1, 3.2, and 11]

Siebert, K. J. (2001). Quantitative structure-activity relationship modeling of peptide and protein behavior as a function of amino acid composition. *J Agric Food Chem*, **49**(2), 851–858. [citations in 3.2 and 4.2]

Siebert, K. J. (2003). Modeling protein functional properties from amino acid composition. *J. Agric. Food Chem. 2003, 51, 7792-7797*, **51**, 7792–7797. [citations in 3.2 and 4.2]

Sjstrm, M., Rnnar, S., and Wieslander, . (1995). Polypeptide sequence property relationships in escherichia coli based on auto cross covariances. *Chemometr. Intell. Lab Syst.*, **29**, 295–305. [citations in 1.1 and 11]

Skiles, J. W., Gonnella, N. C., and Jeng, A. Y. (2001). The design, structure, and therapeutic application of matrix metalloproteinase inhibitors. *Curr Med Chem*, **8**(4), 425–474. [citation in 8.3]

Skiles, J. W., Gonnella, N. C., and Jeng, A. Y. (2004). The design, structure, and clinical update of small molecular weight matrix metalloproteinase inhibitors. *Curr Med Chem*, **11**(22), 2911–2977. [citation in 8.3]

Taylor, B. L. and Zhulin, I. B. (1999). Pas domains: internal sensors of oxygen, redox potential, and light. *Microbiol Mol Biol Rev*, **63**(2), 479–506. [citation in 8.3]

Todeschini, R. and Consonni, V. (2000). *Handbook of Molecular Descriptors*. Wiley - VCH. [citations in 4, 4.1, and 4.3]

Todeschini, R. and Gramatica, P. (1997a). 3d-modelling and prediction by whim descriptors. part 5. theory development and chemical meaning of whim descriptors. *Quantitative Structure-Activity Relationships*, **16**, 113–119. [citations in 3.5.1, 3.5.4, and 4.4.1]

Todeschini, R. and Gramatica, P. (1997b). 3d-modelling and prediction by whim descriptors. part 6. application of whim descriptors in qsar studies. *Quantitative Structure-Activity Relationships*, **16**, 120–125. [citations in 3.5.1, 3.5.4, and 4.4.1]

Todeschini, R. and Gramatica, P. (1997c). The whim theory: New 3d molecular descriptors for qsar in environmental modelling. *SAR QSAR Environ. Res.*, **7**, 89–115. [citations in 3.5.1, 3.5.4, and 4.4.1]

Todeschini, R. and Gramatica, P. (1998). *3D QSAR in Drug Design*, volume 2, pages 355–380. Kluwer/ESCOM, Dordrecht (The Netherlands). [citations in 3.5.4, 4.4, and 4.4.1]

Todeschini, R., Lasagni, M., and Marengo, E. (1994). New molecular descriptors for 2d-. and 3d-structures, theory. *J.Chemom.*, **8**, 263–273. [citations in 3.2, 3.5.1, 3.5.4, and 4.4.1]

Todeschini, R., Gramatica, P., and Provenzani, R. (1995). Weighted holistic invariant molecular descriptors.part 2.theory development and applications on modeling physicochemical properties of polycyclic aromatic hydrocarbons. *Chemom Intell Lab Syst*, **27**, 221–229. [citations in 3.5.1, 3.5.4, and 4.4.1]

Todeschini, R., Bettiol, C., Giurin, G., Gramatica, P., Miana, P., and Argese, E. (1996a). Modeling and prediction by using whim descriptors in qsar studies: submitochondrial particles (smp) as toxicity biosensors of chlorophenols. *Chemosphere*, **33**, 71–79. [citations in 3.5.1, 3.5.4, and 4.4.1]

Todeschini, R., Vighi, M., Provenzani, R., Finizio, A., and Gramatica, P. (1996b). Modeling and prediction by using whim descriptors in qsar studies: toxicity of heterogeneous chemicals on daphnia magna. *Chemosphere*, **32**, 1527–1545. [citations in 3.5.1, 3.5.4, and 4.4.1]

Todeschini, R., Vighi, M., Finizio, A., and Gramatica, P. (1997). 3d-modelling and prediction by whim descriptors. part 8. toxicity and physico-chemical properties of environmental priority chemicals by 2d-ti and 3d- whim descriptors. *SAR QSAR Environ. Res.*, **7**, 173–193. [citations in 3.5.1, 3.5.4, and 4.4.1]

Todeschini, R., Consonni, V., Mauri, A., and Pavan, M. (2003). *MobyDigs: software for regression and classification models by genetic algorithms*, volume Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial

Neural Networks of *Data Handling in Science and Technology*, chapter 5, pages 141–167. Elsevier. [citations in 5.4.3, 5.4.3, and 9.3]

Tomii, K. and Kanehisa, M. (1996). Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng*, **9**(1), 27–36. [citations in 3.4.2 and 10.1]

Toseland, C. P., Clayton, D. J., McSparron, H., Hemsley, S. L., Blythe, M. J., Paine, K., Doytchinova, I. A., Guan, P., Hattotuwagama, C. K., and Flower, D. R. (2005). Antijen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res*, **1**(1), 4. [citation in 10.1]

Wilchek, M. and Bayer, E. (1989). *Protein Recognition of Immobilized Ligands.*, pages 83–90. Alan R. Liss, Inc. [citation in 7.2]

Wold, S. (1972). S. spline functions, a new tool in data-analysis. *Kemisk Tidskrift*, **84**, 34–37. [citation in 2.1]

Wold, S. (1990). Chemometrics; what do we mean with it, and what do we want from it? *Chemometric and Intelligent Laboratory Systems*, **35**, 109–115. [citation in 2.1]

Wold, S., Esbensen, K. H., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **2**, 37–52. [citation in 5.2]

Wolfenden, R., Andersson, L., Cullis, P. M., and Southgate, C. C. B. (1981). Affinities of amino acid side chains for solvent water. *Biochemistry*, **20**, 849–855. [citation in 3.2]

Zaliani, A. and E., G. (1999). Ms-whim scores for amino acids: A new 3d-description for peptide qsar and qspr studies. *J. Chem. Inf. Comput. Sci.*, **39**, 525–533. [citation in 3.2]

Zimmermann, R. and Cox, E. (1994). Dna stretching on functionalized gold surfaces. *Nucleic Acids Research*, **22**, 492–497. [citation in 7.2]

part III

Appendix

List of Figures

List of Tables

List of molecular descriptors

Introduction

In this chapter the labels and the definitions of all the descriptors used in this PhD thesis are listed. Labels and definitions are reported in a general way, weighted descriptors are identified by the suffix wi. This suffix is replaced in the PhD thesis by the proper identification suffix of every weight. Before the list of the descriptors the list of suffix for every amnino acid weight is reported.

For example if the considered property used to characterise the amino acids is the residue accessible surface area in folded protein by Chothia [Chothia (1976)] the Broto-Moreau autocorrelation of a topological structure (lag 1), that in general way is $ATS1wi$, will become $ATS1ras$ and its definition will be "Broto-Moreau autocorrelation of a topological structure - lag 1 / Weighted by residue accessible surface area in folded protein (Chothia, 1976)".

Table 11.1: Suffixes and descriptions of the physicochemical weights.

Suffix	Description
mw	molecular weight [Fasman (1976)]
p	polarity [Grantham (1974)]
hyb	hydrophobicity [Jones (1975)]
ras	residue accessible surface area in folded protein [Chothia (1976)]
hyl	hydrophilicity scale [Kuhn *et al.* (1995)]

Table 11.2: Suffixes and descriptions of the statistical weights.

Suffix	**Description**
rf_bs	relative frequency in beta-sheet [Prabhakaran (1990)]
rfo	relative frequency of occurrence [Jones *et al.* (1992)]
rm	relative mutability [Jones *et al.* (1992)]
rf_ah	relative frequency in alpha-helix [Prabhakaran (1990)]
rf_rt	relative frequency in reverse-turn [Prabhakaran (1990)]

Table 11.3: Suffixes and descriptions of the WHIM weights.

Suffix	**Description**
Am	WHIM global dimension index weighted by atomic masses
Km	WHIM global shape index weighted by atomic masses
Dm	WHIM global density index weighted by atomic masses

Table 11.4: List of constitutional molecular descriptors

ID	Symbol	Description
1	nAAs	number of AAs
2	Wwi_sum	sum of weight wi
3	Wwi_asum	average sum of weight wi
4	nAla	number of Alanines
5	nArg	number of Arginines
6	nAsn	number of Asparagines
7	nAsp	number of Aspartic acids
8	nCys	number of Cysteines
9	nGln	number of Glutamic acids
10	nGlu	number of Glutamines
11	nGly	number of Glycines
12	nHis	number of Histidines
13	nIle	number of Isoleucines
14	nLeu	number of Leucines
15	nLys	number of Lysines
16	nMet	number of Methionines
17	nPhe	number of Phenylalanines
18	nPro	number of Prolines

Table 11.4: List of constitutional molecular descriptors

ID	Symbol	Description
19	nSer	number of Serines
20	nThr	number of Threonines
21	nTrp	number of Tryptophans
22	nTyr	number of Tyrosines
23	nVal	number of Valines
24	nAla / nAAs	number of Alanines / number of AAs
25	nArg / nAAs	number of Arginines / number of AAs
26	nAsn / nAAs	number of Asparagines / number of AAs
27	nAsp / nAAs	number of Aspartic acids / number of AAs
28	nCys / nAAs	number of Cysteines / number of AAs
29	nGln / nAAs	number of Glutamic acids / number of AAs
30	nGlu / nAAs	number of Glutamines / number of AAs
31	nGly / nAAs	number of Glycines / number of AAs
32	nHis / nAAs	number of Histidines / number of AAs
33	nIle / nAAs	number of Isoleucines / number of AAs
34	nLeu / nAAs	number of Leucines / number of AAs
35	nLys / nAAs	number of Lysines / number of AAs
36	nMet / nAAs	number of Methionines / number of AAs
37	nPhe / nAAs	number of Phenylalanines / number of AAs
38	nPro / nAAs	number of Prolines / number of AAs
39	nSer / nAAs	number of Serines / number of AAs
40	nThr / nAAs	number of Threonines / number of AAs
41	nTrp / nAAs	number of Tryptophans / number of AAs
42	nTyr / nAAs	number of Tyrosines / number of AAs
43	nVal / nAAs	number of Valines / number of AAs

Table 11.5: List of autocorrelation molecular descriptors

ID	Symbol	Description
1	ATS1wi	Broto-Moreau autocorrelation of a topological structure - lag 1 / Weighted by wi
2	ATS2wi	Broto-Moreau autocorrelation of a topological structure - lag 2 / Weighted by wi
3	ATS3wi	Broto-Moreau autocorrelation of a topological structure - lag 3 / Weighted by wi
4	ATS4wi	Broto-Moreau autocorrelation of a topological structure - lag 4 / Weighted by wi
5	ATS5wi	Broto-Moreau autocorrelation of a topological structure - lag 5 / Weighted by wi
6	ATS6wi	Broto-Moreau autocorrelation of a topological structure - lag 6 / Weighted by wi
7	ATS7wi	Broto-Moreau autocorrelation of a topological structure - lag 7 / Weighted by wi
8	ATS8wi	Broto-Moreau autocorrelation of a topological structure - lag 8 / Weighted by wi
9	MATS1wi	Moran autocorrelation - lag 1 / Weighted by wi
10	MATS2wi	Moran autocorrelation - lag 2 / Weighted by wi
11	MATS3wi	Moran autocorrelation - lag 3 / Weighted by wi
12	MATS4wi	Moran autocorrelation - lag 4 / Weighted by wi
13	MATS5wi	Moran autocorrelation - lag 5 / Weighted by wi
14	MATS6wi	Moran autocorrelation - lag 6 / Weighted by wi
15	MATS7wi	Moran autocorrelation - lag 7 / Weighted by wi
16	MATS8wi	Moran autocorrelation - lag 8 / Weighted by wi
17	GATS1wi	Geary autocorrelation - lag 1 / Weighted by wi
18	GATS2wi	Geary autocorrelation - lag 2 / Weighted by wi
19	GATS3wi	Geary autocorrelation - lag 3 / Weighted by wi
20	GATS4wi	Geary autocorrelation - lag 4 / Weighted by wi
21	GATS5wi	Geary autocorrelation - lag 5 / Weighted by wi
22	GATS6wi	Geary autocorrelation - lag 6 / Weighted by wi
23	GATS7wi	Geary autocorrelation - lag 7 / Weighted by wi
24	GATS8wi	Geary autocorrelation - lag 8 / Weighted by wi

Table 11.6: List of WHIM molecular descriptors

ID	Symbol	Description
1	L1wi	1st component size directional WHIM index / weighted by wi
2	L2wi	2nd component size directional WHIM index / Weighted by wi
3	L3wi	3rd component size directional WHIM index / Weighted by wi
4	P1wi	1st component shape directional WHIM index / Weighted by wi
5	P2wi	2nd component shape directional WHIM index / Weighted by wi
6	G1wi	1st component symmetry directional WHIM index / Weighted by wi
7	G2wi	2st component symmetry directional WHIM index / Weighted by wi
8	G3wi	3st component symmetry directional WHIM index / Weighted by wi
9	E1wi	1st component accessibility directional WHIM index / Weighted by wi
10	E2wi	2nd component accessibility directional WHIM index / Weighted by wi
11	E3wi	3rd component accessibility directional WHIM index / Weighted by wi
12	Twi	T total size index / weighted by wi
13	Awi	A total size index / Weighted by wi
14	Gwi	G total symmetry index / Weighted by wi
15	Kwi	K global shape index / Weighted by wi
16	Dwi	D total accessibility index / Weighted by wi
17	Vwi	V total size index / Weighted by wi

Table 11.7: List of GETAWAY molecular descriptors

ID	Symbol	Description
1	ITH	total information content on the leverage equality
2	ISH	standardized information content on the leverage equality
3	HIC	mean information content on the leverage magnitude
4	HGM	geometric mean on the leverage magnitude
5	H0wi	H autocorrelation of lag 0 / Weighted by wi
6	H1wi	H autocorrelation of lag 1 / Weighted by wi
7	H2wi	H autocorrelation of lag 2 / Weighted by wi
8	H3wi	H autocorrelation of lag 3 / Weighted by wi
9	H4wi	H autocorrelation of lag 4 / Weighted by wi
10	H5wi	H autocorrelation of lag 5 / Weighted by wi
11	H6wi	H autocorrelation of lag 6 / Weighted by wi
12	H7wi	H autocorrelation of lag 7 / Weighted by wi
13	H8wi	H autocorrelation of lag 8 / Weighted by wi
14	HTwi	H total index / Weighted by wi
15	HATS0wi	leverage-weighted autocorrelation of lag 0 / Weighted by wi
16	HATS1wi	leverage-weighted autocorrelation of lag 1 / Weighted by wi
17	HATS2wi	leverage-weighted autocorrelation of lag 2 / Weighted by wi
18	HATS3wi	leverage-weighted autocorrelation of lag 3 / Weighted by wi
19	HATS4wi	leverage-weighted autocorrelation of lag 4 / Weighted by wi
20	HATS5wi	leverage-weighted autocorrelation of lag 5 / Weighted by wi
21	HATS6wi	leverage-weighted autocorrelation of lag 6 / Weighted by wi
22	HATS7wi	leverage-weighted autocorrelation of lag 7 / Weighted by wi
23	HATS8wi	leverage-weighted autocorrelation of lag 8 / Weighted by wi

Table 11.7: List of GETAWAY molecular descriptors

ID	Symbol	Description
24	HATSwi	leverage-weighted total index / Weighted by wi
25	RCON	Randic-type R matrix connectivity
26	RARS	R matrix average row sum
27	REIG	first eigenvalue of the R matrix
28	R1wi	R autocorrelation of lag 1 / Weighted by wi
29	R2wi	R autocorrelation of lag 2 / Weighted by wi
30	R3wi	R autocorrelation of lag 3 / Weighted by wi
31	R4wi	R autocorrelation of lag 4 / Weighted by wi
32	R5wi	R autocorrelation of lag 5 / Weighted by wi
33	R6wi	R autocorrelation of lag 6 / Weighted by wi
34	R7wi	R autocorrelation of lag 7 / Weighted by wi
35	R8wi	R autocorrelation of lag 8 / Weighted by wi
36	RTwi	R total index / Weighted by wi
37	R1wi+	R maximal autocorrelation of lag 1 / Weighted by wi
38	R2wi+	R maximal autocorrelation of lag 2 / Weighted by wi
39	R3wi+	R maximal autocorrelation of lag 3 / Weighted by wi
40	R4wi+	R maximal autocorrelation of lag 4 / Weighted by wi
41	R5wi+	R maximal autocorrelation of lag 5 / Weighted by wi
42	R6wi+	R maximal autocorrelation of lag 6 / Weighted by wi
43	R7wi+	R maximal autocorrelation of lag 7 / Weighted by wi
44	R8wi+	R maximal autocorrelation of lag 8 / Weighted by wi
45	RTwi+	R maximal index / Weighted by wi

List of publications

2007

Todeschini R., Ballabio D., Consonni V., Mauri A., Pavan M. (2007). CAIMAN (Classification And Influence Matrix Analysis): a new approach to the classification based on leverage-scaled functions. *Chemometrics and Intelligent Laboratory Systems*, **87**, 3-17

Abstract A new classification method is proposed based on the influence matrix (or leverage matrix). The use of the influence matrix is well known in regression analysis, where the diagonal matrix elements (i.e. the leverages) give information on the role of each sample within the regression model. In effect, the leverages are related to the distance of the sample from the hyperellipsoids defining the model space, to the degree of membership to the model, to the reliability of model predictions. Exploiting the leverage properties, the Classification And Influence Matrix Analysis method (CAIMAN) models each class by means of the class dispersion matrix and calculates the leverage of each sample with respect to each class model space. Unlike other classification methods such as LDA, QDA, and UNEQ, CAIMAN does not require multinormality assumptions. It is developed in three different options: (1) D-CAIMAN, which is a discriminant classification method, (2) M-CAIMAN, which is a class modelling method allowing an object to be classified, not classified at all, or assigned to more than one class, (3) A-CAIMAN, which deals with the asymmetric case, i.e. only a reference class needs to be modelled. Performance of the proposed method has been evaluated by means of several classification data sets taken from literature and compared with the most popular classification methods. Final results seem to indicate that CAIMAN performs well and, in most of the analysed cases, better than the other classification methods.

Todeschini R., Ballabio D., Consonni V., Mauri A. (2007). A new similarity/diversity measure for sequential data. *MATCH Communications in Mathematical and in Computer Chemistry*, **57**, 51-67

Abstract The concept of similarity and its dual concept of diversity play a fundamental role in several QSAR strategies, chemometrics and library searching methods, virtual screening, as well as in relatively new fields such as genomics and proteomics. In this paper, a new flexible similarity/diversity measure is proposed to deal with sequential data, both taking into account the differences in property values of the sequence elements and the ordering relationships among the sequence elements themselves. Data such as DNA sequences, mass and NMR spectra, sequential molecular descriptors are all characterized by an ordering variable (the sequence) and by a property of the sequence elements. Some examples on artificial DNA sequences, mass spectra, molecular descriptors and proteomic maps are given.

2006

Mauri A., Consonni V., Pavan M., Todeschini R. (2006) DRAGON software: an easy approach to molecular descriptor calculations *MATCH Communications in Mathematical and in Computer Chemistry* **56**, 237-248

Abstract Due to the relevance that molecular descriptors gained in several scientific fields, software for the calculation of molecular descriptors has became very important tools for the scientists. In this paper, the main characteristics of DRAGON software for the calculation of molecular descriptors are illustrated.

Ballabio D., Mauri A.,Todeschini R., Buratti S. (2006). Geographical classification of wine and olive oil by means of CAIMAN (Classification And Influence Matrix Analysis). *Analytica Chimica Acta*, **570**, 249-258

Abstract Classification and influence matrix analysis (CAIMAN) is a new classification method, recently proposed and based on the influence matrix (also called leverage matrix). Depending on the purposes of the classification analysis, CAIMAN can be used in three outlines: (1) D-CAIMAN is a discriminant classification method, (2) M-CAIMAN is a class modelling method allowing a sample to be classified, not classified at all, or assigned to more than one class (confused) and (3) A-CAIMAN deals with the asymmetric case, where only a reference class needs to be modelled. In this work, the geographic classification of samples of wine and olive oil has been carried out by means of CAIMAN and its results compared with discriminant analysis, by focusing great attention on the model predictive capabilities. The geographic characterization has been carried out on three different datasets: extra virgin olive oils produced in a small area, with a protected denomination of origin label, wines with different denominations of origin, but produced in enclosed geographical areas, and olive oils belonging to different production areas. Final results seem to indicate that the application of CAIMAN to the geographical origin identification offers several advantages: first, it shows on an average basis good performances; second, it is able to deal in a simple way classification problems related to tipicity, authenticity, and uniqueness characterization, which are of increasing interest in food quality issues.

Todeschini R., Consonni V., Mauri A., Ballabio D. (2006). Characterization of DNA Primary Sequences by a New Similarity/Diversity Measure Based on the Partial Ordering *Journal of Chemical Information and Modeling*, **46**, 1905-1911

Abstract The similarity/diversity measures play a fundamental role in library searching, virtual screening, and quantitative structure-activity relationship / quantitative structure-property relationship modeling as well as in genomics and proteomics. In this paper, a new similarity/diversity measure is proposed as a new approach for the analysis of sequential data, where useful information can be also obtained by the ordering relationships between the sequence elements. This methodology can be applied for evaluating molecular similarity/diversity, using sets of sequential descriptors, and for evaluating the similarity between spectra, sensor arrays, and other sequential data such as DNA and protein sequences. The new proposed distance (weighted standardized Hasse distance) is evaluated between pairs of Hasse matrices derived from the classical partial-ordering rules. It can be naturally standardized, thus allowing the interpretation of these distances as absolute values (e.g., percentage) and deriving simple similarity and correlation indices. A simple example is taken to highlight the behavior of the new similarity/diversity measure on DNA sequences taken from the first exons of the beta-globins for eight different species. Sensitivity analysis has been also performed, showing the high capability of this measure to take into account small modifications of the DNA sequences. Finally, a comparison with results obtained from the literature is given, together with a comparison with matrix invariants derived from the Hasse matrix.

2005

Tetko I.V., Gasteiger J., Todeschini R., Mauri A., Livingstone D., Ertl P., Palyulin V.A., Radchenko E.V., Zefirov N.S., Makarenko A.S., Tanchuk V.Y., Prokopenko V.V. (2005) Virtual computational chemistry laboratory–design and description. *Journal of Computer-Aided Molecular Design*, **19**, 453-463

Abstract Internet technology offers an excellent opportunity for the development of tools by the cooperative effort of various groups and institutions. We have developed a multi-platform software system, Virtual Computational Chemistry Laboratory, http://www.vcclab.org, allowing the computational chemist to perform a comprehensive series of molecular indices/properties calculations and data analysis. The implemented software is based on a three-tier architecture that is one of the standard technologies to provide client-server services on the Internet. The developed software includes several popular programs, including the indices generation program, DRAGON, a 3D structure generator, CORINA, a program to predict lipophilicity and aqueous solubility of chemicals, ALOGPS and others. All these programs are running at the host institutes located in five countries over Europe. In this article we review the main features and statistics of the developed system that can be used as a prototype for academic and industry models.