MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **60** (2008) 671-690

ISSN 0340 - 6253

Peptides multivariate characterisation using a molecular descriptor based approach

A. Mauri*, D. Ballabio, V.Consonni, A. Manganaro and R. Todeschini

Milano Chemometrics and QSAR Research Group, Dept. of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1 – 20126 Milano (Italy)
Web site: http://michem.disat.unimib.it/chm/

*Corresponding author. Tel +39-0264482801; Fax +39-0264482839;
e-mail address: andrea.mauri@unimib.it

(Received April 18, 2008)

Abstract

Peptide sequences with different lengths, available from synthesised peptide libraries and sequenced proteins, are potentially valuable for evaluating structure-activity relationships. However, in order to apply multivariate regression and classification models on such sequences, it is necessary to have a preprocessing method that translates them into a uniform set of variables.

A molecular descriptor based approach can be suitable for the characterisation of peptide sequences and the prediction of their chemical or biological properties. In this paper a novel methodology based on traditional molecular descriptors calculated on a simplified representation of peptides and proteins has been evaluated. 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 successfully applied on a peptide data set taken from the literature.

Introduction

The most used approach for peptide characterisation is based on the modelling of biological properties of small peptides as a function of amino acid principal properties. This approach was introduced by Kidera et al. [1] who 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. [2-5] who developed principal properties for each of 20 natural amino acids and for a series of non-standard ones. These properties were derived by carrying out Principal Components Analysis (PCA) of numerous amino acid properties, such as HPLC retention times, p*K*as, 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 scores were interpreted mainly to represent 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 build and evaluate QSAR models. With the three z-scales it was possible to numerically quantify the structural variation within a series of related peptides, by arranging the z-scales according to the amino acid sequence. This approach produced good models for small peptides but has the disadvantage for those larger than a few amino acids. As a result, in this second case, the number of peptides needed to construct a reliable model has to be large [6,7].

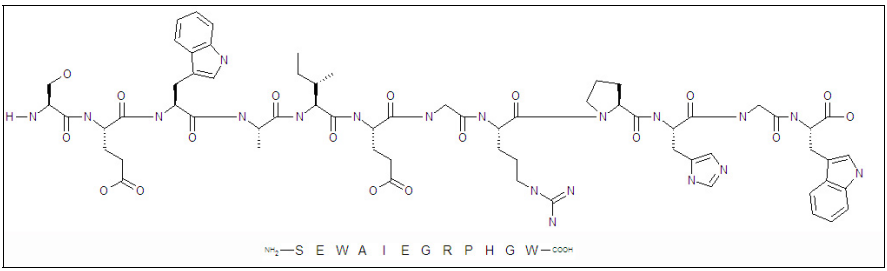

Figure 1.Graphical representation of the same peptide sequence (SEWAIEGRPHGW) using an atom based (upper) and an amino acid based (lower) representation

In this paper a novel methodology based on traditional molecular descriptors calculated on a simplified representation of peptides and proteins is presented. This 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,

similarly the descriptor-based representation is a simplified and 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.

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 peptide representation is indispensable.

Considering a peptide as a topological molecular graph, the more immediate way to represent a peptide is an atom based representation. Using this representation all the atoms belonging to a peptide are considered. The atom-based representation raises one big problem related to the fact that complex descriptors cannot be calculated on structures constituted by thousands of atoms. Another issue is that not necessarily all the information brought by the atom-based representation is directly connected to peptides' properties.

In this paper, an amino acid-based representation has been studied. A topological representation of a peptide using an atom-based approach is a complex molecular graph where atoms are connected to the others by the molecular bonds, while the same peptide using an amino acid-based approach is just a sequence of amino acid types. A graphical comparison of the atom based and amino acid based representations is shown in Figure 1. By means of the amino acid based representation, the peptide description is simplified, considering that a) the physicochemical properties of the amino acids are responsible for the 3D structure and the functionality of the peptide and b) all amino acids share common structural features, including an alpha-carbon to which an amino group, a carboxyl group, and a variable side chain are bonded. The amino acid based representation permits to reduce the complexity of the structures, since the number of amino acids in a peptide is significantly lower than the number of atoms.

In the first part of the paper, the new approach for the calculation and the weighting scheme of molecular descriptors on peptide sequences is presented. Then, the results obtained on a dataset taken from the literature are shown and commented with respect to the regression models for the prediction of two biological responses.

Materials and methods

Molecular Descriptors

Molecular descriptors are formally mathematical representations of a molecule obtained by a well-specified algorithm applied to a defined molecular representation enabling mathematical treatment [8].

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.

For this study amino acids constitutional and 2D autocorrelation descriptors have been used. All the calculations were performed by an ongoing implementation of DRAGON software [9,10], where an extension of its capability in order to calculate descriptors for proteins and peptides has been on purpose added during this study.

Constitutional descriptors are the most simple and commonly used descriptors [8], reflecting the molecular composition of a compound without any information about its molecular geometry; the proposed constitutional descriptors are listed in Table 1.

2D autocorrelation descriptors [11-15] are molecular descriptors which describe how a considered property is distributed along a topological molecular structure; the proposed 2D autocorrelation descriptors are listed in Table 2.

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.

Weighting scheme

The amino acids are the building blocks of proteins and peptides each having different characteristics in terms of shape, volume, and chemical reactivity.

Molecular descriptors can be calculated in an unweighted way, i.e. considering every amino acid equal to the others, or weighting every amino acid by a descriptive property. Consequently, it has been necessary to choose the properties needed to weight 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), that includes numerical indices representing various physicochemical, biochemical and statistical properties of amino acids and pairs of

amino acids [16-19]. AAindex database has been made publicly available by the Japanese GenomeNet database service.

Symbol	Definition
nAAs	number of AAs
Wwi_sum	sum of weight wi
Wwi_asum	average sum of weight w
nAla	number of Alanines
nArg	number of Arginines
nAsn	number of Asparagines
nAsp	number of Aspartic acids
nCys	number of Cysteines
nGln	number of Glutamic acids
nGlu	number of Glutamines
nGly	number of Glycines
nHis	number of Histidines
nIle	number of Isoleucines
nLeu	number of Leucines
nLys	number of Lysines
nMet	number of Methionines
nPhe	number of Phenylalanines
nPro	number of Prolines
nSer	number of Serines
nThr	number of Threonines
nTrp	number of Tryptophans
nTyr	number of Tyrosines
nVal	number of Valines
nAla / nAAs	number of Alanines / number of AAs
nArg / nAAs	number of Arginines / number of AAs
nAsn / nAAs	number of Asparagines / number of AAs
nAsp / nAAs	number of Aspartic acids / number of AAs
nCys / nAAs	number of Cysteines / number of AAs
nGln / nAAs	number of Glutamic acids / number of AAs
nGlu / nAAs	number of Glutamines / number of AAs
nGly / nAAs	number of Glycines / number of AAs
nHis / nAAs	number of Histidines / number of AAs
nIle / nAAs	number of Isoleucines / number of AAs
nLeu / nAAs	number of Leucines / number of AAs
nLys / nAAs	number of Lysines / number of AAs
nMet / nAAs	number of Methionines / number of AAs
nPhe / nAAs	number of Phenylalanines / number of AAs
nPro / nAAs	number of Prolines / number of AAs
nSer / nAAs	number of Serines / number of AAs
nThr / nAAs	number of Threonines / number of AAs
nTrp / nAAs	number of Tryptophans / number of AAs
nTyr / nAAs	number of Tyrosines / number of AAs
nVal / nAAs	number of Valines / number of AAs

Table 1. List of constitutional molecular descriptors

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.

Symbol	Definition
ATS1wi	Broto-Moreau autocorrelation of a topological structure - lag 1 / Weighted by wi
ATS2wi	Broto-Moreau autocorrelation of a topological structure - lag 2 / Weighted by wi
ATS3wi	Broto-Moreau autocorrelation of a topological structure - lag 3 / Weighted by wi
ATS4wi	Broto-Moreau autocorrelation of a topological structure - lag 4 / Weighted by wi
ATS5wi	Broto-Moreau autocorrelation of a topological structure - lag 5 / Weighted by wi
ATS6wi	Broto-Moreau autocorrelation of a topological structure - lag 6 / Weighted by wi
ATS7wi	Broto-Moreau autocorrelation of a topological structure - lag 7 / Weighted by wi
ATS8wi	Broto-Moreau autocorrelation of a topological structure - lag 8 / Weighted by wi
MATS1wi	Moran autocorrelation - lag 1 / Weighted by wi
MATS2wi	Moran autocorrelation - lag 2 / Weighted by wi
MATS3wi	Moran autocorrelation - lag 3 / Weighted by wi
MATS4wi	Moran autocorrelation - lag 4 / Weighted by wi
MATS5wi	Moran autocorrelation - lag 5 / Weighted by wi
MATS6wi	Moran autocorrelation - lag 6 / Weighted by wi
MATS7wi	Moran autocorrelation - lag 7 / Weighted by wi
MATS8wi	Moran autocorrelation - lag 8 / Weighted by wi
GATS1wi	Geary autocorrelation - lag 1 / Weighted by wi
GATS2wi	Geary autocorrelation - lag 2 / Weighted by wi
GATS3wi	Geary autocorrelation - lag 3 / Weighted by wi
GATS4wi	Geary autocorrelation - lag 4 / Weighted by wi
GATS5wi	Geary autocorrelation - lag 5 / Weighted by wi
GATS6wi	Geary autocorrelation - lag 6 / Weighted by wi
GATS7wi	Geary autocorrelation - lag 7 / Weighted by wi
GATS8wi	Geary autocorrelation - lag 8 / Weighted by wi

Table 2. List of autocorrelation molecular descriptors. wi identifies the used weight, the suffixes for the physicochemical weights are reported in Table 3 and the WHIM suffixes are reported in Table 4

The first section (AAindex ver. 9.1) has been considered as a possible resource in order to identify relevant properties of the 20 natural amino acids, since it 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 the 20 natural 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.

The first three classes can be considered as statistical properties of the amino acids, while the fourth and the fifth classes include physicochemical properties.

In order to be able to evaluate how different properties highlight different kind of information, two weighting schemes have been defined.

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: molecular weight [20], polarity [21], hydrophobicity [22], residue accessible surface area in folded protein [23] and hydrophilicity scale [24]. The selected physicochemical weights are listed in Table 3.

Suffix	Description
mw	molecular weight by Fasman (FASG760101) [20]
p	polarity by Grantham (GRAR740102) [21]
hyb	hydrophobicity by Jones (JOND750101) [22]
ras	residue accessible surface area in folded protein by Chothia (CHOC76010) [23]
hyl	hydrophilicity scale by Kuhn (KUHL950101) [24]

Table 3. Suffixes and descriptions of the physicochemical weights, between brackets the AAindex accession number are reported

Aside from the twenty standard amino acids, there is a vast number of non standard amino acids, that are usually formed through modifications to standard amino acids.

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 introduce another weighting scheme, not depending from the amino acid index database.

The adopted weighting scheme reported in Table 4 has been obtained calculating three different global Weighted Holistic Invariant Molecular descriptors (WHIM) descriptors [25] from the molecular 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 order 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.

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 4. Suffixes and descriptions of the WHIM weights

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.

Amino Acid	3-letter	1-letter	mw	p	hyb	ras	hyl	Am	Km	Dm
Alanine	Ala	A	0.651	0.973	0.614	0.57	0.78	0.3634	0.4430	0.2330
Arginine	Arg	R	1.272	1.261	0.6	2.052	1.58	1.9266	0.7980	0.3130
Asparagine	Asn	N	0.965	1.393	0.063	1.437	1.2	0.9274	0.4970	0.2960
Aspartic acid	Asp	D	0.972	1.562	0.466	1.14	1.35	0.8575	0.4290	0.3700
Cysteine	Cys	C	0.885	0.661	1.072	0.433	0.55	0.6683	0.4990	0.2530
Glutamic acid	Glu	E	1.068	1.261	0	1.619	1.19	0.9970	0.5840	0.3810
Glutamine	Gln	Q	1.075	1.477	0.473	1.117	1.45	1.1128	0.4040	0.3260
Glycine	Gly	G	0.548	1.081	0.071	0.525	0.68	0.2343	0.5420	0.3220
Histidine	His	H	1.133	1.249	0.614	0.981	0.99	1.0631	0.7590	0.2740
Isoleucine	Ile	I	0.958	0.625	2.222	0.41	0.47	0.9845	0.5820	0.2660
Leucine	Leu	L	0.958	0.589	1.531	0.525	0.56	1.1486	0.5210	0.3370
Lysine	Lys	K	1.068	1.357	1.157	2.212	1.1	1.5369	0.8120	0.3340
Methionine	Met	M	1.09	0.685	1.178	0.707	0.66	1.0385	0.4610	0.2940
Phenylalanine	Phe	F	1.207	0.625	2.025	0.547	0.47	1.3731	0.5790	0.2710
Proline	Pro	P	0.841	0.961	1.954	1.14	0.69	0.5536	0.4870	0.2910
Serine	Ser	S	0.768	1.105	0.049	1.003	1	0.4656	0.3910	0.2810
Threonine	Thr	T	0.87	1.033	0.049	1.072	1.05	0.6918	0.4850	0.3070
Tryptophan	Trp	W	1.492	0.649	2.66	0.73	0.7	2.3415	0.6410	0.2970
Tyrosine	Tyr	Y	1.324	0.745	1.884	1.368	1	1.6385	0.6620	0.2840
Valine	Val	V	0.856	0.709	1.319	0.41	0.51	0.7066	0.5100	0.2980

Table 5. Weighting scheme values for the 20 AAs. m (Molecular Weight), p (Polarity), hyb (Hydrophobicity), ras (Residue accessible surface area in folded protein), hyl (Hydrophilicity scale), Am (WHIM global dimension descriptor), Km (WHIM global shape descriptor) and Dm (WHIM global density descriptor)

Once selected, the five physicochemical indices have been separately scaled in order to obtain values with mean equal to one. Scaled index values are showed in Table 5. Hydrophilicity has not been scaled due to the fact that this property is already scaled.

WHIM index values for all the 20 amino acids are reported in Table 5.

The two previously described weighting schemes have been used in order to separately calculate two different blocks of molecular descriptors.

Data

The twenty sequences evaluated in this application have been collected from the literature [26]; 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 two biological responses: a) activated partial thromboplastin time (APTT); b) thromboplastin time (TBPL).

Peptide	Sequence	IC_{50} (μM)			
		APTT	TBPL	Log(1+APTT)	Log(1+TBPL)
1	PKPRPDR	5.52	17.4	0.81	1.26
2	SWKHYW	0.58	2.17	0.20	0.50
3	SWKYYW	0.79	2.34	0.25	0.52
4	SWVDAW	1.56	1.26	0.41	0.35
5	RQGRYWL	1.5	6.06	0.40	0.85
6	PPGEMD	2.66	3.04	0.56	0.61
7	EGEGGM	1.58	1.2	0.41	0.34
8	RHWNIEGRPWWS	0.66	0.71	0.22	0.23
9	SEWAIEGRPHGW	1.21	0.58	0.34	0.20
10	FLRGEV	2.32	1.94	0.52	0.47
11	FMHLST	2.26	3.5	0.51	0.65
12	FMRPQM	4.14	54	0.71	1.74
13	FGWGQN	4.87	14.64	0.77	1.19
14	CWPMTRGC	1.09	0.77	0.32	0.25
15	KPRWWMWK	0.05	0.13	0.02	0.05
16	KSWQVWVK	0.8	1.1	0.26	0.32
17	KSWKYYWK	0.04	0.75	0.02	0.24
18	SWKYYWK	0.03	1.5	0.01	0.40
19	KSWKYYW	0.03	0.71	0.01	0.23
20	KMMSWKGK	0.7	0.49	0.23	0.17

Table 6. Peptide sequences and their biological activities (APTT and TBPL)

The partial thromboplastin time (PTT) or activated partial thromboplastine 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, 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. APTT and TBPL values (both original and log-transformed) for the considered peptide sequences are collected in Table 6.

Multivariate modelling

Once calculated the molecular descriptors, regression models have been built using Genetic Algorithms (GAs) [27-30] as implemented in the MobyDigs package [31,32], in order to select subsets of variables that maximise the predictive power of the multivariate models.

An important characteristic of the Genetic Algorithms is that they provide not a single model but a population of acceptable models; this characteristic enables the evaluation of variable relationships with response from different points of view. 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.

Bootstrapping is a modern, computer-intensive, general purpose approach to statistical inference, falling within a broader class of resampling methods. By bootstrap validation technique [33-35], 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. 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 **y**, i.e. by assigning to each object a randomly selected response [36,37]. Usually, the test is repeated several hundreds of times and the mean result is then considered. 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.

Results and Discussion

Constitutional and autocorrelation descriptors have been weighted using the two different weighting schemes previously described (physicochemical and WHIM weighting schemes).

The physicochemical weighting scheme collects five different properties and in this case the constitutional descriptor block consists of 51 descriptors and 2D autocorrelation descriptors consist of 120 molecular descriptors, giving a total of 171 molecular descriptors. On the other side, the WHIM weighting scheme comprises three different properties and the resulting calculated descriptors are 47 constitutional and 72 autocorrelation descriptors, giving a total of 119 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. For both biological responses the following steps have been performed, once considering the molecular descriptors weighted by the physicochemical weighting scheme and once the WHIM weighting scheme:

1. constitution of two different populations of variable, the first one collecting all the constitutional descriptors and the second one collecting all the autocorrelation descriptors;

2. selection of explained variance in validation (Q^2 leave-one-out) as fitness function for GAs;
3. application of all subset model approach for the selection of the best regression models up to two variables;
4. evolution of GAs with a maximum number of variables for each model set to three variables;
5. creation of a new population by merging the constitutional and the autocorrelation populations collecting all variables and preserving the best models;
6. increasing of the number of variables for each regression model up to four variables;
7. selection of the best five models from each population;
8. evaluation of the predictive quality of each selected model by means of bootstrap; evaluation of the stability of each selected model by means of y-scrambling analysis.

Three different model populations have been obtained for each response and for each weighting scheme: one model population collecting constitutional descriptors (one population for physicochemical and one for WHIM weighting scheme), one model population collecting autocorrelation descriptors and one model population collecting both descriptor blocks.

The best selected models are listed in Table 7 and Table 8 for molecular descriptors calculated using the physicochemical weighting scheme while the final models obtained from the molecular descriptors calculated using the WHIM weighting scheme are reported in Table 9 and Table 10.

Both bootstrap and Y-scrambling results look acceptable and indicate that the model quality is good with respect to the absence of overfitting and chance correlation, respectively. In fact, almost all Q^2_{BOOT} values are comparable with the Q^2 and R^2 values (see Table 7-10), indicating that the calculated regression models are not significantly affected by overfitting and that when applying a robust validation approach, the predictive capabilities of the selected models do not decrease.

On the other hand, the Y-scrambling results (not shown) for the reported models are all comprised in the expected limits (0.3 for R^2 and 0.05 for Q^2), showing that no chance correlation is present and highlighting the predictive abilities of the models.

Method	Size	Variables	Descriptors	R^2	Q^2	Q^2_{boot}
GAs	4	nPro nTrp nAla/nAAs nAsp/nAAs	Constitutional	88.3	83.4	81.0
GAs	4	ATS6mw ATS1hyl MATS4hyl GATS4p	ATS	88.8	81.2	76.3
GAs	4	nTrp nLys/nAAs nMet/nAAs GATS1hyb	Constitutional / ATS	89.2	80.4	75.2
GAs	3	ATS6mw ATS1hyl MATS4hyl	ATS	84.8	75.4	74.6
GAs	3	nTrp nLys/nAAs GATS1hyb	Constitutional / ATS	84.0	75.3	73.3
GAs	3	Whyb_sum nPhe nPro	Constitutional	84.7	73.9	72.2
ASM	2	nTrp nAsp/nAAs	Constitutional	73.0	65.8	65.1
ASM	2	nPro/nAAs ATS3hyb	Constitutional / ATS	72.7	60.4	56.7
ASM	2	ATS5hyb GATS5mw	ATS	65.7	55.3	56.1

Table 7. Best models obtained for APTT, using the physicochemical weighting scheme. ASM: all subset models, ATS: autocorrelation descriptor block

Method	Size	Variables	Descriptors	R^2	Q^2	Q^2_{boot}
GAs	4	nPhe nGlu/nAAs nPro/nAAs MATS7ras	Constitutional / ATS	88.2	73.5	67.1
GAs	4	nGln nTrp nArg/nAAs nGlu/nAAs	Constitutional	85.1	73.5	56.2
GAs	4	ATS3ras ATS5ras MATS4ras GATS2ras	ATS	85.3	73.4	68.3
GAs	3	nGlu/nAAs nPro/nAAs MATS7ras	Constitutional / ATS	83.9	68.0	62.8
GAs	3	nGln nAsp/nAAs nGlu/nAAs	Constitutional	65.8	58.1	47.6
GAs	3	ATS1ras ATS5ras GATS5mw	ATS	73.3	54.8	51.3
ASM	2	nGlu/nAAs MATS7ras	Constitutional / ATS	67.1	51.6	51.5
ASM	2	nTrp nGlu/nAAs	Constitutional	65.0	46.0	43.7
ASM	2	ATS3mw ATS1ras	ATS	53.2	36.7	34.3

Table 8. Best models obtained for TBPL, using the physicochemical weighting scheme. ASM: all subset models, ATS: autocorrelation descriptor block

Method	Size	Variables	Descriptors	R^2	Q^2	Q^2_{boot}
GAs	4	nPro nArg/nAAs ATS4Am GATS5Dm	Constitutional / ATS	95.7	92.8	89.5
GAs	4	ATS3Km ATS1Dm ATS2Dm GATS1Dm	ATS	92.1	84.3	81.8
GAs	3	nPro ATS4Am GATS5Dm	Constitutional / ATS	92.9	88.7	87.9
GAs	3	WAm_sum nAsn / nAAs nAsp / nAAs	Constitutional	82.4	73.6	69.4
GAs	3	ATS6Km ATS1Dm MATS2Dm	ATS	80.1	70.8	68.6
ASM	2	GATS2Dm GATS6Dm	ATS	72.4	60.5	60.7

Table 9. Best models obtained for APTT, using the WHIM weighting scheme. ASM: all subset models, ATS: autocorrelation descriptor block

Method	Size	Variables	Descriptors	R^2	Q^2	Q^2_{boot}
GAs	4	nPro nGlu / nAAs ATS2Km ATS4Km	Constitutional / ATS	86.0	71.2	64.7
GAs	4	ATS1Dm ATS2Dm GATS3Km GATS1Dm	ATS	78.4	60.7	55.9
GAs	3	nGlu / nAAs nPro / nAAs MATS7Km	Constitutional / ATS	80.3	61.9	45.1
GAs	3	ATS1Dm ATS2Dm GATS1DM	ATS	72.8	57.9	55.1
ASM	2	nGlu / nAAs MATS7Km	Constitutional / ATS	64.2	48.6	39.6
ASM	2	ATS1Dm ATS2Dm	ATS	55.5	35.0	33.1

Table 10. Best models obtained for TBPL, using the WHIM weighting scheme. ASM: all subset models, ATS: autocorrelation descriptor block

Considering the physicochemical weighting scheme, 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, the APTT response is modelled using different constitutional descriptors, but the most frequent are the number of prolines (nPro) and the number of tryptophan (nTrp). The mostly selected autocorrelation descriptors 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 using four molecular descriptors is represented in Figure 2.

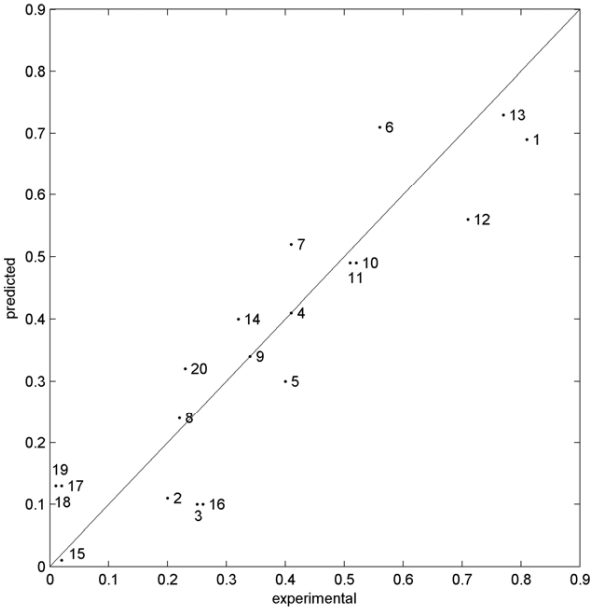

Figure 2. Experimental vs. predicted values of log(1 + APTT) for the best model obtained using the physicochemical weights. (nPro, nTrp, nAla/nAAs, nAsp/nAAs , Q^2 = 83.41)

TBPL on the contrary 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 descriptor is the relative frequency of glutamic acid (nGlu/nAAs) in a single peptide, that is selected in all the models containing at least one constitutional descriptor. The best model obtained for TBPL response using four molecular descriptors is represented in Figure 3.

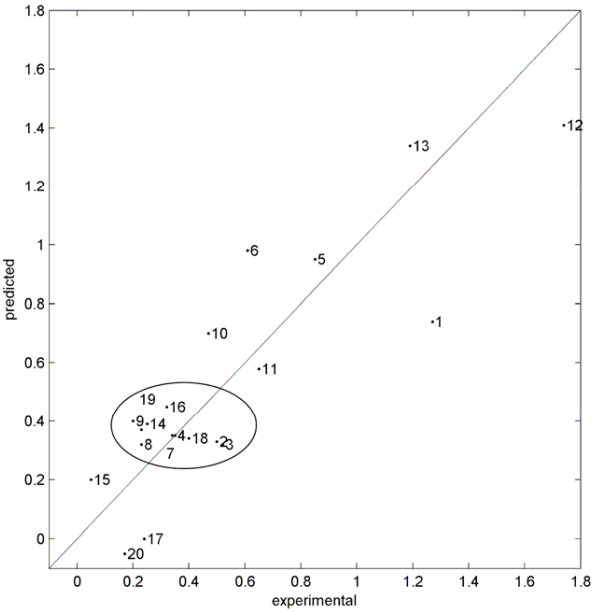

Figure 3. Experimental vs. predicted values of log(1 + TBPL) for the best model using the physicochemical weights. (nPhe, nGlu/nAAs, nPro/nAAs, MATS7ras, $Q^2 = 73.54$)

Considering the WHIM weighting scheme only one model being constituted only by constitutional descriptors is reported in Table 9. It includes the average sum of the WHIM global dimension index (WAm_sum). Anyway, considering APTT as modelled response, the models based on the WHIM weights are significantly better than the models obtained using

the physicochemical weighting scheme. 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 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%.

The mostly selected autocorrelation descriptors are weighted by the 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) occur in two of the five models containing autocorrelation descriptors. Considering the constitutional descriptors, the number of prolines (nPro) occurs in two different models and is always coupled with ATS4Am and GATS5Dm autocorrelation descriptors. The best model obtained for APTT response using four molecular descriptors weighted by the WHIM indices is represented in Figure 4.

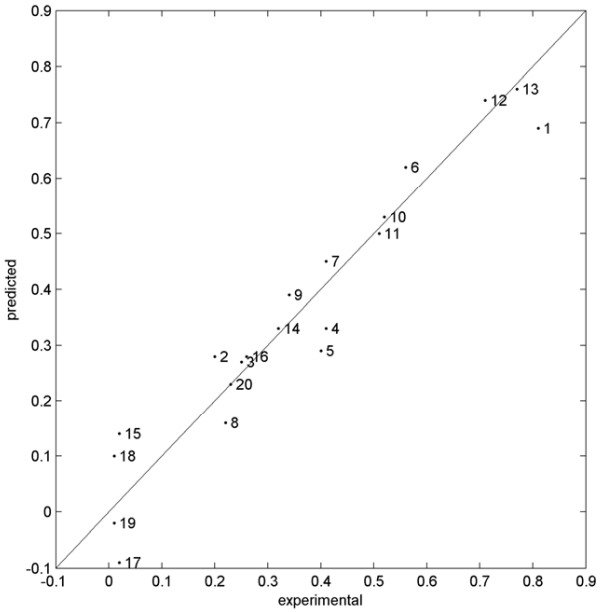

Figure 4. Experimental vs. predicted values of log(1 + APTT) for the best model obtained using the WHIM weights (nPro, nArg/nAAs, ATS4Am, GATS5Dm , Q^2 = 92.86)

Models obtained for TBPL response using the WHIM weighting scheme have a lower predictive power compared to those obtained using the physicochemical weighting scheme. Models constituted only by constitutional descriptors are omitted in Table 10 due to the fact that are the same models reported in Table 8, since no models for TBPL include weighted constitutional descriptors.

APTT is globally better modelled than TBPL; the reason is probably due to the not homogeneous distribution of the response values for TBPL. In Figure 3 a cluster of 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, while the greater part of the response space is not well represented, is usually an obstacle to build good models.

However, the models obtained using the proposed approach look significantly better than the models proposed in the literature. The best model proposed by Andersson et al. [26] was calculated using a modified z-scales approach and gave $R^2 = 86.2\%$ and $Q^2 = 60.3\%$, while the models proposed in this paper have Q^2 significantly higher, both for APTT and TBPL responses.

Conclusions

In this paper a new methodology for the characterisation of peptide sequences using a molecular descriptor based approach is presented.

Constitutional and 2D autocorrelation descriptors have been calculated by applying two different kinds of weights (the first based on physicochemical properties of the amino acids, the second based on WHIM descriptors) and used for the prediction of two biological responses on a dataset of 20 peptide sequences taken from the literature.

The presented 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 and appear stable and with good predictive power.

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, while 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.

References

[1] A. Kidera, Y. Konishi, M. Oka, T. Ooi, H. A. Scheraga, Statistical Analysis of the Physical Properties of the 20 Naturally Occuring Amino Acids, *J. Protein Chem.,* 1985*, 4*, 23-55

[2] S. Hellberg, M. Sjöström, B. Skagerberg, S. Wold, Peptide quantitative structure-activity relationships, a multivariate approach, *J. Med. Chem.,* 1987*, 30*, 1126-1135

[3] J. Jonsson, L. Eriksson, S. Hellberg, M. Sjöström, S. Wold, Multivariate parametrization of 55 coded and noncoded amino acids, *Quant. Struct.-Act. Relat.,* 1989*, 8*, 204-209

[4] S. Hellberg, L. Eriksson, J. Jonsson, F. Lindgren, M. Sjöström, B. Skagerberg, S. Wold, P. Andrews, Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships, *Int. J. Pept. Protein Res.,* 1991*, 37*, 414-424

[5] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids, *J. Med. Chem.,* 1998*, 41*, 2481-2491

[6] K. J. Siebert, Quantitative structure-activity relationship modeling of peptide and protein behavior as a function of amino acid composition, *J. Agric. Food. Chem.,* 2001*, 49*, 851-858

[7] K. J. Siebert, Modeling Protein Functional Properties from Amino Acid Composition, *J. Agric. Food Chem.,* 2003*, 51*, 7792-7797

[8] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, *Wiley - VCH,* 2000

[9] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, DRAGON Software: An Easy Approach to Molecular Descriptor Calculations, *MATCH Commun. Math. Comput. Chem.,* 2006*, 56*, 237-248

[10] Talete srl, DRAGON for Linux - Software for molecular descriptors calculation, 2007

[11] P. Broto, G. Moreau, C. Vandicke, Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies, *Eur. J. Med. Chem.,* 1984*, 19*, 66-70

[12] P. Broto, G. Moreau, C. Vandicke, Molecular structures: perception, autocorrelation descriptor and SAR studies, *Eur. J. Med. Chem.,* 1984*, 19*, 71-78

[13] P. Broto, G. Moreau, C. Vandicke, Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies, *Eur. J. Med. Chem.,* 1984*, 19*, 79-84

[14] P. A. P. Moran, Notes on continuous stochastic phenomena, *Biometrika,* 1950*, 37*, 17-23

[15] R. Geary, The contiguity ratio and statistical mapping, *Incorp. Statist.,* 1954*, 5*, 115-145

[16] S. Nakai, E. Li Chan, S. Hayakawa, Contribution of protein hydrophobicity to its functionality, *Nahrung,* 1986*, 30*, 327-336

[17] K. Tomii, M. Kanehisa, Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, *Protein Eng.,* 1996*, 9*, 27-36

[18] S. Kawashima, H. Ogata, M. Kanehisa, AAindex: Amino Acid Index Database, *Nucleic Acids Res.,* 1999*, 27*, 368-369

[19] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, *Nucleic Acids Res.,* 2000*, 28*, 374

[20] G. Fasman *(Ed.),* Handbook of Biochemistry and Molecular Biology, *CRC Press, Cleveland,* 1976*, 1*

[21] R. Grantham, Amino acid difference formula to help explain protein evolution, *Science,* 1974*, 185*, 862-864

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

[23] C. Chothia, The nature of the accessible and buried surfaces in proteins, *J. Mol. Biol.,* 1976*, 105*, 1-14

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

[25] R. Todeschini, P. Gramatica, The WHIM Theory: New 3D Molecular Descriptors for QSAR in Environmental Modelling, *SAR QSAR Environ. Res.,* 1997*, 7*, 89-115

[26] P. M. Andersson, M. Sjöström, T. Lundstedt, Preprocessing peptide sequences for multivariate sequence-property analysis, *Chemom. Intell. Lab. Syst.,* 1998*, 42*, 41-50

[27] D. E. Goldberg, Genetic algorithms in search, optimization and machine learning, *Addison-Wesley,* 1989

[28] R. Leardi, R. Boggia, M. Terrile, Genetic Algorithms as a strategy for feature selection, *J. Chemom.,* 1992*, 6*, 267-281

[29] R. Leardi, Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection, *J. Chemom.,* 1994*, 8*, 65-79

[30] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, *J. Chemom.,* 2001*, 15*, 559-569

[31] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, MobyDigs: software for regression and classification models by genetic algorithms, in: R. Leardi *(Ed.), Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*, Elsevier Science Inc., New York, NY, 2003, pp. 141-167

[32] Talete srl - MobyDigs for Windows (Software for the calculation of regression models using genetic algorithms for variable selection), Version 1.0, 2007

[33] B. Efron, Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics,* 1979*, 7*, 1-26

[34] B. Efron, The Jackknife, the Boostrap and Other Resampling Methods, *Society for Industrial and Applied mathematics,* 1982

[35] B. Efron, Better Bootstrap Confidence Intervals, *J. Am. Stat. Assoc.,* 1987*, 82*, 171-200

[36] F. Lindgren, B. Hansen, W. Karcher, M. Sjöström, L. Eriksson, Model validation by permutation tests: applications to variable selection, *J. Chemom.,* 1996*, 10*, 521–532

[37] L. Eriksson, E. Johansson, S. Wold, QSAR model validation, in: F. Chen, G. Schuurmann *(Eds.), Quantitative Structure-Activity Relationships in Environmental Sciences—VII,* SETAC Press, Pensacola, FL, 1997, pp. 381-397