MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies

A. Zaliani* and E. Gancia

Computer Chemistry, Italfarmaco Research Center, via dei lavoratori 54, I-20092 Cinisello Balsamo, Milan, Italy

Received June 28, 1998

Several descriptors applied to peptide structure−activity and/or structure−property relationships have been developed in recent years. This report describes 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. MS-WHIM indexes are a collection of 36 statistical indexes aimed at extracting and condensing steric and electrostatic 3D-properties of a molecule. These new descriptors have been developed both on extended side-chain conformation and on rotamer library of natural amino acids. The method appeared to be more potent when describing a single conformation (*i.e*. extended) than when applied collectively on library conformation families. MS-WHIM scores, however, were shown to efficacely describe and correctly classify the natural amino acid features and to provide sound statistical models either predicting activities of two peptide sets taken as a test or correlating amino acid chemicophysical properties, like water-accessible surface area, to general structural features of amino acids.

INTRODUCTION

Coding of 3-D structural information is an open question both for scientific and general data management. Chemists tackled this problem providing successful reports in quantitative structure−activity (QSAR) and structure−property relationship (QSPR) studies which showed how important structural coding, or *description*, may be in relating three-dimensional features to biological or chemicophysical properties. Despite difficulties in QSAR analyses of peptides, there have been encouraging results in this particular field as well.[1]

It has become evident that meaningful quantitative correlations for structure−activity relationship studies must be obtained on the basis of the active, actual or most probable, conformation of the ligands. In the case of peptidic ligands, this is usually a hard task without spectroscopical evidences. Peptide flexibility renders coding of the 3D-structural information an extremely difficult target on the ground of local (*i.e*. oriented) description either for the number of conformations or for the amount of descriptive points sampled.

Chemometric strategies allowing for circumventing problems arising from peptide flexibility and those related to local and precise description have been undertaken. Monodimensional and holistic descriptors were the first to be analyzed due to their reduced number, simplicity, and invariance to rototranslation, especially when applied to peptides. A rather successful approach was based on the most basilar of the peptide structural information: its sequence. This strategy tries to reduce structural complexity by taking advantage of the description of singular residues and translates the peptide sequence into a vector of descriptors. In this way, Kidera *et al*.[2] 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*.[3] who developed principal properties, or *z-scores*, for each of 20 natural amino acids and for a series of unnatural ones. *z*-Scores were extracted through PCA from a collection of experimental data on a number of peptides, like HPLC retention times, pK_as, NMR-derived properties, and other measurable variables related to hydrophobicity, size, and electronic features. 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.[4] Similar results were obtained by Cocchi *et al*.[5] 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,[6] here defined as *t*-scores, turned out to be effective when applied in a QSAR study of a set of dipeptide ACE inhibitors. Recently, Collantes *et al*.[7] showed that two computable 3D-descriptors, Isotropic Surface Area (ISA) and Electronic Charge Index (ECI), may be usefully applied as side-chain descriptors. While ISA correlates well with *z*1-score values and with Fauchere and Pliska's hydrophilicity scale,[8] ECI showed good correlation with amino acid free energy of vaporization.[9]

These results provided evidence that calculated structurally-derived properties can be used to generate robust description for residues in a peptidic sequence. Nevertheless, only Cocchi's *t*-values and ISA and ECI values take explicitly into account 3D-structural features of the residues.

In a recent report, we showed that the original WHIM approach,[10] when applied on Molecular Surfaces (MS-WHIM),[11] may constitute a valid tool for coding both structural and electronic features of a molecule in a multi-

* Corresponding author. Phone: +39-2-64 433 095; fax: +39-2-66 011 579; e-mail: andrea@edith.sublink.org.

10.1021/ci980211b CCC: $18.00
Published on Web 04/17/1999

dimensional vector. The success of this approach in obtaining QSAR models on organic molecules[12] prompted us to apply it to natural amino acids. By applying PCA on a MS-WHIM description matrix thus derived and by extracting scores associated with the three first major components, MSW-scores were obtained for each residue and are here described. This numerical code allows for classifying residues into families and making some considerations on the genetic code, finding interesting structure−property correlation, and constructing quantitative regression models predicting biological activity of test peptide sequences.

METHODS

Structures of Amino Acids. The structures of the 20 natural α-amino acids were built within Sybyl 6.2 package (Tripos Inc., St. Louis, MO) using the BIO BUILD command and the default peptide dictionary. N- and C-terminal backbone groups were set in their neutral form, while side chains were set in their charged form where appropriate. Side chains were built in either a fully extended conformation (extended set) or in the conformations deduced from the rotamer library[13] (rotamer set). Because the torsion angles χ_1 and χ_2 of Asn and Gln were not contained in that library,[13] they were taken from a rotameric library by Ponder and Richards.[14]

Charge and Geometries Optimization. All the amino acid structures were submitted to semiempirical AM1 calculations in MOPAC 6.0 (QCPE Program no. 455) with keywords NOINTER EF GNORM = 0. Where necessary the MMOK keyword for amide bond correction was used. AM1 charges were used in the following step.

Molecular Electrostatic Potential (MEP) Calculation on Connolly Molecular Surface. Optimized structures were submitted to Connolly's surface calculations (MS program QCPE no. 429) with a probe atom of radius 1.5 Å and a density of 10 points/Å^2. An internally developed program[11] computed MEP values on each point of Connolly's surfaces. MEP values were calculated by means of the classical Coulomb equation using a distance-dependent dielectric constant (ϵ)

$$V_p = \sum_i q_i/\epsilon|r_i| \qquad \text{(I)}$$

where V_p is the MEP value (in kcal/mol*Å) relative to point p and r_i is the distance between p and the ith atom.

MS-WHIM Description Matrices. MS-WHIM descriptors for each amino acid contained in the extended set and in the rotamer set were calculated from the Connolly surface files containing MEP values. The following procedure was applied to each molecular surface within each of the three weighting schemes described in ref 11: (1) unweighted case (*i.e.* weight = 1), (2) positive MEP, and (3) absolute value of negative MEP. This gave a total of 12 × 3 = 36 molecular descriptors. The original MS-WHIM procedure applied was as described in ref 11 and here is briefly summarized:

(1) the *x*-*y*-*z*-coordinates of each surface point are centered with respect to their weighted means;

(2) weighted PCA on the centered data is performed to obtain the score matrix **T** in the three principal component axes;

(3) for each weight the following parameters[11] are computed from the **T** matrix ($i = 1 - n$, $m = 1-3$):

(a) variance:

$$\text{PCA eigenvalues} = \lambda_m = \sum_i w_i t_{im}^2 / \sum_i w_i$$

(b) eigenvalue proportions:

$$\theta_m = \lambda_m / \sum_m \lambda_m \qquad \text{(II)}$$

(c) skewness:

$$\gamma_m = |[\sum_i (w_i t_{im}^3) / \sum_i w_i]| * 1/\lambda_m^{3/2} \qquad \text{(III)}$$

(d) kurtosis:

$$\kappa_m = [\sum_i (w_i t_{im}^4) / \sum_i w_i] * 1/\lambda_m^2 \qquad \text{(IV)}$$

The PCA eigenvalues (λ_m-I) physically refer to the coordinate variances of the molecules. Eigenvalue proportions (θ_m-II) are related to surface shape as surfaces of flat molecules (*e.g.* aromatic rings) have only two major components.

θ_3 has been replaced by the acentric factor $\omega = \theta_1 - \theta_3$;[15] spherical surfaces have null acentric factor, while linear ones will have $\omega = \theta_1 = 1$. Skewness (γ_m-III) can be associated with the surface distribution symmetry along each component. Being a third-order moment it could assume negative values; to preserve the invariance to rotation its absolute value was considered. This theoretical implication precludes the possibility to discriminate between surfaces derived from enantiomeric structures through the sign of symmetry without taking into account their spatial orientation. To avoid alignment problems and to preserve the holistic character of the description we preferred not to use signed skewnesses, even though, for example, amino acids of the D-series would have been easily differentiated through skewness signs from those of the L-series.

Finally, the Kurtosis index (κ_m-IV) relates to points distribution on the Connolly's surface and their density around the center of mass and along principal axes. To avoid problems related to infinite κ_3 values, obtained, for example, with "flat" surfaces, the reciprocal of this entity, $\eta_m = 1/\kappa_m$, was used as defined.[15] A total of 12 MS-WHIM indices are thus computed for each weight:

$$\lambda 1, \lambda 2, \lambda 3, \theta 1, \theta 2, \omega, \gamma 1, \gamma 2, \gamma 3, \eta 1, \eta 2, \eta 3$$

In relation to the kind of weights assigned to points, different types of information can be extracted. In the unitary case (*i.e.* weighting scheme $w_{ii} = 1$, $i = 1 - n$) size info can be extracted, as differently charged points are not distinguished. MEP values as weights have a more direct physical meaning as they depict the distribution of the electrostatic potential onto the surface. The information obtained within these schemes (unitary + MEP values) may be thus referred to (i) the molecular volume distribution and (ii) the electrostatic potential distribution. The need to split MEP values in two complementary matrices arises from the need to have a semipositive weight scheme. Thus the description matrix for the 20 amino acids was defined as having 20 rows × 36 columns (the MS-WHIM molecular descriptors).

MS-WHIM Description of the Amino Acid Rotameric Collection. The torsion angles χ_1 and χ_2 of each side chain

Table 1. Three MS-WHIM Scores for Amino Acid Extended Conformation

amino acid	one-letter code	MS-WHIM scores[a] first	second	third	design coordinates[b]		
Val	V	−1.00	0.79	−0.58	−	+	−
Ile	I	−0.91	0.83	−0.25	−	+	−
Leu	L	−0.74	0.72	−0.16	−	+	−
Met	M	−0.70	1.00	−0.32	−	+	−
Ala	A	−0.73	0.20	−0.62	−	+	−
Pro	P	−0.43	0.73	−0.60	−	+	−
Cys	C	−0.66	0.26	−0.27	−	+	−
Ser	S	−0.80	0.61	−1.00	−	+	−
Thr	T	−0.58	0.85	−0.89	−	+	−
Lys	K	−0.51	0.08	0.60	−	+	+
Arg	R	−0.22	0.27	1.00	−	+	+
Asn	N	0.14	0.20	−0.66	+	+	−
Gln	Q	0.30	1.00	−0.30	+	+	−
His	H	0.84	0.67	−0.78	+	+	−
Phe	F	0.76	0.85	−0.34	+	+	−
Trp	W	1.00	0.98	−0.47	+	+	−
Tyr	Y	0.97	0.66	−0.16	+	+	−
Asp	D	0.11	−1.00	−0.96	+	−	−
Glu	E	0.24	−0.39	−0.04	+	−	−
Gly	G	−0.31	−0.28	−0.75	−	−	−

[a] The range-scaled scores (within −1/+1) derived from autoscaled PCA upon MS-WHIM indexes description matrix for 20 natural amino acids in extended side-chain conformation (see Methods section). [b] Design coordinates are related to the eight possible octants in the three-dimensional factorial design cube.

were varied following the collection of Schrauber's library.[13] Each conformer was submitted to the previously described calculations for charges, surface, and MEP values determination. The resulting output files of each conformation, containing the surface points with their own weights (unitary and MEP values), were concatenated in order to obtain only one text file for residue. Any of these residue files collects all the different surfaces the residue possesses. Each point belonging to the same surface (*i.e.* each residue conformer) was weighted for the population percentage of the conformer to which it belongs. Finally, MS-WHIM calculation applied on files so manipulated led to a MS-WHIM vector for each residue type and, collectively, to a description matrix of 20 rows × 36 descriptors similar to the previous one from the extended set. Both matrices underwent the following PC analysis.

Principal Component Analysis and Amino Acid MSW-Scores Extraction. Amino acid description matrices described above were submitted to PCA within the Tripos package Sybyl 6.2. Autoscaling was always applied. The PCA scores were extracted for all those components which were identified to be necessary to explain a minimum percentage of 60% of the total variance. Range scaling between −1 and 1 was applied to the MSW-scores for a homogeneous graphical representation.

PLS Models and Multiple Regressions. All the PLS models were obtained within QSAR module of Sybyl 6.2 package. Multiple regressions were obtained with BMDP 1.0 package (BMPD Inc., Los Angeles, CA) through stepwise multilinear regression analysis.

RESULTS AND DISCUSSION

Extended Conformations. Table 1 shows the range-scaled MSW-scores from cross-validated PC analysis of the matrix containing MS-WHIM descriptors for the 20 natural amino acids in extended side-chain conformation. The first three components explain 61% of the total variance, while this percentage increases to 87% when the first seven components, which entered the definition of *t*-scores,[5] are considered.

The first component (see Table 2 for loadings) contains as more informative descriptors those derived from electrostatic potential, both in terms of MEP sign and size of the side chains. This component discriminates between positive and negative charged residues and between aromatic and bulkier aliphatic. Interestingly, the component is only slightly

Table 2. PCA Loadings for the Extended Set

	λ1_O	λ2_O	λ3_O	θ1_O	θ2_O	ω_O
1 Factor1	0.569	0.388	−0.341	0.626	−0.454	0.668
1 Factor2	0.133	0.624	0.334	−0.081	0.209	−0.030
1 Factor3	**0.763**[a]	0.113	0.127	0.738	**−0.751**	0.705
1 Factor4	0.125	0.150	0.727	0.066	−0.230	0.000417
1 Factor5	0.140	0.115	0.140	0.063	−0.070	0.055
1 Factor6	0.084	−0.099	−0.045	0.040	−0.070	0.027
1 Factor7	0.034	0.466	−0.047	−0.129	0.282	−0.065

	γ1_O	γ2_O	γ3_O	η1_O	η2_O	η3_O
1 Factor1	−0.324	−0.125	0.267	−0.046	0.377	−0.401
1 Factor2	0.001	−0.530	−0.067	0.163	0.437	−0.140
1 Factor3	0.291	0.172	−0.386	0.538	−0.175	0.391
1 Factor4	−0.005	0.596	−0.646	0.285	−0.357	0.441
1 Factor5	−0.389	−0.396	0.204	0.345	0.536	−0.252
1 Factor6	0.543	−0.075	0.103	0.387	0.083	0.253
1 Factor7	0.236	−0.146	−0.238	0.059	−0.172	0.345

	λ1_P	λ2_P	λ3_P	θ1_P	θ2_P	ω_P
1 Factor1	0.617	0.365	−0.511	0.414	−0.029	0.568
I Factor2	0.310	**0.847**	0.746	**−0.820**	**0.903**	−0.710
1 Factor3	0.588	0.092	0.123	0.284	−0.242	0.281
1 Factor4	−0.185	0.004	0.064	−0.094	0.070	−0.097
1 Factor5	−0.005	0.072	0.069	−0.038	0.017	−0.044
1 Factor6	−0.160	0.029	0.025	−0.167	0.169	−0.153
1 Factor7	0.137	0.226	−0.330	0.091	0.139	0.196

	γ1_P	γ2_P	γ3_P	η1_P	η2_P	η3_P
1 Factor1	−0.140	0.231	0.250	−0.075	0.029	−0.632
1 Factor2	−0.679	−0.462	−0.183	0.144	0.792	−0.036
1 Factor3	−0.173	−0.525	−0.323	0.065	0.423	0.350
1 Factor4	0.395	0.366	−0.110	−0.552	−0.278	0.118
1 Factor5	0.420	0.035	0.324	−0.402	−0.081	0.031
1 Factor6	0.227	0.230	0.630	−0.412	0.132	−0.348
1 Factor7	0.033	−0.322	0.207	0.218	−0.006	−0.216

	λ1_N	λ2_N	λ3_N	θ1_N	θ2_N	ω_N
1 Factor1	**0.905**	0.393	0.691	**0.904**	**−0.863**	0.908
1 Factor2	0.187	−0.425	0.124	0.126	−0.193	0.084
1 Factor3	−0.122	−0.405	−0.453	−0.117	0.207	−0.059
1 Factor4	0.158	0.189	0.372	0.118	−0.213	0.059
1 Factor5	−0.239	0.144	0.103	−0.258	0.193	−0.290
1 Factor6	0.031	−0.261	−0.004	0.007	−0.031	−0.008
1 Factor7	−0.023	0.569	0.239	−0.135	0.099	−0.154

	γ1_N	γ2_N	γ3_N	η1_N	η2_N	η3_N
1 Factor1	−0.756	−0.641	0.039	0.759	**0.836**	−0.037
1 Factor2	−0.173	0.319	0.416	0.179	−0.202	−0.467
1 Factor3	0.441	−0.182	−0.208	−0.326	−0.081	0.195
1 Factor4	−0.094	0.211	0.506	0.085	−0.044	−0.616
1 Factor5	0.345	−0.463	0.376	−0.408	0.385	−0.099
1 Factor6	−0.116	0.103	−0.372	0.162	−0.130	0.346
1 Factor7	0.116	0.152	−0.073	−0.175	0.114	0.063

[a] Bold typeface for the most representative in the first three components.

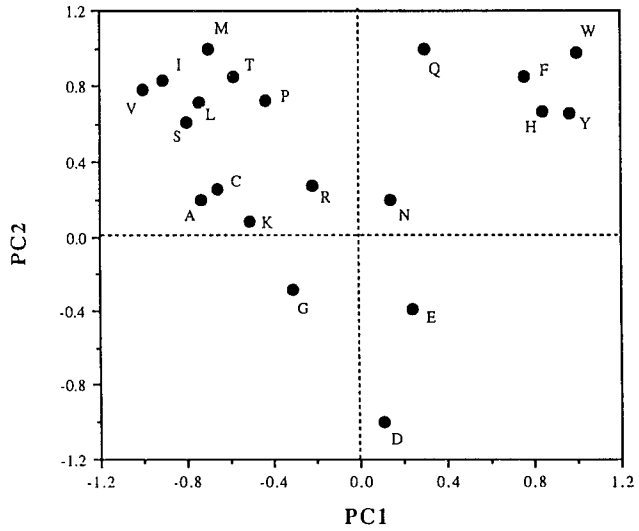

Figure 1. Score plot of the first and second components of Principal Component Analysis applied on the MS-WHIM description matrix of the extended conformers of the 20 natural amino acids. Notice the nonhomogeneous distribution of the scores in the component space (see text for details).

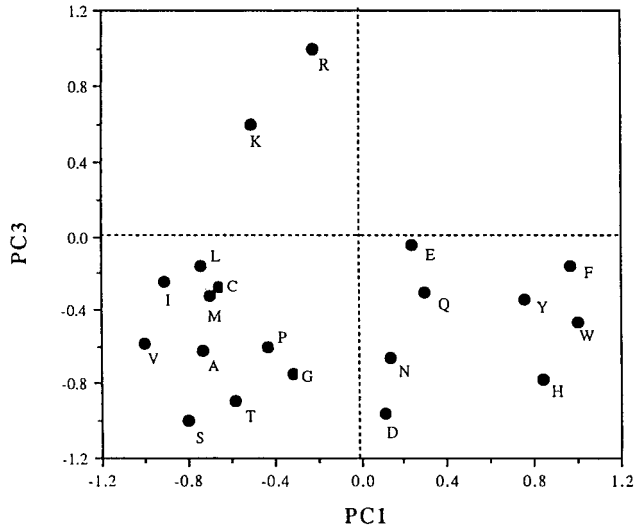

Figure 2. Score plot as Figure 1 but of the first and third components. The third component provides a neat differentiation for the positively charged residues only.

influenced by the 12 unitary descriptors, while the eigenvalues (λs) derived from negative and positive MEP surfaces are well represented. Figures 1 and 2 map graphically the 20 residues in the principal component space. Positively charged residues cluster together with aliphatic residues, while negatively charged residues with their corresponding amidated derivatives occupy the opposite side of the map. These latter form a tight group of their own. It is likely that surface planarity and charge distribution of the amide group render Gln and Asn more similar to aromatic than aliphatic residues. Within the aromatic cluster, histidine and tryptophan are well positioned and are not outliers.[5]

The second component is influenced by descriptors derived from positive MEP values and thus allows for discrimination of Asp and Glu from the other residues. Besides these, only glycine has a negative second score being the only residue described with all negative component values. In this respect, Gly appears to be an outlier; the likely reason for that being the lack of a side chain.

Intriguingly, no residue can be described with triple positive scores or a minus−minus−plus combination. Finally, the third component is generally negative and, similarly to the second, is not influenced by descriptors derived from negative MEP. Arg and Lys differentiate well from the other residues both because of their positive charge and, above all, for their threadlike shape. In fact, the third component is negatively influenced by $\theta 2$ and positively by $\theta 1$. Assuming that signs of range-scaled MSW-scores codify for the eight possible octants in the three-dimensional factorial design cube, one can in principle correlate the different triplets with the extent of molecular dissimilarity. In other words, if all the available chemical space can be described by the first three components, it is clear that amino acids do not cover the entire diversity space. Of the eight possible triplets, only six are represented. The lack of the $+++$ and of the $--+$ triplets suggests that, within this type of description, side chains do not span all the theoretical possibilities. Selection may therefore have occurred during evolution. Codon analysis previously suggested similar considerations.[16] For example, arginine may be considered an exception in the context of genetic code,[17] and the high number of its codons conflicts with its abundance in naturally occurring proteins.[17]

MSW-scores from the extended conformation set showed no significant (*i.e.* $r^2 > 0.6$) direct correlations with other known scores such as *z*-scores[4] or *t*-scores.[5] Only the second MSW-score showed a correlation ($r^2 = 0.62$) with nonpolar water-accessible surface area[13] which may be obvious given the definition of MS-WHIM indexes.

Rotamer Library Conformations. The rotameric library from Schrauber *et al.*[3] was used. Accordingly, each conformer has a percentage of population associated with it. Glycine, alanine, and proline have only one possible rotamer, while isoleucine counts up to seven different families just considering the two first χs. So, for each residue, we took into account all the reported rotamer surfaces summed up in one, and we weighted any points on them for their population percentages. Table 3 contains range-scaled MSW-scores from rotamer set. The total amount of explained variance from the first three components was 72%, better than with the extended set (61%), while to reach the same information content of 87% five, instead of seven, components had to be selected. The higher extent of the explained variance does not guarantee *per se* a more precise description; instead it may be a sign of loss of structural information caused by a severe drop in variance due to the sum of geometrically very different surfaces within each residue family. The more flexible the side chain is the more dramatically this consideration might apply. For this point, refer to the discussion at the end of the section and to Table 7 for statistical comparison of the two parent MS-WHIM descriptions.

Figures 3 and 4 map the residue scores in the principal component space, and in Table 3 the results are compared with those from the extended set: the descripting and recognizing power of MSW-scores was greatly reduced in the rotamer case. Thus large hydrophilic residues like Asn and Gln cannot be easily differentiated from positively charged ones and from Met or, surprisingly, from Pro. However, negatively charged residues are found opposite to the positively charged ones on all of the first three compo-

Table 3. Three MS-WIIM Scores for Amino Acid Rotameric Families[a]

amino acid	one-letter code	MS-WHIM scores[b] first	second	third	design coordinates[c]		
Leu	L	−0.05	−0.62	−0.45	−	−	−
Ile	I	−0.07	−0.29	−0.08	−	−	−
Met	M	0.29	−0.39	−0.29	+	−	−
Pro	P	0.13	−1.00	−0.98	+	−	−
Asn	N	0.06	−0.33	−0.25	+	−	−
Gln	Q	0.41	−0.02	−0.50	+	−	−
Lys	K	0.29	−0.98	−0.70	+	−	−
Arg	R	0.35	−0.94	0.39	+	−	+
Gly	G	−1.00	−0.43	0.84	−	−	+
Ala	A	−0.73	−0.55	0.62	−	−	+
Cys	C	−0.36	−0.02	0.17	−	−	+
Ser	S	−0.36	−0.46	0.28	−	−	+
Val	V	−0.33	−0.33	0.17	−	−	+
Thr	T	−0.30	−0.29	0.22	−	−	+
Asp	D	−0.53	0.90	−0.53	−	+	−
Glu	E	−0.32	1.00	−1.00	−	+	−
His	H	0.53	−0.10	0.27	+	−	+
Phe	F	0.72	0.00	0.43	+	+/−	+
Trp	W	0.97	0.26	0.24	+	+	+
Tyr	Y	1.00	0.07	1.00	+	+	+

[a] See ref 13 for rotamer library used. [b] The range-scaled scores (within −1/+1) derived from autoscaled PCA upon MS-WHIM indexes description matrix for 20 natural amino acids in extended side-chain conformation (see Methods section). [c] Design coordinates are related to the eight possible octants in the three-dimensional factorial design cube.

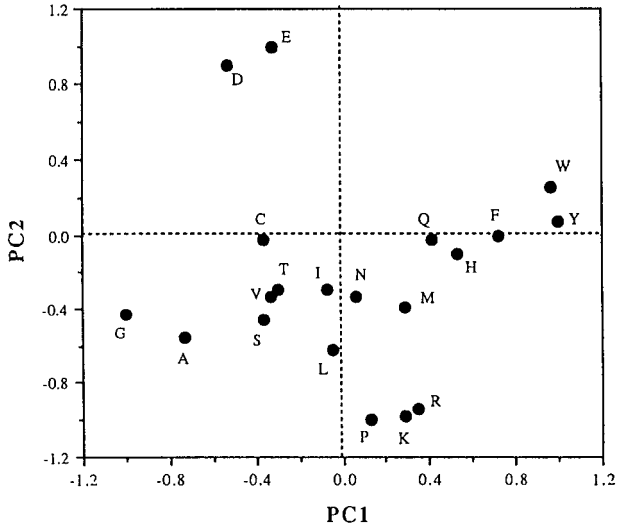

Figure 3. Score plot of the first and second components of Principal Component Analysis applied on the MS-WHIM description matrix of the 20 natural amino acids from a weighted rotameric library. Comparison with Figure 1 provides evidence for a less pronounced clustering of different classes of amino acids (see Results and Discussion).

nents, while aromatic residues are still completely clusterized and separated on the positive first and third component. A certain ambiguity between small lipo- and small hydrophilic residues is not resolved with this description, and this fact may be likely explained by their very similar shape and size (*i.e.* eigenvalues [λs] and proportions [θs] of unitary weight of Thr and Leu are, of course, very similar to having similar graph representations and dimensions). These strict similarities are even more pronounced with rotameric characterization as these residues possess only one side-chain torsion and their possibility of shape differentiation are so limited.

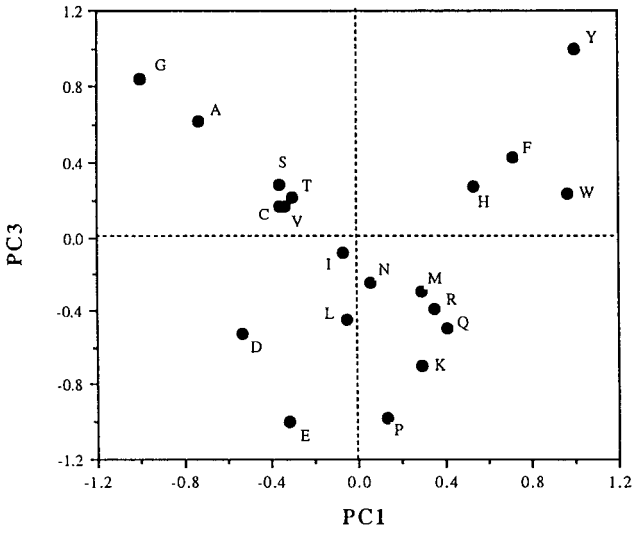

Figure 4. Score plot as Figure 3 but of the first and third components. The third component provides a neat differentiation for aromatics and small-sized residues.

The first MSW-score in the rotameric set turned out to be correlated with the first *t*-scores[5] ($r^2 = 0.68$). As with the extended set, the first MSW-score correlates to nonpolar water-accessible surface area[9] (ASAnpol) but in a more intriguing and detailed way. The following function (V), which relates to the first MSW-score and ASAnpol and hydration enthalpies[19,20] (DHydr), was found:

$$\text{MSW-score}_1 = -1.136 + 0.008*\text{ASAnpol} - 0.108*\text{DHydr}$$
$$N = 20, \quad F = 37.99, \quad r = 0.82 \qquad \text{(V)}$$

This function adds evidence to the previously mentioned mix of electronic and shape effects which seems to most influence the first MSW-score and indirectly underlines the known reciprocal and opposite effects that water-accessible surface has on hydration enthalpies.

Peptide QSAR Application. 1. ACE Dipeptide Inhibitors. To test the MSW-scores description in peptide QSAR, we applied them to two panels of dipeptides. In both dipeptide sets the sequences were coded replacing each residue with the first three MSW-scores for a total of just six columns of descriptors. Only Collantes *et al.*[7] introducing Isotropic Surface Area (ISA) and Electronic Charge Index (ECI) used fewer descriptor columns. The first panel was a series of 58 ACE inibitor dipeptides taken from Cocchi's report.[5] Table 4 shows the list of the dipeptides together with observed and calculated activities, while Table 5 summarizes the most important statistical parameters of PLS regression compared to those obtained with *t*-scores.

PLS models from MSW-scores derived from extended conformation maintained a superior modeling power relative to rotameric MSW-scores. This is evident from both the number of components needed and *r*-squared values. Even though MSW-scores performed similar or inferior fitting regression indexes, deeper comparisons should take into account cross-validated parameters and predictions. Because for the former no indications have been published, Table 4 shows that MSW-scores performed satisfactorily in predicting ACE inhibiting potencies of the dipeptides. Descriptor weights of the PLS model revealed that descriptors belonging

Table 4. Dipeptides Panel and Observed and Calculated log[1/(IC_{50})] Values for Inhibition of ACE

no.	peptide	obs[a] act.	calcd[b] act. *t*-scores	calcd[c] act. MSW-scores
1	VW	5.80	5.45	4.63
2	IW	5.70	5.49	4.74
3	IY	5.43	4.12	4.45
4	AW	5.00	4.81	4.30
5	RW	4.80	5.53	4.75
6	VY	4.66	4.08	4.34
7	GW	4.52	4.61	3.92
8	VF	4.28	4.12	4.34
9	AY	4.06	3.45	4.00
10	IP	3.89	3.36	3.71
11	RP	3.74	3.40	3.72
12	AF	3.72	3.49	4.04
13	GY	3.68	3.24	3.63
14	AP	3.64	2.69	3.26
15	RF	3.64	4.20	4.50
16	VP	3.38	3.33	3.60
17	GP	3.35	2.48	2.89
18	GF	3.20	3.28	3.67
19	IF	3.03	4.16	4.61
20	VG	2.96	2.87	2.76
21	IG	2.92	2.90	2.87
22	GI	2.92	2.97	2.72
23	GM	2.85	2.77	2.99
24	GA	2.70	2.32	2.24
25	YG	2.70	2.32	2.28
26	GL	2.60	2.69	2.72
27	AG	2.60	2.23	2.42
28	GH	2.51	3.18	3.53
29	GR	2.49	2.70	2.67
30	KG	2.49	2.84	2.74
31	FG	2.43	2.30	2.36
32	GS	2.42	2.34	2.55
33	GV	2.34	3.10	2.85
34	MG	2.32	2.61	2.32
35	GK	2.27	2.66	2.29
36	GE	2.27	2.56	2.29
37	GT	2.24	2.67	2.87
38	WG	2.23	2.58	2.30
39	HG	2.20	2.15	2.10
40	GQ	2.15	2.37	3.54
41	GG	2.14	2.03	2.05
42	QG	2.13	2.33	2.57
43	SG	2.07	2.43	2.47
44	LG	2.06	2.40	2.80
45	GD	2.04	3.28	1.63
46	TG	2.00	2.97	2.55
47	EG	2.00	2.19	2.09
48	DG	1.85	2.13	1.55
49	PG	1.77	1.94	2.56
50	LA	3.51	2.70	3.00
51	KA	3.42	3.13	2.94
52	RA	3.34	3.24	3.08
53	YA	3.34	2.61	2.48
54	AA	3.21	2.52	2.61
55	FR	3.04	2.97	2.98
56	HL	2.49	2.81	2.77
57	DA	2.42	2.42	1.75
58	EA	2.00	2.48	2.29

[a] Inhibitory activity taken from ref 5. [b] Taken from ref 5. The first six *t*-scores were used. [c] The first three component of the extended set were used.

to the first residue of the sequence do not contribute to the models, while the first two MSW-scores of the second had the highest weights. This correlates to what is already known about this data set;[5] thus, aromatic residues possessing positive first and second MSW-scores turned out to be the preferred C-terminal residues, while the aliphatic side chain provided intermediate inhibitory activity having a negative first and a positive second MSW-scores. Moreover, regression eq V showed that improved first MSW-scores can be obtained with high values of ASAnpol. This fact applies well to aromatic residues, expecially tryptophan, which possess the highest ASAnpol value.[20]

Peptide QSAR Application. 2. Bitter Tasting Dipeptides. The second panel used to test MS-WHIM scores was the set of 48 bitter tasting dipeptides.[4] Table 6 shows the dipeptides together with observed and calculated activities from different sources. Table 5 reports the most significant statistical parameters of PLS regression. As in the previous study, MSW-scores derived from extended conformation performed better than the corresponding scores from rotameric library. This confirms the worse quality of the original description contained in the latter. Comparison with results from Collantes *et al*. is difficult as few details of the PLS model were given by the authors. Nevertheless, PLS model from MSW-scores description turned out to be again slightly less fitting in statistics (Table 5) but completely reliable for what concerns predictions (Table 6). *z*-Scores performed here better than MSW-scores, as referred by cross-validated regression index. The weights of descriptors in the PLS model revealed that the third MSW-score of the first residue and the second MSW-score of the second residue were the most influencing. This fact relates to the size dependence of the second MSW-scores (see in Table 1 large aliphatic and aromatic residues) and with the charge/shape dependence of the third MSW-scores. Similar findings were outlined in Collantes *et al.* where ECI for position 1 and ISA for position 2 were found to be the most influencing descriptors of a very accurate PLS model ($r^2 = 0.847$). The predictions of the activities made with MSW-scores model were accurate in a similar extent relative to those of literature, but, unexpectedly, peptides containing isoleucine in position 1 provided the worst predictions of the series. In contrast to what was seen by Collantes *et al*., the residue in position 2 should be positively and not negatively charged. Unfortunately, the data set contained only one such dipeptide (IK) and only three dipeptides containing Asp or Glu residues in position 2. Further reasoning upon them may be misleading although it should be noticed that the IK dipeptide showed the highest bitterness relative to these three cases.

Comparison between Extended and Rotamer Library Conformations. MSW-scores derived from amino acids in one conformation performed better than those derived from the weighted sum of conformer families in both of the above QSAR studies. This apparently surprising result needs more comments. The strong dependence of MS-WHIM from molecular geometry[11,12] is crucial to understand the behavior of the rotameric set. In this case, the larger amount of points, their larger scattering (due to the copresence of multiple and different surfaces), and, above all, the weighting have certainly reduced the geometrical specificity of each residue. For comparison, in Table 7 basic statistics of some descriptors of MS-WHIM matrices, the eigenvalues (λs), where MSW-scores come from have been summarized. The eigenvalues contain variance information by definition. Eigenvalues from a rotameric set are always smaller than those from an extended set, and this applies both on their mean and on their range. The reason for that is mainly due to the weighting for rotameric population applied to each point of the multisurface collections; each weight has to be always

Table 5. Statistical Parameters of PLS Model in QSAR of 58 ACE Inhibitors and 48 Sweetener Dipeptides

peptide set (reference)	descriptors per residue	no. components	r^2 (fitting)	r^2 (cross-validated)[a]
ACE (Cocchi *et al.*)	7	1	0.744	nd[b]
ACE (Collantes *et al.*)	2	nd[b]	0.700	nd[b]
ACE (MSW-scores extended)	3	2	0.708	0.637
ACE (MSW-scores rotameric)	3	6	0.657	0.541
sweeteners (Jonsson *et al.*)	3	1	nd[b]	0.780
sweeteners (Collantes *et al.*)	2	2	0.847	nd[b]
sweeteners (MSW-scores extended)	3	3	0.754	0.710
sweeteners (MSW-scores rotameric)	3	3	0.704	0.633

[a] Leave-group-out cross-validation performed 100 times through five groups of 10−12 randomly chosen objects. [b] Not determined.

Table 6. Dipeptides Panel and Observed and Calculated log[1/*T*] Values for Bitter Tasting Activity

no.	peptide	obs act.[a]	calcd act.[b] *z*-scores	calcd act.[c] MS-WHIM scores
1	GV	1.13	1.02	1.23
2	GL	1.68	1.35	1.35
3	GI	1.70	1.32	1.38
4	GP	1.35	1.21	1.40
5	GF	1.80	1.68	1.84
6	GW	1.89	1.87	1.97
7	GY	1.77	1.30	1.80
8	AV	1.16	1.57	1.49
9	AL	1.70	1.90	1.54
10	AF	1.72	2.23	1.99
11	VG	1.19	0.95	1.04
12	VA	1.16	1.54	1.33
13	VV	1.71	1.79	1.66
14	VL	2.00	2.12	1.77
15	LG	1.72	1.25	1.44
16	LA	1.72	1.84	1.68
17	LL	2.35	2.42	2.46
18	LF	2.75	2.75	2.65
19	LW	3.40	2.94	2.68
20	LY	2.46	2.37	2.57
21	IG	1.68	1.22	1.39
22	IA	1.68	1.80	1.65
23	IV	2.05	2.06	2.03
24	IL	2.26	2.39	2.07
25	II	2.26	2.36	2.11
26	IP	2.40	2.25	2.06
27	IW	3.05	2.91	2.70
28	IN	1.49	1.69	1.80
29	ID	1.37	1.64	0.87
30	IQ	1.49	1.69	2.64
31	IE	1.37	1.59	1.71
32	IK	1.65	1.64	2.11
33	IS	1.49	1.58	1.76
34	IT	1.49	1.62	2.14
35	PA	1.32	1.70	1.42
36	PL	2.22	2.28	2.03
37	PI	2.33	2.26	1.94
38	PY	1.80	2.23	2.40
39	PF	2.80	2.61	2.33
40	FG	1.77	1.54	1.88
41	FL	2.87	2.71	2.63
42	FP	2.70	2.57	2.58
43	FF	3.10	3.04	3.14
44	FY	3.13	2.66	3.04
45	WE	1.56	2.11	1.96
46	WW	3.60	3.42	3.25
47	YL	2.40	2.42	2.83
48	SL	1.49	1.71	1.43

[a] Taken from ref 7. [b] Taken from ref 4. The three *z*-scores were used. [c] The first three components of the extended set were used.

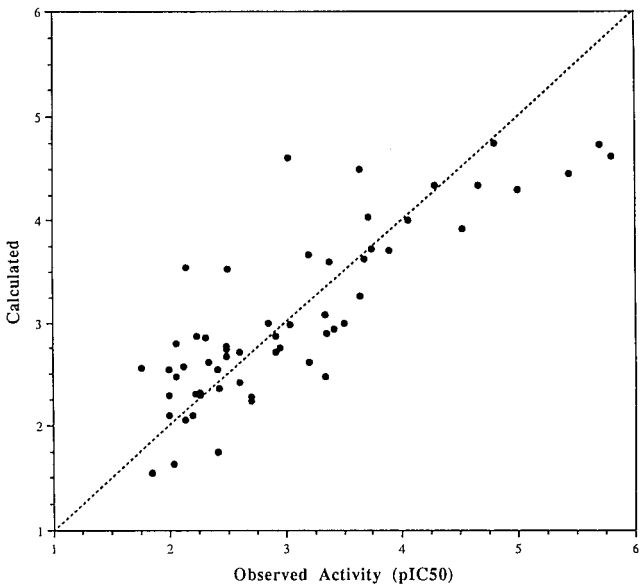

Figure 5. Plot of observed and calculated pIC50 derived from the best PLS regression model for a set of 58 ACE inhibitor dipeptides described with MSW-scores.

less than 1 (100%). This fact is fairly reflected on all the other MS-WHIM descriptors (not shown) which are, by definition, strongly dependent on eigenvalues. This may be a crucial effect as variance reflects the amount of information which PCA extracts. In this case, PCA applied to rotameric matrix to extract MSW-scores only needed three components to explain 72% of variance, while the extended MS-WHIM matrix provided just 67%. As shown in Table 7 by the respective eigenvalues, the total amount of variance is higher in the extended set than that from rotameric case.

In other words, the amount of structural information condensed by the MS-WHIM formalism is deeply coupled to the distribution of molecular surface points. Merging different surfaces for each residue so that the resulting 3D-object, even though weighted, could be a realistic picture of the allowed conformations might be theoretically correct. Practically, moving from extended to rotameric conformations the structural information (variance of coordinates) lowers, and hence the numerical differences between residues flatten instead of raising.

In summary, the comparison between extended/rotamer-library conformations highlighted two points worth stressing. (i) Different results can be obtained starting with different molecular geometries. This proves that MSW-scores are deeply affected by the molecular geometry used for their calculation. This finding is encouraging and makes MSW-scores potentially useful for investigating the "active" geometry in a series of peptides. (ii) The presented technique of summing up different surfaces (those from the rotameric collection) and applying the MS-WHIM method and score extraction seems not to be the right answer to the confor-

Table 7. Descriptive Statistics of Eigenvalues in MS-WHIM Description Matrices

	λ1_O[a]	λ2_O[a]	λ3_O[a]	λ1_P[a]	λ2_P[a]	λ3_P[a]	λ1_N[a]	λ2_N[a]	λ3_N[a]
				Extended Conformation MS-WHIM Matrix					
mean	6.58	2.74	1.85	6.72	3.02	1.48	5.47	1.58	0.98
std dev	2.48	0.47	0.23	2.20	1.30	0.52	4.22	0.22	0.31
high	12.42	4.14	2.45	11.10	6.44	2.38	13.81	2.11	1.41
low	3.56	1.89	1.16	2.40	0.46	0.31	1.85	1.21	0.52
				Rotamer-Weighted Conformation[b] MS-WHIM Matrix					
mean	1.81	0.90	0.42	2.16	0.82	0.36	1.71	0.42	0.19
std dev	0.99	0.38	0.24	1.17	0.47	0.22	0.95	0.21	0.10
high	4.45	2.19	1.23	5.22	2.44	1.06	4.18	0.87	0.40
low	0.75	0.51	0.21	1.03	0.24	0.10	0.73	0.10	0.06

[a] λ1,2,3 stand for the eigenvalues in the first, second, and third dimension according to MS-WHIM formalism. O stands for unitary field, while P and N stand for positive and negative electrostatic fields, respectively. [b] Refer to text for the construction of the rotamer-weighted multiple surfaces.

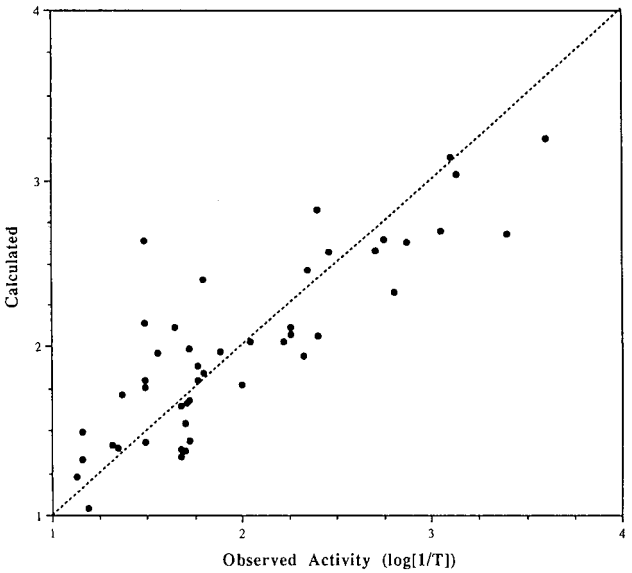

Figure 6. Plot of observed and calculated activities (log[1/*T*]) derived from the best PLS regression model for a set of 48 sweetener dipeptides described with MSW-scores.

mational problem. Thus, it gave QSAR results not better than those derived by simply using, for each peptide, its extended side-chain conformation.

We are presently investigating other approaches to the conformational problem of peptides. It is certainly possible to codify each rotamer and to set up different conformational description vector database for each peptide sequence. The only required hypothesis is that the studied peptides behave homogeneously for what concerns backbone folding or binding mode, which is not always the case for medium-sized and long sequences. We are also expanding MSW-scores to unnatural side chains in order to tackle this problem with a specific peptide set. Furthermore, it is certainly conceivable that MSW-scores, due their compactness, may be used as fingerprints for rotamer selection. Our preliminary results in this field are encouraging. It may be straightforward, in fact, selecting the optimal rotamer for a known binding site on a recursive trial-and-error basis. This possibility sheds, on the other hand, some light on one of the most severe drawbacks of the approach. As pointed out by some referees, it is often difficult, when not yet impossible, to clearly state a physical interpretation of a QSAR based on MSW-scores. In the same way, it seems hardous to assess a "friendly" physical meaning of a QSAR based on, for example, the symmetry or on the kurtosis of a MEP surface. The problem arises from the original WHIM definition and is based on the statistics of these chemometric tools. Furthermore, as back-transformation of a MS-WHIM vector in its parent conformer is not automatic and can only be afforded by optimization on a recursive trial-and-error basis,[11,12] it is often impossible to "visualize" the physical content of this description. Nevertheless, the reduced number of descriptors per residue and the ease of automation of the whole procedure make the recursive screening of large peptide libraries, coded by MSW-scores, possible to implement on any modem software.

CONCLUSIONS

In this report new 3D-descriptors for natural amino acids have been described. They have been developed by applying the MS-WHIM approach to the point coordinates of water-accessible surface calculated either in extended conformation or as a weighted average among members of Schrauber's rotamer library of the 20 natural amino acids. Principal Component Analysis of the so-obtained description matrices allowed for extraction of MSW-scores which are shown to be useful 3D-descriptors. They were used to clusterize similar residues, to relate them to classical property scales like hydrophobicity, and to obtain sound statical models for peptide QSAR studies. Multiple uses may be envisioned for such an approach including improved similarity matrices for alignment algorithms, similarity measurements for unnatural side chains, 3D-focused libraries design, and peptide fingerprint for rotamer selection. These studies are currently under development, and for some of them we are collecting encouraging results, which will be published in due course.

ACKNOWLEDGMENT

We are indebted with Dr. P. Mascagni, Dr. G. Bravi, and Dr. G. Sandrone for deep revision of the manuscript and helpful and motivating discussions.

REFERENCES AND NOTES

(1) Sneath, P. H. A. Relations Between Chemical Structure and Biological Activity in Peptides. *J. Theor. Biol.* **1966**, *12*, 157−195. Nadasdi, L.; Medzihradszky, K. A Study of the Applicability of QSAR Calculation for Peptide Hormones. *Biochem. Biophys. Res. Commun.* **1981**, *99*, 451−457. Borea, P. A.; Santo, G. P.; Salvadori, S.; Tomatis, R. Opioid Peptides. Pharmacological Activity and Lipophilic Character of Dermorphin Oligopeptides. *Farmaco Ed. Sci.* **1983**, *38*, 521−526. Asao, M.; Iwamura, H.; Akamatsu, M.; Fujita, T. Quantitative

Structure−Activity Relationships of the Bitter Thresholds of Amino Acids, Peptides and their Derivatives. *J. Med. Chem.* **1987**, *30*, 1873−1879. Fauchere, J.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Amino Acid Side Chain Parameters for Correlation Studies in Biology and Pharmacology. *Int. J. Pept. Protein Res.* **1988**, *32*, 269−278. Charton, M. The Quantitative Description of Amino Acid, Peptide and Protein Properties and Bioactivities. *Prog. Phys. Org. Chem.* **1990**, *18*, 163−284. DePriest, S. A.; Mayer, D.; Naylor, C. D.; Marshall, G. R. 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimental Determined Active Site Geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384. Waller, C. L.; Oprea, T. L.; Giolitti, A.; Marshall, G. R. Three-Dimensional QSAR of Human Immunodeficiency Virus (I) Protease Inibitors. 1. A CoMFA Study Employing Experimentally-Determined Alignment Rules. *J. Med. Chem.* **1993**, *36*, 4152−4160. Waller, C. L.; Marshall, G. R. Three-Dimensional Quantitative Structure−Activity Relationship of Angiotensin-Converting Enzyme and Thermolysin Inhibitors. II. A Comparison of CoMFA Models Incorporating Molecular Orbital Fields and Desolvation Free Energy Based on Active-Analog and Complementary-Receptor-Field Alignment. *J. Med. Chem.* **1993**, *36*, 2390−2403.

(2) Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. A Statistical Analysis of the Physical Properties of the 20 Naturally Occuring Amino Acids. *J. Protein Chem.* **1985**, *4*, 23−55.

(3) Hellberg, S.; Sjostroem, M.; Skagerberg, B.; Wold, S. Peptide Quantitative Structure−Activity Relationships, a Multivariate Approach. *J. Med. Chem.* **1987**, *30*, 1126−1135.

(4) Hellberg, S.; Sjostroem, M.; Wold, S. The Predicition of Bradykinin Potentiating Potency of Pentapeptides. An Example of a Peptide Quantitative Structure-Activity Relationship. *Acta Chem. Scand. B* **1986**, *40*, 135−140. Wold, S.; Eriksson, L.; Jonsson, J.; Sjostroem, M.; Hellberg, S.; Skagerberg, B.; Wikstroem, C. Principal Property Values for Six Non-natural Amino Acids and their Application to a Structure−Activity Relationship for Oxytocin Peptide Analogues. *Can. J. Chem.* **1987**, *65*, 1814−1820. Jonsson, J.; Eriksson, L.; Hellberg, S.; Sjostroem, M.; Wold, S. Multivariate Parametrization of 55 Coded and Non-coded Amino Acids. *Quant. Struct.-Act. Relat.* **1989**, *8*, 203−209.

(5) Cocchi, M.; Johansson, E. Amino Acids Characterization by GRID and Multivariate Data Analysis. *Quant. Struct.-Act. Relat.* **1993**, *12*, 1−8.

(6) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(7) Collantes, E. R.; Dunn III, W. J. Amino Acids Side Chain Descriptors for Quantitative Structure−Activity Relationship Studies of Peptide Analogues. *J. Med. Chem.* **1995**, *38*, 2705−2713.

(8) Fauchere, J.; Pliska, V. Hydrophobic Parameters of Amino Acid Side Chain from the Partitioning of N-Acetyl-Amino-Acid Amides. *Eur. J. Med. Chem.* **1983**, *4*, 369−375.

(9) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. Affinities of Amino Acid Side Chains for Solvent Water. *Biochemistry* **1981**, *20*, 849−855.

(10) Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D and 3D Structures. Theory. *J. Chemometrics* **1994**, *8*, 263−272.

(11) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. 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.* **1997**, *11*, 79−92.

(12) Gancia, E.; Bravi, G.; Mascagni, P.; Zaliani, A. Global 3D-QSAR Methods: MS-WHIM and Autocorrelation. *J. Comput.-Aided Drug Design* Submitted for publication.

(13) Schrauber, H.; Eisenhaber, F.; Argos, P. Rotamers: To Be or Not To Be? An Analysis of Amino Acid Side-chain Conformations in Globular Proteins. *J. Mol. Biol.* **1993**, *230*, 592−612.

(14) Ponder, J. W.; Richards, F. M. Tertiary Templates for Proteins. Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes. *J. Mol. Biol.* **1987**, *193*, 775−791.

(15) Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E. Weighted Holistic Invariant Molecular (WHIM) descriptors. Part 2. Thery Development and Application on Modeling Physico-chemical Properties of PolyAromatic Hydrocarbons. *Chemometrics and Intelligent Laboratory Systems* **1995**, *27*, 221−229.

(16) Todeschini, R.; Vighi, M.; Provenzani, R.; Finizio, A.; Gramatica, P. Modeling and Prediction by Using WHIM Descriptors in QSAR Studies: Toxicity of Heterogeneous Chemicals on *Daphnia Magna*. *Chemosphere* **1996**, *8*, 1527.

(17) Tolstrup, N.; Toftgaard, J.; Engelbrecht, J.; Brunak, S. Neural Network Model of the Genetic Code is Strongly Correlated to the GES Scale of Amino Acid Transfer Free Energies. *J. Mol. Biol.* **1994**, *243*, 816−820.

(18) Schneider, F. Die Funktion des Arginins in den Enzymes. *Naturwissenschaften* **1978**, *65*, 376−381.

(19) Gill, S. J.; Nichols, N. F.; Wadsoe, I. Calorimetric Determination of Enthalpies of Solution of Slightly Soluble Liquids. 2. Enthalpy of Solution of Some Hydrocarbons in Water and Their Use in Establishing Temperature Dependence of Their Solubilities. *J. Chem. Thermodynam.* **1976**, *8*, 445−452.

(20) Makhatadze, G. I.; Privalov, P. L. Contribution of Hydration to Protein Folding Thermodynamics I. The Enthalpy of Hydration. *J. Mol. Biol.* **1993**, *232*, 639−659.

CI980211B