New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids

Maria Sandberg,*,† Lennart Eriksson,† Jörgen Jonsson,‡ Michael Sjöström,§ and Svante Wold§

Research Group for Chemometrics, Department of Organic Chemistry, Umeå University, S-901 87 Umeå, Sweden, Umetri AB, Box 7960, S-907 19 Umeå, Sweden, and Eurona Medical AB, Box 398, S-751 06 Uppsala, Sweden

Received January 29, 1997

In this study 87 amino acids (AA.s) have been characterized by 26 physicochemical descriptor variables. These descriptor variables include experimentally determined retention values in seven thin-layer chromatography (TLC) systems, three nuclear magnetic resonance (NMR) shift variables, and 16 calculated variables, namely six semiempirical molecular orbital indices, total, polar, and nonpolar surface area, van der Waals volume of the side chain, log *P*, molecular weight, and four indicator variables describing hydrogen bond donor and acceptor properties, and side chain charge. In the present study, the data from a previous characterization of 55 AA.s from our laboratory have been extended with data for 32 additional AA.s and 14 new descriptor variables. The new 32 AA.s were selected to represent both intermediate and more extreme physicochemical properties, compared to the 20 coded AA.s. The new extended and updated principal property scales, the z-scales, were calculated and aligned to previously reported z(old)-scales. The appropriateness of the extended z-scales were validated by the use in quantitative sequence−activity modeling (QSAM) of 89 elastase substrate analogues and in a QSAM of 29 neurotensin analogues.

Introduction

Peptides and peptidomimetic compounds have attracted considerable pharmacological interest in recent years.[1−3] Following the isolation of natural and biologically active peptides, many examples have been reported that aim at the synthesis of analogues for pharmacological purposes. Since peptides composed of amino acids (AA.s) coded by mRNA have limited life spans in an organism due to their biodegradability, they are usually less suitable for therapeutic purposes. One possibility to circumvent this problem is to incorporate noncoded AA.s in the design of new peptide analogues, with possibly reduced sensitivity to biodegradation by peptidases.[4] In the case of peptidomimetics and combinatorial peptide libraries, noncoded AA.s will also provide a larger span and diversity in physicochemical properties and thereby increase the diversity of conceivable compounds.

The introduction of a noncoded AA into a peptide sequence would probably change the biological activity of the peptide. It is of great interest to be able to model and predict this change in activity. One way to accomplish this is to use quantitative sequence−activity models (QSAM.s),[5] which are crucial cases of quantitative structure−activity relationships (QSAR.s). A QSAM (or QSAR) will indicate how the change in peptide chemistry, i.e., sequence, is correlated with the change in biological performance, i.e., activity. It will also indicate how to modify the sequence to achieve improved performance. The basic assumption in QSAR is that the biological activity (BA) within a set of compounds is related to the structural variation of the compounds, i.e., the BA can be modeled as a function of molecular structure. In this context, quantitative amino acid descriptor variables have shown to be valuable.[6−10] These descriptor variables provide quantitative scales so that each AA position in a peptide sequence can be translated into the corresponding descriptor variables for the actual AA. Furthermore, by applying multivariate design—design in principal properties—the problem of introducing multiple positional AA changes and modeling the effects of such changes can be handled.[11] To be able to use statistical experimental design, a quantification of the possible modifications is needed. This can be achieved by a parametrization of the AA.s by a set of orthogonal quantitative descriptor scales. This paper addresses the derivation of such updated AA scales for 87 AA.s based on measured and theoretical AA descriptor variables.

A high-quality QSAR should be validated, be easy to interpret, and give reliable predictions of the activity of new compounds not originally present in the model.[12−14] Hence the nature of the chemical descriptor variables used and the extent to which they encode the structural features of the pertinent molecules are important for the quality of the QSAR. In principle, the descriptor variables should together contain physical and chemical information of the main types of interactions, such as lipophilicity, steric properties, hydrogen bonding, and electrostatic interactions, that might be responsible for molecular bioactivity.[15−17] Since the pioneering work of Sneath,[18] who derived amino acid descriptor variables from physicochemical semiqualitative data for the 20 coded AA.s and used them in a QSAM analysis of oxytocin−vasopressine analogues, a number of quantitative amino acid descriptor variables have been proposed for the 20 coded amino acids.[6,19−21]

* Corresponding author.
† Umetri AB.
‡ Eurona Medical AB.
§ Umeå University.

S0022-2623(97)00057-5 CCC: $15.00
Published on Web 06/04/1998

Previously, Hellberg et al. developed three AA scales, z_1-z_3, for the 20 coded AA.s, for use in peptide QSAR.s.[6] The scales, hereafter referred to as z(old), were calculated by principal component analysis (PCA) from a multiproperty matrix with 29 physicochemical variables. The three resulting principal components, so-called principal properties, are linear combinations of the primary data and were tentatively interpreted as reflecting lipophilicity (z_1(old)), steric properties (z_2(old)), and electronic properties (z_3(old)). By using only 12 physicochemical variables, Jonsson et al.[20] took a first step toward expanding these scales to encompass 35 noncoded AA.s. In this paper we present an expansion of the amino acid principal properties. Thus, we have used the same 12 descriptor variables as Jonsson et al. In addition to this, 14 supplementary theoretical descriptor variables have been applied, namely six semiempirical molecular orbital derived variables, total, polar and nonpolar molecular surface area, calculated log *P*, and four indicator variables describing hydrogen bonding donor and acceptor properties, and side chain charge (in all, 26 descriptors). According to the current knowledge, these variables will together capture lipophilic, steric, and electronic properties of the AA.s. Furthermore, 32 new AA.s have been characterized using the 26 variables eventually forming an 87 × 26 data matrix. This data matrix was analyzed to extract updated and expanded (from three to five) AA scales, denoted z-scales, for the 87 AA.s (Table 1).

Partial least squares projections to latent structures (PLS)[22,23] was used to align the extended z-scales to the z(old)-scales previously reported for the 20 coded AA.s.[6] The new z-scales can be interpreted as lipophilicity, size/polarizability, and electronic properties of the 87 AA.s.

To explore the validity and information content in the z-scales, they were used in two QSAM.s. The first QSAM consists of a model of 89 synthetic peptide substrates for the elastase enzyme. The second QSAM is based on a series of 29 neurotensin analogues. The models are interpreted in terms of how the physicochemical properties of the amino acids may be altered in the different positions in order to enhance the biological activity.

Methods

The characterized amino acids are presented in Table 1 and the used variables in Table 2. The structural formulas of the 87 amino acids are presented in Chart 1 (Supporting Information).

NMR Measurements. Seventeen of the 32 new AA.s were commercially available from Sigma and the remaining 15 (numbers 73−87 in Table 1) were synthesized and characterized (with descriptors 1−12) at our department by Larsson et al.[24−26] These latter 15 were explicitly made to show chemical properties filling gaps in the z_1, z_2, and z_3 space, thus getting AA.s with unique properties when compared to existing ones, particularly to the 20 coded AA.s.

For each amino acid, three NMR spectra were recorded, at pD 2.0, 7.0, and 12.5, and α-proton shifts were determined according to a procedure by Jonsson et al.[20] Three AA.s (numbers 47, 50, and 71) were not sufficiently soluble in D_2O, and amino acid number 52 lacked the α-proton. This means that it was not possible to determine the α-proton shifts for these four AA.s. For the NMR experiments a Bruker 250 MHz instrument was used. α-Proton shifts as well as other characterization data are presented in Table 3 (Supporting Information).

Thin Layer Chromatography (TLC). Each AA was investigated in seven TLC systems. The standardized experimental details of the chromatography are given elsewhere.[27]

Calculated Descriptor Variables. For the molecular orbital calculations, the AM1 Hamiltonian in the SPARTAN[28] molecular orbital calculation framework was used. Appropriate starting geometries of the AA.s were obtained through a molecular mechanics (force field) minimization using SYBYL, implemented in SPARTAN. Standard bond lengths and bond angles of L-amino acids from the SPARTAN database were used to construct the molecules, followed by the appropriate modifications of the side chain. The amide angles were kept frozen, and only the side chain was allowed to rotate during the geometry optimization. The reason for this restriction was that in our view the calculated properties should reflect only the structural change in the side chain. In the case of isoserine and *β*-alanine, however, the angle between the α-carbon and nitrogen was not kept constant since they lack nitrogen coupled to the α-carbon. Also for the AA.s where the α-carbon and nitrogen are part of a cyclic side chain (for example, numbers 68, 83, and 87), these atoms were not frozen during the geometry optimization. Six global descriptor variables (representing the whole molecule) were calculated: heat of formation (HOF), energy of the highest occupied molecular orbital (EHOMO), energy of the lowest unoccupied molecular orbital (ELUMO), electronegativity (EN), hardness (HA), and polarizability (POLAR). The first three variables were directly available from the SPARTAN output file. EN and HA were calculated as described by Schüürmann,[29] and polarizability (POLAR) was calculated by MOPAC.[30] After that the energetically most favored conformation was reached with the SPARTAN calculation; the following parameters were calculated using PCMODEL:[31] total surface area (Stot), total polar surface area (Spol), total nonpolar surface area (Snp). MacLogP was used for the calculation of log *P*.[32]

QSAM Data Sets. 1. Elastase Substrates. Elastase is a serine protease, which is considered to participate in the pathogenesis of some diseases, e.g., emphysema. The elastase substrates data originates from a study of 89 synthetic peptide substrates of porcine pancreatic elastase, reported by Nomizu et al.[33] The general formula of the synthetic substrate is expressed as Suc-x^1 -x^2 -Ala-pNa (Suc, succinyl; pNa, *p*-nitroanilide). The 89 peptides were modified in two positions, x^1 and x^2.

Amidolytic activity by elastase for each peptide substrate was assayed by monitoring the production rate of *p*-nitroaniline spectrophotometrically, and the kinetic parameters k_{cat} and k_{cat}/K_m were determined. K_m relates to the binding of the substrate to the enzyme and k_{cat} to the preparation of acylated enzyme and the release of *p*-nitroaniline from the enzyme. These two parameters were used as *y*-variables (responses) in the QSAM analysis. All variables were scaled to unit variance prior to the QSAM. The used sequences and activity data from Nomizu et al. are summarized in Table 4 (Supporting Information).

2. Neurotensin Analogues. Neurotensin (NT) is a tridecapeptide found to be important in the mammalian central nervous system with several effects, i.e., sedation and muscle relaxation. A set of 29 NT(8−13) peptide receptor analogues, varied in three positions (positions 8, 9, and 11), compiled from Cusack et al.,[34,35] was here used in a QSAM. The original NT-(8−13) peptide has the sequence Arg^8-Arg^9-Pro^{10}-Tyr^{11}-Ile^{12}-Leu^{13}, where the numbering originates from the native NT sequence. The binding potency at the human neurotensin receptor (hNTR) and the rat neurotensin receptor (rNTR) were evaluated as equilibrium dissociation constants (K_d values) from radioligand binding assays. The ^{10}logarithm of $(1/K_d)$ for these two receptors, hNTR and rNTR, were used as *y*-variables in the QSAM analysis. The three varied amino acid positions were described by the five z-scales. The configuration of the α-carbon of the amino acids in position 1 and 2 was assigned an indicator variable with the value of 1 if in D-configuration and a value of 0 if in L-configuration. In total 17 variables

Table 1. Descriptor Scales for the Characterized Amino Acids

no.	abbrev	name[a]	z_1	z_2	z_3	z_4	z_5
1	Ala	alanine	0.24	−2.32	0.60	−0.14	1.30
2	Arg	arginine	3.52	2.50	−3.50	1.99	−0.17
3	Asn	asparagine	3.05	1.62	1.04	−1.15	1.61
4	Asp	aspartic acid	3.98	0.93	1.93	−2.46	0.75
5	Cys	cysteine	0.84	−1.67	3.71	0.18	−2.65
6	Gln	glutamine	1.75	0.50	−1.44	−1.34	0.66
7	Glu	glutamic acid	3.11	0.26	−0.11	−3.04	−0.25
8	Gly	glycine	2.05	−4.06	0.36	−0.82	−0.38
9	His	histidine	2.47	1.95	0.26	3.90	0.09
10	Ile	isoleucine	−3.89	−1.73	−1.71	−0.84	0.26
11	Leu	leucine	−4.28	−1.30	−1.49	−0.72	0.84
12	Lys	lysine	2.29	0.89	−2.49	1.49	0.31
13	Met	methionine	−2.85	−0.22	0.47	1.94	−0.98
14	Phe	phenylalanine	−4.22	1.94	1.06	0.54	−0.62
15	Pro	proline	−1.66	0.27	1.84	0.70	2.00
16	Ser	serine	2.39	−1.07	1.15	−1.39	0.67
17	Thr	threonine	0.75	−2.18	−1.12	−1.46	−0.40
18	Trp	tryptophan	−4.36	3.94	0.59	3.44	−1.59
19	Tyr	tyrosine	−2.54	2.44	0.43	0.04	−1.47
20	Val	valine	−2.59	−2.64	−1.54	−0.85	−0.02
21	Acpa	α-aminocaprylic acid	−4.38	1.92	2.14	−2.61	−4.93
22	Aecys	*(S)-2-aminoethyl-*L*-cysteine·HCl*	3.03	2.60	0.50	2.65	−1.55
23	Afa	aminophenylacetate	−3.51	2.93	2.94	1.17	1.22
24	Aiba	α-aminoisobytyric acid	−1.33	−2.80	−0.61	−0.55	0.40
25	Aile	*alloisoleucine*	−4.09	−1.28	−1.40	−0.63	0.94
26	Alg	L-*allylglycine*	−2.31	−1.35	−0.05	0.05	1.25
27	Aba	α-aminobutyric acid	−1.22	−2.44	−0.38	−0.51	0.65
28	Aphe	*p-aminophenylalanine*	−0.62	3.28	−0.11	3.24	−1.51
29	Bal	*β-alanine*	2.16	−6.54	−4.46	−2.66	−5.93
30	Brphe	*p*-bromophenylalanine	−5.62	3.18	0.29	0.54	−1.10
31	Cha	cyclohexylalanine	−6.26	0.30	−2.58	−0.67	1.01
32	Cit	citrulline	1.31	1.47	−2.76	−2.10	0.42
33	Clala	*β*-chloroalanine	−0.66	0.30	2.65	−0.47	1.92
34	Cle	cycloleucine	−2.95	−2.16	−1.66	−0.65	0.19
35	Clphe	*p*-chlorophenylalanine	−5.31	2.66	0.99	0.02	−1.76
36	Cya	cysteic acid	4.20	3.59	3.76	−5.09	−1.36
37	Dab	2,4-diaminobutyric acid	3.69	−0.53	−0.24	1.03	−0.15
38	Dap	2,3-diaminopropionic acid	4.34	−0.54	0.96	1.04	0.24
39	Dhp	3,4-dehydroproline	−1.24	0.40	2.50	1.48	1.53
40	Dhphe	3,4-dihydroxyphenylalanine	−0.45	3.32	−0.07	−0.33	−1.95
41	Fphe	*p*-fluorophenylalanine	−4.58	2.26	1.28	−0.70	−1.58
42	Gaa	D-*glucoseaminic acid*	4.90	3.91	−1.98	−4.18	0.89
43	Hag	homoarginine	2.70	3.06	−4.15	2.32	−0.46
44	Hlys	*δ-hydroxylysine·HCl*	3.98	1.67	−2.51	0.32	0.08
45	Hnvl	DL-*β-hydroxynorvaline*	−0.85	−1.08	−1.10	−1.73	−0.04
46	Hog	*homoglutamine*	1.33	1.19	−2.14	−1.61	0.59
47	Hoph	*homophenylalanine*	−5.86	2.95	0.37	1.03	0.32
48	Hos	homoserine	0.93	−0.71	−0.01	−1.58	0.94
49	Hpr	hydroxyproline	−0.24	2.27	2.47	0.18	2.94
50	Iphe	*p-iodophenylalanine*	−6.23	6.88	3.01	1.52	1.05
51	Ise	*isoserine*	3.78	2.82	2.55	0.27	2.96
52	Mle	*α-methylleucine*	−5.40	−2.07	−2.86	−1.15	−0.27
53	Msmet	DL-*methionine-s-methylsulfoniumchloride*	1.22	1.89	−0.91	3.75	−1.25
54	1Nala	3-(1-naphthyl)alanine	−5.67	6.31	3.43	3.51	−0.47
55	2Nala	3-(2-naphthyl)alanine	−6.48	6.37	2.81	3.02	−0.49
56	Nle	norleucine (or 2-aminohexanoic acid)	−4.33	−1.30	−1.54	−0.85	0.74
57	Nmala	*N*-methylalanine	−1.30	−3.13	−0.65	0.04	−0.16
58	Nva	norvaline (or 2-aminopentanoic acid)	−3.08	−1.76	−0.98	−0.68	0.87
59	Obser	*O*-benzylserine	−5.20	2.54	−0.60	0.32	−0.48
60	Obtyr	*O*-benzyltyrosine	−7.71	7.33	−1.81	2.39	0.11
61	Oetyr	*O*-ethyltyrosine	−5.62	3.33	−0.75	0.71	−1.17
62	Omser	*O*-methylserine	−1.02	−0.30	0.36	−0.97	1.70
63	Omthr	*O*-methylthreonine	−1.75	−1.63	−1.55	−1.60	−0.20
64	Omtyr	*O*-methyltyrosine	−4.28	3.05	−0.03	0.72	−1.11
65	Orn	ornithine	3.09	0.17	−1.85	1.46	0.42
66	Pen	penicillamine	0.15	−0.76	0.42	0.67	−2.79
67	Pga	pyroglutamic acid	−3.56	2.88	2.82	1.09	3.10
68	Pip	pipecolic acid	−2.66	−2.29	−1.57	0.20	−0.39
69	Sar	sarcosine	0.30	−3.55	−0.09	0.29	−0.35
70	Tfa	*3,3,3-trifluoroalanine*	−1.47	1.11	3.66	−4.70	2.13
71	Thphe	6-*hydroxydopa*	1.29	5.13	0.89	−0.93	−2.06
72	Vig	L-*vinylglycine*	−0.81	1.17	3.54	1.20	3.43
73	Aaspa	(−)-(2*R*)-2-amino-3-(2-aminoethylsulfonyl)propanoic acid dihydrochloride	5.35	6.24	2.92	−1.44	−2.26
74	Ahdna	(2*S*)-2-amino-9-hydroxy-4,7-dioxanonanoic acid	−1.40	3.33	−2.51	−2.81	1.96
75	Ahoha	(2*S*)-2-amino-6-hydroxy-4-oxahexanoic acid	0.05	1.17	−0.74	−1.96	1.64
76	Ahsopa	(−)-(2*R*)-2-amino-3-(2-hydroxyethylsulfonyl)propanoic acid	3.01	5.82	3.85	−3.86	−1.72

Table 1 (Continued)

no.	abbrev	name[a]	z_1	z_2	z_3	z_4	z_5
77	Ahspa	(−)-(2*R*)-2-amino-3-(2-hydroxyethylsulfanyl)propanoic acid	−0.43	1.61	0.66	−0.21	−1.40
78	Ahtda	(2*S*)-2-amino-12-hydroxy-4,7,10-trioxadodecanoic acid	−2.72	5.23	−4.36	−3.62	2.07
79	Dadna	(2*S*)-2,9-diamino-4,7- dioxanonanoic acid	2.02	3.79	−3.05	0.06	0.70
80	Datda	(2*S*)-2,12-diamino-4,7,10- trioxadodecanoic acid	0.87	6.12	−4.66	−0.70	1.14
81	Dfnl	(*S*)-5,5-difluoronorleucine	−4.24	−0.29	−1.66	−2.94	1.01
82	Dfnv	(*S*)-4,4-difluoronorvaline	−3.04	0.55	0.58	−1.99	2.17
83	Dtca	(3*R*)-1-1-dioxo-[1,4]thiaziane-3-carboxylic acid	0.31	3.32	3.48	−2.87	−2.42
84	Hfnl	(*S*)-4,4,5,5,6,6,6-heptafluoronorleucine	−4.22	3.19	1.81	−8.32	−0.96
85	Pfnl	(*S*)-5,5,6,6,6-pentafluoronorleucine	−5.03	0.86	−1.61	−7.17	−0.68
86	Pfnv	(*S*)-4,4,5,5,5-pentafluoronorvaline	−3.30	2.22	2.59	−6.36	0.16
87	Tca	(3*R*)-1,4-thiazinane-3-carboxylic acid	−2.51	−0.54	0.58	1.61	−1.82

[a] Names in italics correspond to new amino acids experimentally characterized in this paper.

Table 2. Used Variables

no.	descriptor variables	abbrev
1	molecular weight (g/mol)	MW
2	TLC % migration on silica gel, ethanol/water (70/30)[a]	TL1
3	TLC, silica gel, 1-butanol/acetic acid/water (40/10/10)	TL2
4	TLC, silica gel, phenol/water (75/25)	TL3
5	TLC, silica gel, butanone/pyridine/acetic acid/water (70/15/2/15)	TL4
6	TLC, cellulose, ethanol/water (70/30)	TL5
7	TLC, cellulose, pyridine/isoamyl alcohol/water (35/30/30)	TL6
8	TLC, kiselguhr, butanone/water/phenol/acetone/ethanol (1/1)	TL7
9	side chain van der Waals volume (cm^3/mol)	vdW
10	NMR α-proton shift at pD = 2 (ppm)	NM1
11	NMR α-proton shift at pD = 7 (ppm)	NM7
12	NMR α-proton shift at pD = 12.5 (ppm)	NM12
13	^{10}log(octanol/water) partition coefficient	logP
14	energy of highest occupied molecular orbital (eV)	EHOMO
15	energy of lowest unoccupied molecular orbital (eV)	ELUMO
16	heat of formation (kcal)	HOF
17	α-polarizability (Å^3)	POLAR
18	absolute electronegativity (eV)	EN
19	absolute hardness (eV)	HA
20	total accessible molecular surface area (log Å^2)	Stot
21	polar accessible molecular surface area (log Å^2)	Spol
22	nonpolar accessible molecular surface area (log Å^2)	Snp
23	number of hydrogen bond donors	HDONR
24	number of hydrogen bond acceptors	HACCR
25	indicator of positive charge in side chain	Chpos
26	indicator of negative charge in side chain	Chneg

[a] All mobile phase compositions are given in volume parts except for variable 4 for which weight composition is given.

were used to describe each of the 29 neurotensin analogues. All variables were scaled to unit variance prior to the QSAM. The sequences for the NT analogues together with the binding potencies are presented in Table 5 (Supporting Information).

Partial Least Squares Projections to Latent Structures (PLS) and Principal Component Analysis (PCA). PLS correlates dependent variables **Y**, to a predictor matrix **X**.[22,23] PLS calculates latent variables (t_a) as linear combinations of **X** so that they well approximate **X** and well correlate with **Y**. Since PLS is a projection method, it can handle collinear data having many more variables (here sequence descriptor variables, K) than observations (here sequences, N), as long as the resulting components (*A*) are few compared to *N*. The result is a stable model of the correlation structure between **X** and **Y**. The predictor variables can be expanded by their squared and/or cross terms if desired, to account for curvature and/or interaction in the relationship. The statistical significance of the PLS model is determined by cross-validation.[36,37] The predictive validity may also be checked with a combination of cross-validation and a response permutation test.[38,39] SIMCA-P 3.0 was used for the PLS analysis.[40]

In peptide QSAMs, the structural change within a series of peptides is described by the five z-values in each varied amino acid position, which gives a peptide descriptor variable matrix **X**. The relation between the biological activity, *y*, and the peptide descriptor variable matrix **X** is then modeled by PLS. Alternatively, the structural change in each amino acid position in the peptide can be described by all 26 descriptor variables (here referred to as the whole matrix description) constituting the basis for the five z-scales.

PCA[41] summarizes one data matrix and is conceptually similar to PLS. The major difference lies in that PCA calculates latent vectors for only one data matrix (e.g., **X**). These latent vectors are the directions in space that have the largest variation, and represent the data matrix as well as possible.

Updating the z-Scales−Estimation Procedures. 1. General Considerations. In the work of Hellberg et al., three continuous scales, the z-scales, were introduced as descriptors for peptide QSAM. These scales, in the present paper referred to as z(old)-scales, have shown to work well in a number of peptide−QSAM applications. However, these scales were initially restricted to the 20 coded amino acids. In this work, we have extended and enriched the multivariate description of the amino acids. Because we want to express this new multiproperty matrix in terms of the new, updated AA scales, we decided to include also a fourth and fifth scale and assess their relevance.

There exist several alternatives for calculating the updated z-scales. We here discuss some of them and motivate why we selected one particular alternative.

2. PLS-Based Estimation. Estimation of z_1−z_3. In this approach the 20 coded AA.s are used as the training set. The **X** matrix comprises the 26 descriptor variables and the **Y** matrix contains the three z(old)-scales of Hellberg et al. PLS is then used to relate **X** and **Y**, and predicted z-scales are then obtained for the 67 noncoded AA.s. The aim of this PLS calibration is to calculate the latent variables so that they are

Table 6. SDEP[a] of Different Sets of Scales

	peptide QSAM data sets					
scale	penta[b]	angiotensin[c]	elastase[d] k_{cat}	elastase[d] k_{cat}/K_m	bitter[e]	SUMSQ rows
$z_{(1-5)}$	0.48	0.48	0.31	0.34	0.31	0.77
$z_{(1-4)}$	0.39	0.51	0.23	0.31	0.31	0.66
$z_{(1-3)}$	0.38	0.54	0.24	0.33	0.31	0.70
$2z3t_{(1-5)}$	0.56	0.73	0.33	0.38	0.32	1.20
$2z3t_{(1-4)}$	0.53	0.71	0.37	0.42	0.32	1.20
$2z3t\ _{(1-3)}$	0.78	0.72	0.39	0.41	0.31	1.54
$PP87_{(1-5)}$	0.69	1.08	0.3	0.37	0.31	1.97
$PP87_{(1-4)}$	0.63	0.53	0.25	0.29	0.31	0.92
$PP87_{(1-3)}$	0.45	0.59	0.23	0.30	0.30	0.78
whole matrix description	0.43	0.53	0.25	0.32	0.24	0.69

[a] Standard deviation of predicted value. [b] Dataset from ref 20. [c] Dataset from ref 44. [d] Dataset from ref 33. [e] Dataset from ref 20.

aligned with the z(old)-scales and thereby the interpretation of the z(old)-scales is retained. However, these three scales may be too few to account for all information included in the present **X** matrix, since both new amino acids and descriptor variables are included. Indeed, two additional scales were found significant when computed by means of PCA on the orthogonalized residual **X** matrix for all 87 AA.s. This residual matrix is the result after the information of z_1-z_3 has been removed from the original **X** matrix, for details see below. The five resulting scales are referred to as the z-scales in Table 6.

A second PLS calibration was performed with the 20 coded AA.s as training set, the 26 descriptor variables as **X** but with only z_1(old) and z_2(old) from Hellberg as Y. A subsequent PCA on the residual matrix from this PLS calibration resulted in three additional scales (referred to as the 2z3t-scales in Table 6).

Estimation of z_4 and z_5. With the extension of our data to new descriptor variables and new AA.s, we have found it of interest to extract a few additional scales for validation in peptide QSAM.s. These scales were calculated from the residual matrix after prediction of z_1-z_3. Multiple linear regression (MLR) was used to find the information in the 26 physicochemical descriptor variable matrix that is related to z_1-z_3. MLR modeling with z_1-z_3 as predictor variables (**X**) and each variable in the physicochemical descriptor variable matrix (centered and scaled to unit variance) as dependent variable **y**, were computed. This gave a residual vector, **e**, for each descriptor variable, which is orthogonal to z_1-z_3, and the coefficients, *h*, in eq 1.

$$\mathbf{e} = \mathbf{y} - (\mathbf{z_1}h_1) - (\mathbf{z_2}h_2) - (\mathbf{z_3}h_3) \qquad (1)$$

A PCA on the residual matrix **E**, formed by the 26 **e** vectors, was then calculated and the first and second principal component score vectors gave z_4 and z_5. More details are given in the Supporting Information.

3. PCA-Based Estimation. PCA of the whole set of 87 AA.s: This approach corresponds to a straightforward PCA of the 87 × 26 descriptor matrix. Here each main type of amino acid (coded/noncoded) will influence the PC model. The resulting scores are based on the largest amount of chemical and structural information, and might therefore be more “stable” than other versions of the z-scales. These five principal property scales are referred to as PP87 in Table 6.

4. Selection of One Set of Scales. In preliminary trials we have applied the different scales, in five peptide QSAM examples and compared their performance on the basis of standard error of prediction (SDEP), i.e., $((y_{\text{observed}} - y_{\text{predicted}})^2/n)^{1/2}$, for the external test set (n = number of test objects). The SDEP results for the different sets of scales are presented in Table 6. Of these approaches we finally preferred the PLS calibration versus z_1(old)-z_3(old) for several reasons: (i) It gives resulting scales for the 20 coded AA.s most consistent with previously published values,[6] which also have been extensively used in QSAM.s. (ii) The scales are clear and interpretable. (iii) They perform the best in the validation of the different sets of scales (see Table 6).

Results

Estimated z-Scales. PLS was used to align the extended z-scales to the z(old)-scales previously reported by Hellberg et al.,[6] with the 20 coded AA.s comprising the training set. The z(old)-scales were used as *y* variables and the here reported physicochemical characterization with 26 variables (k = 26) was used as x variables. The PLS computation resulted in a five-dimensional model explaining 92% of the sum of squares in the **X** matrix and 95% (84% cross-validated) of the sum of squares in the **Y** matrix. This PLS model formed the basis for the prediction of the z_1-z_3-scales for the 67 noncoded AA.s. The z_4- and z_5-scales were calculated using PCA on the residual **X** matrix following the removal of the information related to z_1-z_3. The information related to z_1-z_3 corresponded to 68% of sum of squares of the original centered and scaled **X** matrix. The additional z_4- and z_5-scales accounted for an additional 13% and 6%, respectively, of the initial sum of squares of the original **X** matrix. The correlation coefficients between the z_1(old)-z_3(old) and the new z_1-z_3-scales were 0.98, 0.95 and 0.92, respectively. In Figure 1a−c (Supporting Information) the old z are plotted against the new updated ones. The z-scales are interpreted in detail in section 4 and presented in Table 1 and in Figures 2−4.

Peptide QSAM for 89 Elastase Substrates. To validate the z-scales they were first used in a QSAM of 89 synthetic peptide substrates for porcine pancreatic elastase, reported by Nomizu et al.[33] Recall that only two AA positions were varied.

Initial modeling showed that z_5 was not of relevance, hence only z_1-z_4 were used. The best model was obtained when the *x* variables were expanded with four quadratic terms (i.e. $(z_{1pos1})^2$, $(z_{2pos1})^2$, $(z_{1pos2})^2$, $(z_{2pos2})^2$) and the five cross-terms (i.e. $(z_1 \times z_2)$ and $(z_1 \times z_4)$ of positions 1 and 2 and $(z_2 \times z_4)$ of position 1). Hence, the **X** matrix consisted of 17 *x* variables and 89 objects. PLS was applied to calculate a reference QSAM relating the peptide sequence, **X**, to the logarithm of the two biological activities log(k_{cat}) and log(k_{cat}/K_m), **Y**. This resulted in a model with four latent variables and an explained sum of squares in **Y** (R^2Y) of 0.83. The corresponding predicted sum of squares in Y (Q^2Y) according to cross-validation was 0.77. The relationship between observed and calculated activity for the reference QSAM is visualized in Figure 5a and b.

1. Model Validation. The predictive power of the elastase peptide QSAM model was further validated in the following way: A D-optimal[42] subset of 32 peptides

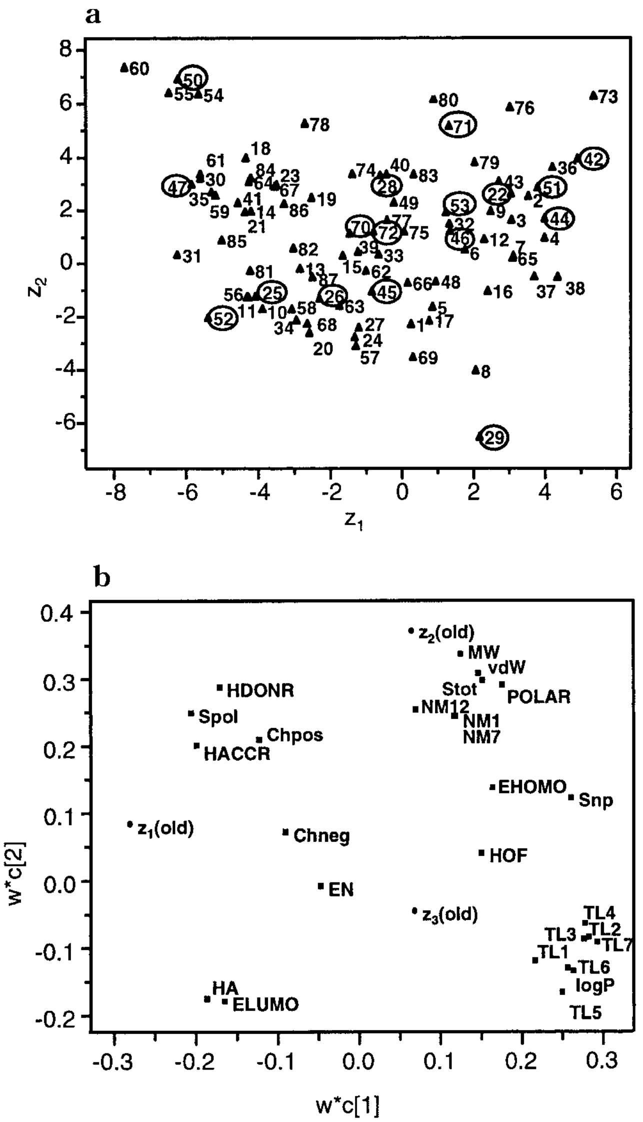

Figure 2. (a) Scatter plot of z_2-scale versus z_1-scale for the 87 amino acids. Large lipophilic AA.s can be found in the upper left quadrant. Polar, hydrophilic AA.s are situated in the upper right quadrant. New AA.s in this study are encircled. For numbering of the AA.s, see Table 1. (b) PLS weights plot of w*c[2] versus w*c[1]. The lipophilicity variables are dominating w*c[1] and steric bulk/polarizability variables contribute to w*c[2]. The 26 variables are denoted as in Table 2.

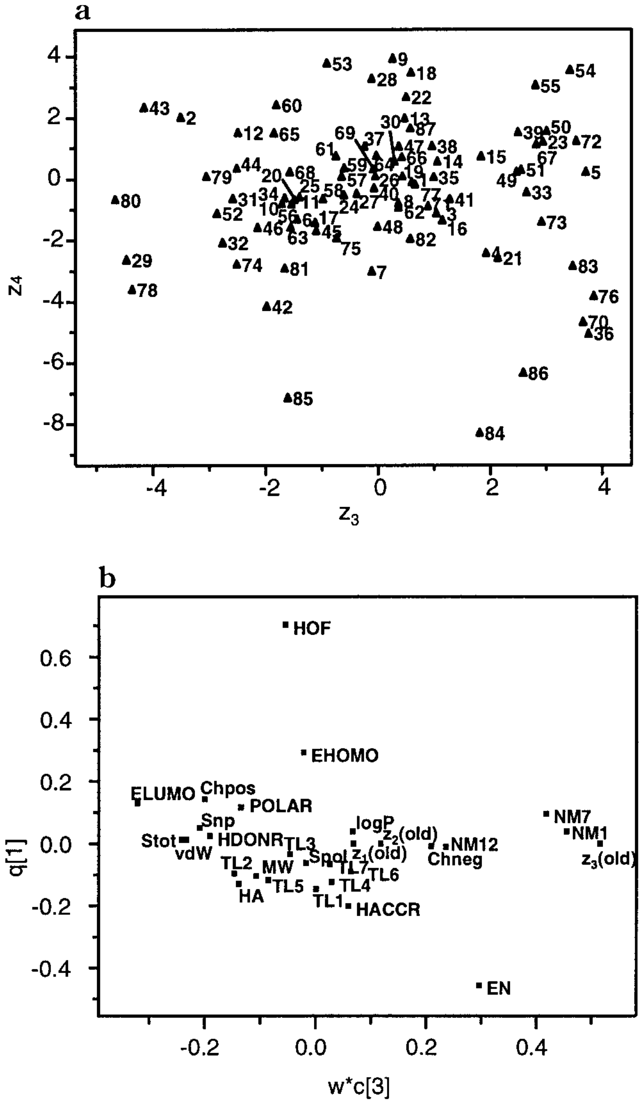

Figure 3. (a) Scatter plot of the z_4-scale versus the z_3-scale. Amino acids are marked in accordance with Table 1. (b) The first PCA loading q[1] (complementary to z_4) plotted versus the PLS weight w*c[3] (complementary to z_3). The NMR variables and ELUMO, dominate the w*c[3] while EN and HOF explore the q[1] vector. Variable abbreviations as in Table 2.

was selected as a training set (peptide numbers 1−32 in Table 4, Supporting Information). The D-optimal training set was based on a selection from the peptide space described by the z-scales and computed with the software MODDE[43] assuming a quadratic model. On the basis of modeling 32 peptides, with a resulting $R^2Y = 0.93$ and $Q^2Y = 0.78$, the biological activity of the 57 remaining peptides (validation set) was predicted. The validation set consists of peptide numbers 33−89 in Table 4 (Supporting Information). The standard error of prediction (SDEP) is a measure of the predictive power of the QSAM. For the test set these are SDEP(log(k_{cat})) = 0.23, and SDEP(log(k_{cat}/K_m)) = 0.30. These values compare well with the training set values (based on cross-validation), SDEP(log(k_{cat})) = 0.29, and SDEP(log(k_{cat}/K_m)) = 0.30, indicating a QSAM of sound predictive power. In Figure 6a and b, the observed log(k_{cat}) and log(k_{cat} /K_m) are plotted versus the predicted values for the test set and calculated values for the training set. Note the similarity with the reference model based on all 89 objects, Figure 5a and b.

Another way to investigate whether the predictive capacity of a model could be obtained by chance, is to permute the *y* values a number of times and compute a QSAR model for each permutation.[38] Twenty permuted models were computed which gave the resulting R^2 intercepts of 0.05 (k_{cat}) and 0.04 (k_{cat}/K_m), and the Q^2 intercepts were −0.04 (k_{cat}) and −0.05 (k_{cat}/K_m), see Figure 7a and b (Supporting Information). The satisfactory results from the permutation testing demonstrate that the good predictive capacity of the QSAM model of elastase substrates is not influenced by chance factors.

Interpretation of the Reference QSAM. The magnitude and sign of the PLS regression coefficients plotted in Figure 8 revealed a complex pattern with large contributions to the biological activity from linear, square and cross-terms of the z-scales. These suggest

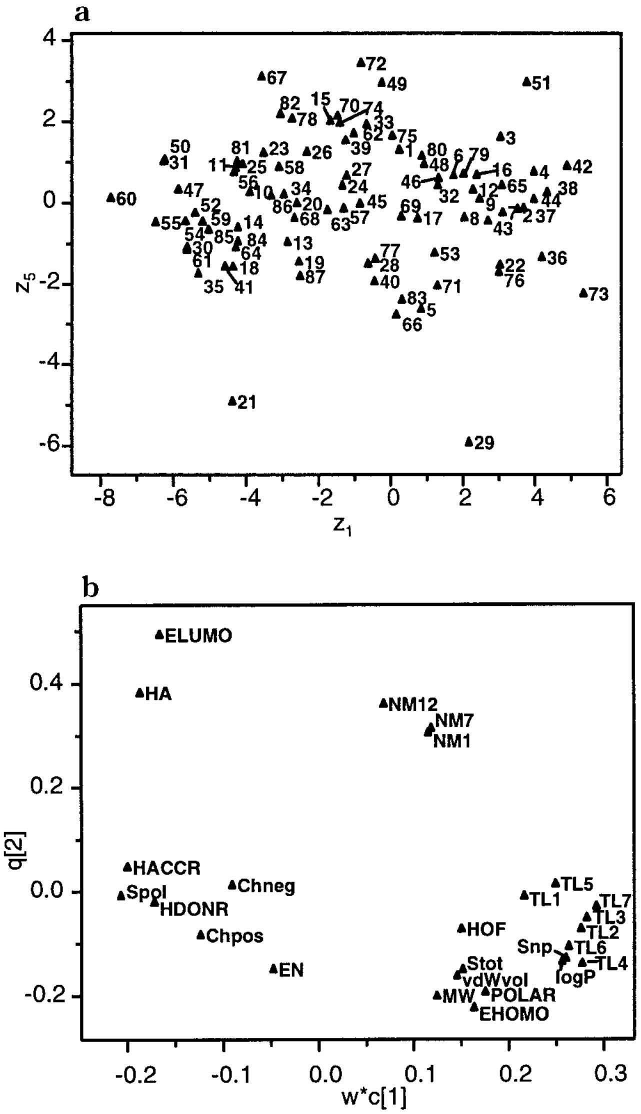

Figure 4. (a) Scatter plot of the AA z_5-scale versus the z_1-scale. The 87 AA.s are numbered as in Table 1. (b) Variable loading plot of the PCA loading q[2] (complementary to z_5) versus the first PLS weight w*c[1] (complementary to z_1). Notation as in Table 2.

how the four z-scales in both positions should be changed for a higher biological activity. However, since these values cannot be varied independently of each other, due to the discrete nature of the amino acids, we have made a compilation of feasible combinations of amino acids in the two positions and predicted their activity. Many of the studied new combinations do not fit well to the model, which is revealed by their large *x* residual standard deviations. However, Lys, Arg, or Msmet in position 1 and Vig in position 2 are interesting alternatives. Insertion of Lys-Vig combination in the QSAM gives a predicted log(k_{cat}) of 2.98 and fits fairly well into the model.

Peptide QSAM for 29 Neurotensin Analogues. In the second QSAM, 29 neurotensin analogues were modeled.[34,35] Qualitative variables describing D- or L-amino acids together with the five z-scales were used to describe the amino acid variation of the neurotensin (NT) analogues. In total 17 *x* variables were used to describe the three varied positions in the peptide sequences. The resulting five-component PLS model

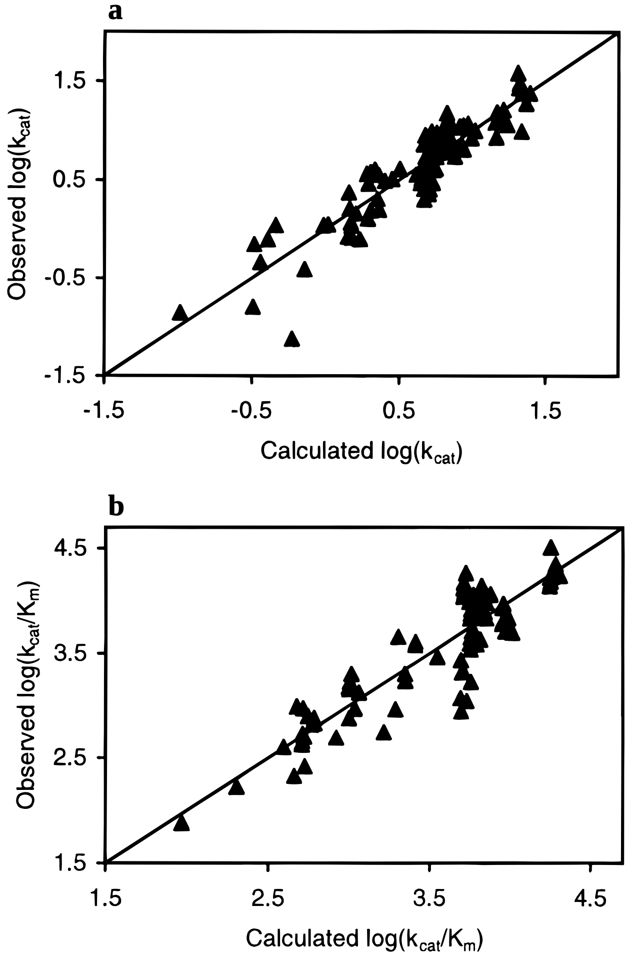

Figure 5. (a) Observed log(k_{cat}) plotted versus the corresponding calculated values from the reference QSAM of all 89 elastase substrates. (b) Observed log(K_{cat}/K_m) plotted versus the calculated values for the reference QSAM of elastase substrates.

explained 94% of the sum of squares in **Y** (78% cross-validated), using 78% of **X**. Here, the model was validated by cross-validation and permutation of **Y**, which shows significant predictive capability for the original QSAM, and Q^2 intercepts that are satisfactory low; −0.17 for the $1/K_d$(hNTR) and −0.1 for the $1/K_d$(rNTR), see Figure 9a and b (Supporting Information). In Figure 10a and b, the observed binding potencies are plotted against the corresponding calculated values for the hNTR and rNTR, respectively. We did not subdivide the data set into separate training and test sets because of scarcity of observations (peptides).

Interpretation of the QSAM of the Neurotensin Analogues. This concerns the influence of various amino acids with respect to their binding potency to the NT receptors. First, the model coefficients for hNTR in Figure 11 indicates the unfavorable influence of D-amino acids in position 9 (seven out of 29 peptides have a D-amino acid in position 9). Furthermore, high values for z_1 in position 11 and low values for z_5 in position 11 are favorable. At the same time z_5 in position 9 should be low to give high binding potency. We can now ask what kinds of AA.s correspond to these properties and how a new peptide should be synthesized to have enhanced activity. In position 11, the model indicates that a side chain with electron-donating

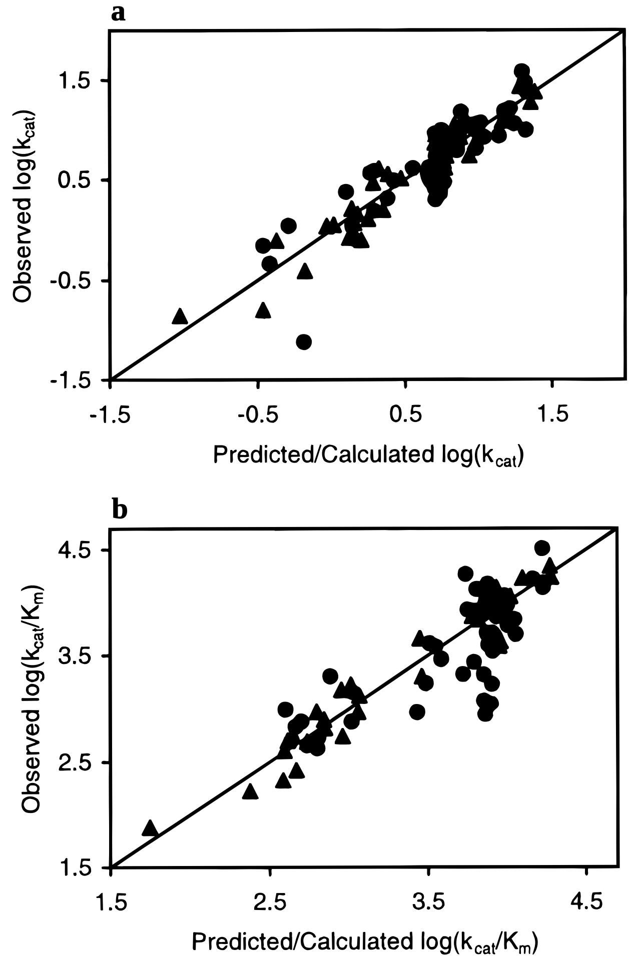

Figure 6. Validation QSAM of elastase substrates based on a D-optimal training set. (a) Relationship between the observed log(k_{cat}) values and those predicted/calculated by the model. Filled circles correspond to the test set peptides and triangles correspond to the work set peptides. (b) Observed log(k_{cat}/K_m) plotted versus the corresponding predicted/calculated values for the validation model. Filled circles correspond to the test set peptides and triangles correspond to the work set peptides.

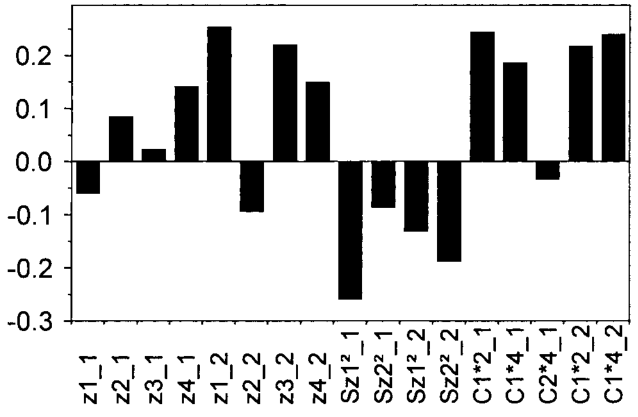

Figure 8. PLS regression coefficients of the z-scales used in the QSAM of 89 elastase substrates. The coefficients are shown for the k_{cat} variable. S stands for "squared" term and C stands for "crossed" terms. Underscore, _1 or _2, denote the position in the peptide.

atoms, such as sulfur or oxygen, is favorable. This corresponds to amino acids such as glutamic acid, cysteic acid, and serine. In position 9 an amino acid with a large side chain and/or electron-donating atoms is favorable, e.g., tyrosine and homoarginine. An aromatic or an aliphatic amino acid is favored in position 8, e.g., phenylalanine, leucine, or isoleucine. New peptides predicted to exhibit enhanced potency are, for example, [Ile^8, Hag,^{9} Sar^{11}]NT(8−13) and [Leu^8, Tyr^9, Ser^{11}]NT(8−13). These new peptides have predicted potencies for ^{10}log(1/K_d(hNTR)) of 2.26 and 3.58, respectively. Due to the unique properties for these peptides, the residuals in the QSAM are high and the exact value

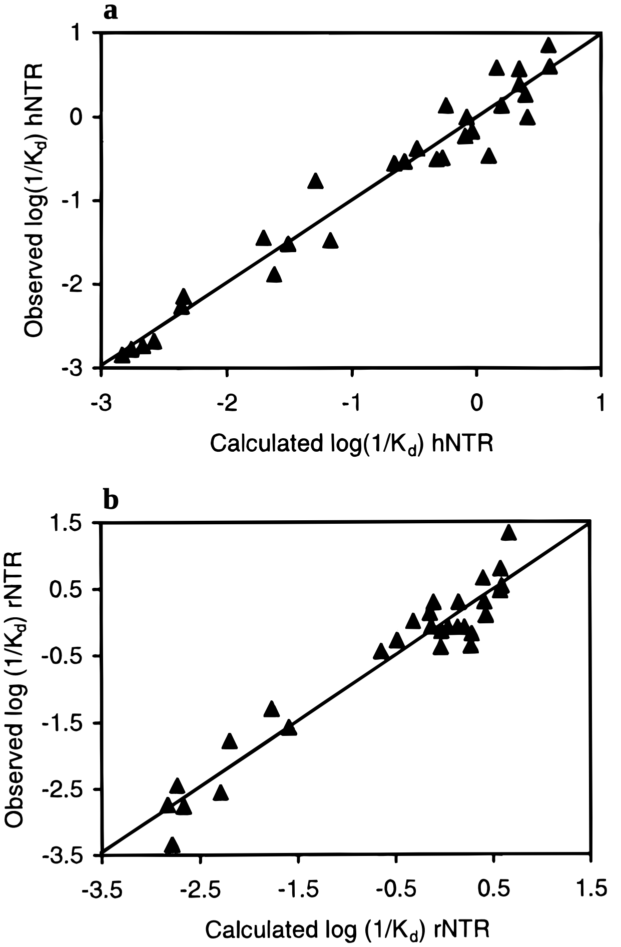

Figure 10. (a) Observed log(1/K_d(hNTR)) plotted versus the calculated values of the QSAM of 29 neurotensin analogues. (b) Observed log(1/K_d(rNTR)) plotted versus the calculated values of the neurotensin QSAM.

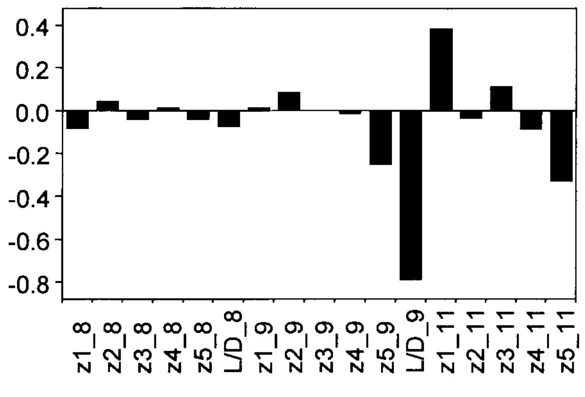

Figure 11. PLS regression coefficients of the z-scales used in the QSAM of 29 neurotensin analogues. The coefficients are shown for the hNTR-variable. S = squared term, C = crossed terms, underscore _8, _9, _11 denote the AA position in the peptide.

of the predictions should naturally be regarded with some caution. To investigate these predictions, it is recommended that the direction in AA property space indicated by the QSAM is adopted. This can be achieved by constructing a fractional factorial design or D-optimal design in the z-scales in this region.[11,14,38,44,45]

Discussion

The approach of describing each amino acid position in peptides and proteins in terms of quantitative scales goes back to the derivation of substituent scales by multivariate analysis.[46−49] We have presented here a way to calculate new scales for AA.s. These AA-scales continue the physical−organic−chemical tradition of Hammett,[50] Taft,[51] and Hansch[52] and has the same theoretical foundation in terms of similarity related modeling.[46−49] One benefit with the new z-scales is that they are interpretable in terms of physicochemical properties.

The similarities and differences in physicochemical properties between the 87 AA.s are illustrated by score plots of the z-scales (Figures 2a, 3a, and 4a). The complementary loading plots (Figures 2b, 3b and 4b) reveal information of which variables that contribute to each z-scale. In the upper left quadrant in Figure 2a, large and lipophilic amino acids are situated, e.g. number 60, *O*-benzyltyrosine. By moving to the right in this plot, the amino acids become more hydrophilic, and AA.s with large and polar side chains are situated in the upper right corner in Figure 2a. The first scale (z_1) can be interpreted as a lipophilicity scale, since the TLC variables, log *P*, and nonpolar surface area (S_{np}) have large positive loadings and polar surface area (S_{pol}) in combination with the number of proton accepting electrons in the side chain (HACCR) have negative loadings along the w*c[1] vector (see Figure 2b). Note that the z_1(old) variable has a negative loading weight in w*c[1] in Figure 2b. Hence, a large negative value of z_1 corresponds to a lipophilic amino acid, and a large positive z_1 value corresponds to a polar, hydrophilic amino acid.

AA.s with a negative z_2 value, have low molecular weight and small surface area, and are situated in the lower part of Figure 2a. The second scale (z_2) can be viewed as summarizing steric bulk/polarizability, since molecular weight (MW), van der Waals volume (vdW), total surface area (Stot) and polarizability (POLAR) have the largest contribution to z_2, see Figure 2b.

The third scale, z_3, mainly describes polarity (Figure 3a). It has negative loadings for the electrophilicity (ELUMO) and positive loadings for NMR at pD 1 and 7 and electronegativity (EN), visualized in Figure 3b.

The fourth and fifth scales, z_4 and z_5, are more difficult to interpret. They relate to such properties as electronegativity (EN), heat of formation (HOF) and electrophilicity (ELUMO), hardness (HA), and NMR at pD = 1 and 7. See loading plots in Figures 3b and 4b.

Most peptide QSAMs are dominated by changing the steric bulk and lipophilicity properties of AA.s. This is unfortunate and may result in an ignorance of possibly important electronic and polar effects of AA.s. Specific peptide−enzyme interactions often have a more polar and electronic character. Hence, we find it important to report the derivation and use of the z_4- and z_5-scales. The use of the fourth and fifth scales needs to be further studied and they should be used in the preliminary analysis of a QSAM to examine if polar and electronic effects influence the system under investigation.

An inspection of the residuals of the model in **X** after that the z_1−z_5-scales have been extracted, shows that no AA has exceptionally large residual standard deviations(data not shown). The AA.s numbers 67 (pyroglutamic acid), 71 (6-hydroxy-dopa), and 84 (heptafluoronorleucine) have, however, moderately larger residual standard deviation compared to the others. Variables contributing to the residuals of these objects are heat of formation (HOF), electronegativity (EN) and NMR at pD = 12 for AA number 67 and electrophilicity, EN for AA number 71 and HOF for AA number 84. These AA.s have somewhat unusual side chains, e.g., a trihydroxylated benzene-ring (no 71) and a heptafluorinated side chain (number 84). Thus, the analysis of the residuals provides a tool for detecting outliers and for determining which variables that are causing this behavior.

The principles described here represent a fast and reproducible methodology for quantifying physicochemical properties of AA.s. The new AA.s are well distributed within the amino acid property space, as shown in Figures 2a, 3a, and 4a, where the coded AA.s are numbers 1−20, and many of the noncoded (no 21−87) have properties which are different from the coded ones. Furthermore, both more hydrophilic (e.g., number 42), larger and more lipophilic (e.g., number 50) AA.s have been better mapped in relation to the 20 coded AA.s.

A procedure for the estimation of z-scales for new AA.s not listed in the present collection, together with necessary parameters (Tables 7−9), is described in the Supporting Information. Since some of the *x* variables that describe lipophilicity are highly correlated, some of these might be possible to omit in the estimation of the z scales (see Figure 2b). If desired, from the TLC variables TL1, TL4, and TL7 may be chosen to be measured. If many variables are systematically omitted, it might be better to perform a target rotation.[53] Here a PLS model is computed with the **X** matrix consisting of the reduced number of variables and all 87 AA.s in this collection, and z_1−z_5 as **Y**. The z-scales for the new AA.s may then be predicted by this model.

In conclusion, we here present new and extended AA z-scales for 87 AA.s, including the 20 coded AA.s and some interesting AA.s explicitly synthesized by Larsson et al.[24−26] to have side chains with unique properties. The new scales have been tentatively interpreted as quantitatively measuring lipophilicity, size and polarity of the AA side chain. We have also illustrated their validity in two peptide QSAM examples. The z-scales are also suitable for the auto-*cross*-covariance approach,[54] where peptide/protein sequences of different lengths can be compared and modeled. The scales can also be used as design variables in peptide design[11,44,45,55] and for construction of combinatorial libraries that effectively span the property space.[56,57] When a more extensive description of AA properties in a QSAM model is needed, the whole characterization matrix may be used in combination with the z-scales.

Acknowledgment. Financial support from the Swedish Natural Science Council of Research (NFR) is gratefully acknowledged. We also thank Dr. Ingemar Sethson, Patrik Boman, and Dr. Mehrdad Arshadi (Department of Organic Chemistry, Umeå University) for assistance with the NMR measurements and Stina Johansson for the TLC measurements.

Supporting Information Available: The structural formulas of the 87 amino acids, the whole physicochemical characterization matrix, procedure for estimation of z-scales for new AA.s and the corresponding parameters that are necessary for calculation of z-scales for new amino acids, the biological activity data and sequences for the elastase substrates and neurotensin analogues, correlation plots of new and old scales and permutation plots (19 pages). Ordering information is given on any current masthead page.

References

(1) Marraud, M.; Aubry, A. Crystal Structures of Peptides and Modified Peptides. *Biopolymers (Peptide Sci.)* **1996**, *40*, 45−83.

(2) Sefler, M. A.; He, J. X.; Sawyer, T. K.; Holub, K. E.; Omecinsky, D. O.; Reily, M. D.; Thanabal, V.; Akunne, H. C.; Cody, W. L. Design and Structure−Activity Relationships of C-Terminal cyclic Neurotensin Fragment Analogues. *J. Med. Chem.* **1995**, *38*, 249−257.

(3) Jirácek, J.; Yiotakis, A.; Vincent, B.; Lecoq, A.; Checler, F.; Dive, V. Development of Highly Potent and Selective Phosphinic Peptide Inhibitors of Zinc Endopeptidase 24-15 Using Combinatorial Chemistry. *J. Biol. Chem.* **1995**, *270*, 21701−21706.

(4) Blondelle, S. E.; Takahashi, E.; Weber, P. A.; Houghten, R. A. Identification of Antimicrobial Peptides by Using Combinatorial Libraries Made Up of Unnatural Amino Acids. *Antimicrob. Agents Chemother.* **1994**, *38*, 2280−2286.

(5) Jonsson, J.; Norberg, T.; Carlsson, L.; Gustavsson, C.; Wold, S. Quantitative Sequence-Activity Models (QSAM)-tools for Sequence Design. *Nucl. Acid Res.* **1993**, *21*, 733−739.

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

(7) Norinder, U. Theoretical Amino Acid Descriptors. Application to Bradykinin Potentiating Peptides. *Peptides* **1991**, *12*, 1223−1227.

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

(9) Kimura, T.; Miyashita, Y.; Funatsu, K.; Sasaki, S. Quantitative Structure−Activity Relationships of the Synthetic Substrates for Elastase Enzyme Using Nonlinear Partial Least Squares Regression. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 185−189.

(10) Norinder, U.; Karlsson, B. G.; Sjölin, L.; Pascher, T.; Bonander, N. Quantitative Structure−Property Relationships of Azurin Mutants from *Pseudomonas Aeruginosa*. *Quant. Struct.-Act. Relat.* **1996**, *15*, 475−479.

(11) Sjöström, M.; Eriksson, L. Applications of Statistical Experimental Design and PLS Modeling in QSAR. In *Chemometric Methods in Molecular Design*; Waterbeemd van de, H., Ed.; VCH Verlagsgesellshaft mbH.: Weinheim, 1995; Vol. 2, pp 63−90.

(12) Ståhle, L.; Wold, S. Multivariate Data Analysis and Experimental Design in Biomedical Research. In *Progress in Medicinal Chemistry*; Ellis, G. P., West, G. B., Eds.; Elsevier Science Publishers, B.V.: Amsterdam, 1988; pp Vol. 25, 291−338.

(13) Dunn, W. J. III; Wold, S. Pattern Recognition Techniques in Drug Design. In *Comprehensive Medicinal Chemistry The Rational Design, Mechanistic Study & Therapeutic Applications of Chemical Compounds*; Hansch, C., Sammes, P. G., Taylor, J. B., Ramsden, C., Eds.; Pergamon Press: Oxford, 1990; Vol. 4, pp 691−713.

(14) Eriksson, L.; Johansson, E. Multivariate Design and Modeling in QSAR. *Chemometr. Intell. Lab. Syst.* **1996**, *34*, 1−19.

(15) Hansch, C.; Leo, A. In *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*; Heller, S. R., Ed.; American Chemical Society: Washington, DC, 1995.

(16) Bowden, K. Electronic Effects in Drugs. In *Comprehensive Medicinal Chemistry, The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds, Quantitative Drug Design*; Hansch, C., Sammes, P., Taylor, J. B., Eds.; Pergamon Press: Oxford, 1990; Vol. 4, pp 205−239.

(17) Taylor, P. J. Hydrophobic Properties of Drugs. In *Comprehensive Medicinal Chemistry, The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds, Quantitative Drug Design*; Hansch, C., Sammes, P., Taylor, J. B., Eds.; Pergamon Press: Oxford, 1990; Vol. 4, pp 241−294.

(18) Sneath, P. H. A. Relations between Chemical Structure and Biological Activity in Peptides. *J. Theor. Biol.* **1966**, *12*, 157−195.

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

(20) Jonsson, J.; Eriksson, L.; Hellberg, S.; Sjöström, M.; Wold, S. Multivariate Parametrization of 55 Coded and Non-Coded Amino Acids. *Quant. Struct.-Act. Relat.* **1989**, *8*, 204−209.

(21) Fauchère, J.-L.; Charton, M.; Kier, L. B.; Verlooop, A.; Pliska, V. Amino Acid Side Chain Parameters for Correlation Studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **1988**, *32*, 269−278.

(22) Wold, S.; Johansson, E.; Cocchi, M. PLS−Partial Least-Squares Projections to Latent Structures. In *3D QSAR in Drug Design Theory Methods and Applications*; Kubinyi, H., Ed.; Escom Science Publishers B.V.: Leiden, 1993; pp 523−550.

(23) Wold, S. PLS for Multivariate Linear Modeling. In *Chemometric Methods in Molecular Design*; Waterbeemd van de, H., Ed.; VCH Verlagsgesellshaft mbH: Weinheim, 1995; Vol. 2, pp 195−218.

(24) Larsson, U.; Carlson, R. Synthesis of Amino Acids with Modified Principal Properties 1. Amino Acids with Fluorinated Side Chains. *Acta Chem. Scand.* **1993**, *47*, 380−390.

(25) Larsson, U.; Carlson, R. Synthesis of Amino Acids with Modified Principal properties 2: Amino Acids with Polar Side Chains. *Acta Chem. Scand.* **1994**, *48*, 511−516.

(26) Larsson, U.; Carlson, R. Amino Acids with Modified Principal Properties 3: sulfur-containing Amino Acids. *Acta Chem. Scand.* **1994**, *48*, 517−525.

(27) Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. Multivariate Parametrization of Amino Acid Properties by Thin Layer Chromatography. *Quant. Struct.-Act. Relat.* **1988**, *7*, 144−150.

(28) SPARTAN, version 4.1, Wavefunction, Inc.: Irvine, CA, 1995.

(29) Shüürmann, G. QSAR Analysis of the Acute Fish Toxicity of Organic Phosphorothionates using Theoretically Derived Molecular Descriptors. *Environ. Toxicol. Chem.* **1990**, *9*, 417−428.

(30) Stewart, J. J. P. MOPAC version 6.0, QCPE 455, Bloomington, 1991.

(31) PCMODEL, Serena software, Bloomington, 1988.

(32) MaclogP, v.1.0. BioByte Corp., Claremont, CA, 1995.

(33) Nomizu, M.; Iwaki, T.; Yamashita, T.; Inagaki, Y.; Asano, K.; Akamatsu, M.; Fujita, T. Quantitative Structure−Activity Relationship (QSAR) Study of Elastase Substrates and Inhibitors. *Int. J. Pept. Protein Res.* **1993**, *42*, 216−226.

(34) Cusack, B.; McCormick, D. J.; Pang, Y.-P.; Souder, T.; Garcia, R.; Fauq, A.; Richelson, E. Pharmacological and Biochemical Profiles of Unique Neurotensin 8-13 Analogs Exhibiting Species Selectivity, Stereoselectivity, and Superagonism. *J. Biol. Chem.* **1995**, *270*, 18359−18366.

(35) Cusack, B.; Groshan, K.; McCormick, D. J.; Pang, Y.-P.; Perry, R.; Phung, C.-T.; Souder, T.; Richelson, E. Chimeric Rat/Human Neurotensin Receptors Localize a Region of the Receptor Sensitive to Binding of a Novel, Species-Specific, Picomolar Affinity Peptide. *J. Biol. Chem.* **1996**, *271*, 15054−15059.

(36) Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Statistical Soc. Ser. B. Methodological* **1974**, *36*, 111−133.

(37) Wakeling, I. N.; Morris, J. J. A Test of Significance for Partial Least Squares Regression. *J. Chemomet.* **1993**, *7*, 291−304.

(38) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometric Methods in Molecular Design*; Waterbeemd van de, H., Ed.; VCH Verlagsgesellshaft mbH.: Weinheim, 1995; Vol. 2, pp 309−318.

(39) Eriksson, L.; Johansson, E.; Wold, S. QSAR Model Validation. In *Proceedings of the seventh international workshop on QSARs in environmental sciences*; SETAC Press: Pensacola, 1997; pp 381−397.

(40) SIMCA-P, version 3.0. Umetri AB, Umeå, 1996.

(41) Jackson, J. E. *A Users Guide to Principal Components*; Wiley: New York, 1991.

(42) Baroni, M.; Clementi, S.; Cruciani, G.; Kettaneh-Wold, N.; Wold, S. D-Optimal Designs in QSAR. *Quant. Struct.-Act. Relat.* **1993**, *12*, 225−231.

(43) MODDE, version 3.0. Umetri AB, Umeå, 1996.

(44) Hellberg, S.; Eriksson, L.; Jonsson, J.; Lindgren, F.; Sjöström, M.; Skagerberg, B.; Wold, S.; Andrews, P. Minimum Analogue Peptide Sets (MAPS) for Quantitative Structure−Activity Relationships. *Int. J. Pept. Protein Res.* **1991**, *37*, 414−424.

(45) Clementi, S.; Cruciani, G.; Baroni, M.; Costantino, G. Series Design. In *3D QSAR in Drug Design Theory Methods and Applications*; Kubinyi, H., Ed.; Escom Science Publishers B.V.: Leiden, 1993; pp 567−582.

(46) Wold, S.; Sjöström, M. Statistical Analysis of the Hammett Equation, I. Methods and Model Calculations. *Chem. Scr.* **1972**, *2*, 49−55.

(47) Wold, S. A Theoretical Foundation of Extrathermodynamic Relationships (Linear Free Energy Relationships). *Chem. Scr.* **1974**, *5*, 97−106.

(48) Wold, S.; Sjöström, M. Linear Free Energy Relationships as Tools for Investigating Chemical Similarity—Theory and Practice. In *Correlation Analysis in Chemistry*, Chapman, N. B., Shorter, J., Eds.; Plenum Press.: New York, 1978; pp 1−54.

(49) Sjöström, M.; Wold, S. Linear Free Energy Relationships. Local Empirical Rules—Or Fundamental Laws of Chemistry? *Acta Chem. Scand. B* **1981**, *35*, 537−554.

(50) Hammett, L. P. *Physical Organic Chemistry, Reaction Rates, Equilibria, and Mechanisms;* McGraw-Hill Book Co., Inc.: New York, 1940.

(51) Taft, R. W., Jr., Linear Free Energy Relationships from Rates of Esterification and Hydrolysis of Aliphatic and Ortho-substituted Benzoate Esters. *J. Am. Chem. Soc.* **1952**, *74*, 2729−2732.

(52) Hansch, C.; Fujita, T. ρ-σ-π-Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(53) Kvalheim, O. M.; Karstang, T. V. Interpretation of Latent-Variable Regression Models. *Chemometr. Intell. Lab. Syst.* **1989**, *7*, 39−51.

(54) Sjöström, M.; Rännar, S.; Wieslander, Å. Polypeptide Sequence Property Relationships in Esceria coli based on Auto Cross Covariances. *Chemometr. Intell. Lab. Syst.* **1995**, *29*, 295−305.

(55) Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wikström, C.; Wold, S. On the Design of Multipositionally Varied Test Series for Quantitative Structure−Activity Relationships. *Acta Pharm. Jugosl.* **1987**, *37*, 53−65.

(56) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Mol. Divers.* **1996**, *2*, 64−74.

(57) Lundstedt, T.; Andersson, P. M.; Clementi, S.; Cruciani, G.; Kettaneh, N.; Linusson, A.; Norden, B.; Pastor, M.; Sjöström, M.; Wold, S. Intelligent Combinatorial Libraries. In *Proceedings of the 11th European Symposium on Quantitative Structure−Activity Relationships: Computer-Assisted Lead Finding and Optimization. Current Tools for Medicinal Chemistry*. Waterbeemd van de, H., Testa, B., Folkers, G. Eds.; Wiley-VCH: Weinheim, 1997; pp 189−208.

JM9700575