accounts for 80% of the variation in partitioning data while the other factor accounts for an additional 15%. Solute size is identified as the major factor. This can be parameterized by what we term the isotropic surface area of a solute. The isotropic surface area is that area of the solute that interacts with the solvent water in a "nonspecific" manner.

The log P of a compound, which is considered a measure of its "lipophilicity", is proportional to the free energy of distribution of the solute between water and the nonpolar phase. The analysis carried out here deconvolutes this free energy into two components, a result consistent with cavity-based theoretical treatments of aqueous solubility and partitioning from an aqueous phaase into a nonpolar phase.[22,23] The driving force for this latter process is considered to be the increase in entropy associated with desolvation of the solute on transfer from the aqueous to the nonpolar phase.[6,21] The use of the isotropic surface area of a solute supermolecule to represent solute structure is consistent with this view.

The treatment presented considers solute hydration, and therefore intramolecular hydrogen bonding, explicitly. The number and positions of waters of hydration in the supermolecule are treated, at this point, as adjustable parameters. In order to generalize the approach, a suitable function must be developed that will theoretically determine the number and positions of hydration in a given solute. This is presently under study.

An advantage of this approach is that the isotropic surface area is a function of solute conformation. For the limited set of solutes on which this first report is based, the members are of limited flexibility. In order to fully understand the role of solute size on partitioning, it will be necessary to consider size as a function of conformation.

The second factor, while relatively small in its general contribution to partitioning, must also be identified in order to completely understand structural effects of the solute on partitioning behavior. Work on these aspects of the problem are under investigation.

Registry No. Methanol, 67-56-1; ethanol, 64-17-5; propanol, 71-23-8; butanol, 71-36-3; pentanol, 71-41-0; hexanol, 111-27-3; heptanol, 111-70-6; acetic acid, 64-19-7; propionic acid, 79-09-4; butyric acid, 107-92-6; hexanoic acid, 142-62-1; pentanoic acid, 109-52-4; trichloroacetic acid, 76-03-9; dichloroacetic acid, 79-43-6; chloroacetic acid, 79-11-8; methyl acetate, 79-20-9; ethyl acetate, 141-78-6; acetone, 67-64-1; ethylamine, 75-04-7; propylamine, 107-10-8; trimethylamine, 75-50-3; *n*-butylamine, 109-73-9; diethylamine, 109-89-7; pyridine, 110-86-1; aniline, 62-53-3; phenol, 108-95-2; benzoic acid, 65-85-0; benzamide, 55-21-0; 2-naphthol, 135-19-3; hydroquinone, 123-31-9; *p*-hydroxybenzaldehyde, 123-08-0; *o*-hydroxybenzoic acid, 69-72-7; *p*-hydroxybenzoic acid, 99-96-7; *o*-hydroxyanisole, 90-05-1; *p*-hydroxyanisole, 150-76-5; *o*-nitrophenol, 150-76-5; *m*-nitrophenol, 554-84-7; *p*-nitrophenol, 100-02-7; *m*-nitrobenzoic acid, 121-92-6; *o*-aminobenzoic acid, 118-92-3; *p*-aminobenzoic acid, 150-13-0; *m*-nitroaniline, 99-09-2; *o*-nitroaniline, 88-74-4; *p*-nitroaniline, 100-01-6; vanillin, 121-33-5; *o*-vanillin, 148-53-8; isovanillin, 621-59-0; isobutyl alcohol, 78-83-1; phenobarbital, 50-06-6; pentobarbital, 76-74-4; octanol, 111-87-5; ether, 60-29-7; chloroform, 67-66-3; benzene, 71-43-2; carbon tetrachloride, 56-23-5; hexane, 110-54-3; progesterone, 57-83-0; hydroxyprogesterone, 68-96-2; cortexone, 64-85-7; cortexolone, 152-58-9; cortisone, 53-06-5; cortisol, 50-23-7; testosterone, 58-22-0; pregnenolone, 145-13-1; corticosterone, 50-22-6; aldosterone, 52-39-1; hydroxypregnenolone, 12041-98-4; water, 7732-18-5.

Peptide Quantitative Structure–Activity Relationships, a Multivariate Approach

Sven Hellberg, Michael Sjöström,* Bert Skagerberg, and Svante Wold

Research Group for Chemometrics, Umeå University, S-901 87 Umeå, Sweden. *Received March 3, 1986*

The variation in amino acid sequence within sets of peptides is described by three principal properties, z_1, z_2, and z_3, per varied amino acid position. These principal properties are derived from a principal components analysis of a matrix of 29 physicochemical variables for the 20 coded (in mRNA) amino acids. The scales z_1, z_2, and z_3 are used to construct informative sets of analogues for exploring and developing quantitative structure–activity relationships (QSAR) of peptides. For the QSARs, the multivariate partial least squares (PLS) method is used. Multivariate QSARs are developed for four families of peptides, and it is shown how these QSARs can predict the activity of new peptide analogues.

Peptides are of central importance in all living systems. Hence, they may be considered to be the drugs of the future. In drug development, quantitative structure–activity relationships (QSARs) are essential to optimize the structure to give desired biological activities. Here we present a strategy for developing peptide QSAR.

The quantitative description of amino acids is crucial for QSARs of peptides. In a pioneering work Sneath[1] derived amino acid descriptors from qualitative (interval) data for the 20 coded amino acids. In a recent paper[2] we extended the multivariate approach of Sneath to continuous amino acid properties. The scales derived from this matrix are relevant in peptide QSAR.[3] Here we have further expanded the property matrix by including nine HPLC measurements of dansylated amino acids at different pH and eluent mixtures[4] (see Table I). The new multiproperty matrix (available as supplementary mate-

(1) Sneath, P. H. A. *J. Theoret. Biol.* **1966**, *12*, 157.
(2) Sjöström, M.; Wold, S. *J. Mol. Evol.* **1985**, *22*, 272.
(3) Hellberg, S.; Sjöström, M.; Wold, S. *Acta Chem. Scand., Ser. B* **1986**, *40*, 135.
(4) Skagerberg, B.; Sjöström, M.; Wold, S., manuscript in preparation.
(5) *The Merck Index*, 9th ed., 1977.
(6) *Handbook of Biochemistry*; CRC: Baca Raton, FL, 1968.
(7) Seydel, J. K.; Schaper, K.-J. *Chemische Struktur und biologische Aktivität von Wirkstoffen*; Verlag Chemie: Weinheim, 1979.
(8) Roberts, G. C. K.; Jardetzky, O. *Adv. Protein Chem.* **1970**, *24*, 447.
(9) Horsley, W.; Sternlicht, H.; Cohen, J. S. *J. Am. Chem. Soc.* **1970**, *92*, 680.
(10) Rosenthal, S. N.; Fendler, J. H. *Adv. Phys. Org. Chem.* **1976**, *13*, 279.
(11) Aboderin, A. A. *Int. J. Biochem.* **1971**, *2*, 537.
(12) Woese, C. R.; Drugre, D. H.; Saxinger, S. A. *Proc. Natl. Acad. Sci. U.S.A.* **1966**, *55*, 966.
(13) Jones, D. D. *J. Theor. Biol.* **1975**, *50*, 167.
(14) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. *Biochemistry* **1981**, *20*, 849.

0022-2623/87/1830-1126$01.50/0

Table I. Variables Used To Characterize the Amino Acids

variable no.	ref	property
1		molecular weight
2	5	pK_{COOH} (COOH on C_α)
3	5	pK_{NH_2} (NH_2 on C_α)
4	6	pI, pH at the isoelectric point
5	7	substituent van der Waals volume
6	8	^{1}H NMR for C_α-H (cation)
7	8	^{1}H NMR for C_α-H (dipolar)
8	8	^{1}H NMR for C_α-H (anion)
9	9, 10	^{13}C NMR for C=O
10	9, 10	^{13}C NMR for C_α-H
11	9, 10	^{13}C NMR for C=O in tetrapeptide
12	9, 10	^{13}C NMR for C_α-H in tetrapeptide
13	11	R_f for 1-N-(4-nitrobenzofurazono)amino acids in ethyl acetate/pyridine/water
14	12	slope of plot $1/(R_f - 1)$ vs. mol % H_2O in paper chromatography
15	13	dG of transfer of amino acids from organic solvent to water
16	14	hydration potential or free energy of transfer from vapor phase to water
17	15	R_f, salt chromatography
18	16	log P, partition coefficient for amino acids in octanol/water
19	17	log D, partition coefficient at pH 7.1 for acetylamide derivatives of amino acids in octanol water
20	18	dG = RT ln f; f = fraction buried/accessible amino acids in 22 proteins
21–29	4	HPLC retention times for nine combinations of three different pH and three eluent mixtures

Table II. Descriptor Scales z_1, z_2, and z_3 for Amino Acids[a]

amino acid	z_1	z_2	z_3
Ala (A)	0.07	-1.73	0.09
Val (V)	-2.69	-2.53	-1.29
Leu (L)	-4.19	-1.03	-0.98
Ile (I)	-4.44	-1.68	-1.03
Pro (P)	-1.22	0.88	2.23
Phe (F)	-4.92	1.30	0.45
Trp (W)	-4.75	3.65	0.85
Met (M)	-2.49	-0.27	-0.41
Lys (K)	2.84	1.41	-3.14
Arg (R)	2.88	2.52	-3.44
His (H)	2.41	1.74	1.11
Gly (G)	2.23	-5.36	0.30
Ser (S)	1.96	-1.63	0.57
Thr (T)	0.92	-2.09	-1.40
Cys (C)	0.71	-0.97	4.13
Tyr (Y)	-1.39	2.32	0.01
Asn (N)	3.22	1.45	0.84
Gln (Q)	2.18	0.53	-1.14
Asp (D)	3.64	1.13	2.36
Glu (E)	3.08	0.39	-0.07

[a] The first three score vectors of a principal component analysis of the amino acid data.

rial) was extracted by principal components analysis (PCA) to give three scales z_1, z_2, and z_3 (see Table II). We call these "principal properties" of the amino acids and tentatively interpret them as related to hydrophilicity (z_1), bulk (z_2), and electronic properties (z_3). Cramer[19] has in a similar way developed chemical descriptor scales (BCDEF) for common organic compounds.

Prior to the introduction of the solid-phase technique, the synthesis of peptides was a severely limiting factor in peptide research. The development of automated peptide synthesis has now made it possible to synthesize a large number of analogues to an interesting "lead" peptide. The major problem is no longer *to make* peptides, but rather *which* peptides to prepare. For example, if a "lead" peptide is varied in four amino acid positions, it is possible to synthesize 160 000 different analogues using only the 20 coded amino acids. Hence, given the time and economical constraint to make only a certain number of analogues, it is important to change the structure of the "lead" peptide according to an informationally optimal scheme. Such a scheme should allow the capture of as much information as possible about which chemical properties (and in which combinations) that are important for the biological effects, i.e., the construction of a QSAR. We note that the common practice to change one position at a time gives a set of analogues that contains *the least* possible amount of information. By applying simple principles of statistical design, we demonstrate that more informative sets can easily be constructed.

When the biological activities of the synthesized peptides have been measured, multivariate data analytic methods are used to model the relation between the structural modifications of the peptides and the biological measurements. The resulting QSAR can predict new more potent and selective analogues in the given family of peptides. We here demonstrate the use of the recently developed PLS model[20,21] for relating multivariate descriptor data (**X**) to uni- or multivariate biological activity data (**Y**). The PLS model also applies when the number of descriptors is larger than the number of analogues.

Peptide QSAR

The development of a QSAR for a series of peptides can, like other QSAR, be divided into five steps.

(1) The description of the change in chemical structure within the series (here of peptide analogues). (2) The selection (design) of a series of analogues to be synthesized and tested. (3) The synthesis and the biological testing of the analogues (peptides). (4) The construction of a model that relates the change in chemical structure to the change in biological activity in the series. (5) The postulation of new and possibly more active and selective analogues to be synthesized and tested.

Step 1: Structural Description. a. Qualitative and Semiquantitative Structural Description. The traditional approach to the structural description of peptides regards each amino acid as having unique qualitative features, e.g., an aromatic ring, an amide group, a hydroxyl group, a sulfhydryl group, β-branching, ionized or nonionized, and hydrophobic side chains. The structure–activity relationship is then discussed in terms of how the biological activity changes when certain features are introduced or deleted in various positions in the peptides.[22]

Molecular modelling gives the investigator a visual representation of the structure of several amino acid positions simultaneously. This approach may be relevant for finding "lead" compounds, but probably not for the lead optimization.[23]

(15) Weber, A. L.; Lacey, J. C., Jr. *J. Mol. Evol.* **1978**, *11*, 199.
(16) Pliška, V.; Schmidt, M.; Fauchère, J.-L. *J. Chromatogr.* **1981**, *216*, 79.
(17) Fauchère, J.-L.; Pliška, V. *Eur. J. Med. Chem.* **1983**, *18*, 369.
(18) Janin, J. *Nature (London)* **1979**, *277*, 491.
(19) (a) Cramer, R. D., III *J. Am. Chem. Soc.* **1980**, *102*, 1837. (b) Cramer, R. D., III *J. Am. Chem. Soc.* **1980**, *102*, 1849.
(20) Wold, H. In *Systems under Indirect Observation*; Jöreskog, K. G., Wold, H., Eds.; North Holland: Amsterdam, 1982; Part II, pp 1–54.
(21) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735.
(22) Sawyer, W. H.; Manning, M. *Annu. Rev. Pharmacol.* **1973**, *13*, 5.
(23) Marshall, G. R. In *Drug Design: Fact or Fantasy?*; Jolles, G., Wooldridge, K. R. H., Eds.; Academic: London, 1984; pp 35–46.

The traditional qualitative approach has been semiquantified by Sneath,[1] who assembled a matrix consisting mainly of qualitative variables related to the presence or absence of functional groups and various other features of the coded amino acids. By multivariate techniques he derived a similarity index, a dissimilarity index, and four scales (v_I–v_{IV}), which he thereafter related to the biological activity of peptides.

A similar approach was used by Simon[24] who used 10 indicator variables for the presence or absence of functional groups and features relating to intermolecular forces. From these variables he derived a dissimilarity index by comparing all peptides to the most active peptide in the set. This index he then related to the biological activity of the peptides. However, this approach has the drawback that it cannot be used to predict peptides with higher activity than the most active already included in the set.

Additivity schemes such as the Free–Wilson/Fujita–Ban methods[25] have been used to relate the change in amino acid composition of peptides to the biological activity.[26] That approach does not allow predictions to be made about peptide activity for peptides substituted with amino acids not already incorporated in the set.

b. Quantitative Description. Univariate quantitative description of amino acid properties have been used by, for example, Burton[27] and Borea et al.,[28] who used π (lipophilicity of the side chain of the amino acid) as a descriptor of peptide structure and regression analysis to model the relationship to the biological activity. A multivariate quantitative approach using multiple regression analysis (MRA) has been used by Nadasdi et al.,[29] who used 13–15 analogues and four to six descriptors, Charton,[30] who used 6–13 analogues and seven descriptors, and Fauchere et al.,[31] who used seven analogues and five descriptors. However, these authors studied peptide sets with only one varying amino acid position.

c. Quantitative, Multivariate, and Multipositional Description. In our approach the variation of the chemical structure of the individual amino acids is quantitatively reflected in a multitude of different chemical measurements. For practical reasons this multitude of data is then contracted by a statistical analysis to give three scales, z_1, z_2, and z_3. The scales are then used for the structural description of peptide analogues and also, as described in "Step 2: Design", to construct test series of analogues. We presently investigate how much information that may be lost in this contraction. For peptide sets with up to four varying positions, a full description with 29 variables per position is certainly feasible.

I. Derivation of Scales z_1, z_2, and z_3. Each of the 20 coded amino acids was described by 29 measures of various properties (see Table I). The variables were scaled to unit variance, except for the nine HPLC variables, which was scaled to variance 0.33. In this way the variables are given the same possibility to influence the statistical analysis and the nine HPLC variables will not dominate. The PCA of the scaled data gave three components z_1, z_2, and z_3, significant according to cross validation[32] (Table II). The three z values can be regarded as "principal properties" of the amino acids summarizing all 29 measurements. As seen from the loadings (Table III) of the PCA, the first component, z_1, is mainly related to hydrophilicity, z_2 is additionally influenced by the size, ^{1}H NMR, and some hydrophobicity/hydrophilicity scales, while z_3 contains information from the pK_a, pI, and ^{1}H NMR variables. We are presently working on an extension of the data matrix by including noncoded amino acids and variables such as ^{1}H and ^{13}C NMR data, IR data, and TLC data recorded at different pH and solvent mixtures.

Table III. Loadings (p_{ak}) from PCA of the Descriptor Matrix for Amino Acids[a]

variable (k)	p_{1k}	p_{2k}	p_{3k}
1	−0.09	0.41	−0.10
2	−0.20	−0.10	−0.20
3	−0.06	−0.16	0.35
4	0.02	0.09	−0.34
5	−0.10	0.37	−0.25
6	−0.01	0.33	0.36
7	0.01	0.34	0.36
8	0.02	0.26	0.41
9	−0.06	0.17	−0.13
10	−0.13	0.11	−0.03
11	−0.11	0.25	−0.21
12	−0.13	0.10	−0.05
13	−0.31	−0.05	−0.02
14	0.28	0.07	0.14
15	−0.28	0.16	0.04
16	−0.21	−0.31	0.12
17	0.23	−0.24	−0.08
18	−0.31	−0.07	0.09
19	−0.31	−0.04	0.15
20	−0.22	−0.22	0.26
21	−0.19	0.01	−0.04
22	−0.18	0.00	−0.01
23	−0.19	−0.02	−0.01
24	−0.18	0.00	−0.04
25	−0.18	−0.03	−0.02
26	−0.18	−0.03	0.00
27	−0.18	−0.01	−0.05
28	−0.18	−0.02	−0.03
29	−0.18	−0.01	−0.05

[a] The loadings reflect the relative contribution of each variable (k) to the three z values.

II. Peptide Description. For a set of peptide analogues, the chemical structure can now be quantified by describing each varied amino acid position with the three z values. Thus, a set of peptide analogues varied in m positions is described by $3m$ variables.

Other possibly relevant descriptors may also be added, e.g., squared terms, cross terms, or descriptors of properties of the whole peptide, such as R_f values from TLC (see example III in the Results section).

Step 2: Design. The selection of the peptide analogues (design) in a series is of the same importance as the selection of compounds in other QSAR.[33] A bad design will

(24) Simon, Z. *Rev. Roum. Biochim.* **1968**, *5*, 319.
(25) (a) Free, S. M.; Wilson, J. W. *J. Med. Chem.* **1964**, *7*, 395. (b) Fujita, T.; Ban, T. *J. Med. Chem.* **1971**, *14*, 148.
(26) Schaper, K.-J. *Eur. J. Med. Chem.* **1980**, *15*, 449.
(27) Burton, J. In *Peptides 1982*; Walter de Gruyter: Berlin, 1983; pp 629–633.
(28) Borea, P. A.; Sarto, G. P.; Salvadori, S.; Tomatis, R. *Farmaco, Ed. Sci.* **1983**, *38*, 521.
(29) Nadasdi, L.; Medzihradszky, K. *Peptides* **1983**, *4*, 137.
(30) Charton, M. In *QSAR and Strategies in the Design of Bioactive Compounds*; Seydel, J. K., Ed.; VCH Verlagsgesellschaft: Weinheim, 1985; pp 260–263.
(31) Fauchère, J.-L.; Lauterwein, J. *Quant. Struct.-Act. Relat.* **1985**, *4*, 11.
(32) Wold, S.; Albano, C.; Dunn, W. J., III; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjöström, M. In *Chemometrics–Mathemathics and Statistics in Chemistry, NATO ASI Series C No. 138*; Kowalski, B. R., Ed.; Reidel: Dordrecht, 1984; pp 17–95.
(33) (a) Hansch, C.; Unger, S. *J. Med. Chem.* **1973**, *16*, 1217. (b) Wootton, R.; Cranfield, R.; Sheppey, G. C.; Goodford, P. J. *J. Med. Chem.* **1975**, *18*, 607. (c) Austel, V. *Eur. J. Med. Chem.* **1982**, *17*, 9. (d) Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wikström, C.; Wold, S., accepted for publication in *Acta Pharm. Jugosl.*

Table IV. Number of Peptide Analogues in Test Series Constructed by Fractional Factorials When Each Varied Position Is Described by Three z Values

no. of varied positions	minimum no. of analogues
1	4
2	8
3–5	16
6–10	32
11–21	64

give data containing no or only little information concerning the structure–activity relationship, whereas a good design will give data containing much information and may result in a successful QSAR.

However, the design problem is generally overlooked. Thus the peptide sets with literature data used as examples in this article are not constructed by any design (to our knowledge).

The intuitive way to select a set of peptide analogues is to change one amino acid position at a time.[34] This "design", or rather lack of design, is *inefficient*. This is because *the resulting data will not contain any information about the joint influence of the substituted positions on the peptide activity*. This inefficiency of "one feature at a time" designs is well-known in chemical engineering and statistics[35] but seems to be unrecognized in peptide chemistry.

Instead all positions of interest in the peptide should be varied simultaneously over all principal properties of the amino acids. One way of making such a design, introduced in QSAR by Austel,[33c] is by a *fractional factorial design*,[35] with 2^{q-r} analogues. Here $q = mj$, where m is the number of varied positions and j the number of descriptors of each amino acid and r the reduction factor. This factor r must be chosen so that 2^{q-r} is larger than q. With these designs, deviations from additivity can be detected and interactions between different positions can be estimated. Even if the substitution at different positions has just an additive influence on the activity, factorial designs give data that can model the structure–activity relationship with higher precision.[35] Thus, the intuitive "one position at a time" design has the disadvantage to give data with less information and "consume" more peptide analogues than a proper design based on, for instance, fractional factorials.

Our approach to the design problem is based on the result that each varied amino acid position, in a series of peptide analogues, can be approximately characterized by three principal properties, z_1, z_2, and z_3 (Table II). In the case when, for example, four positions are varied ($m = 4$), and each is described by three z values ($j = 3$), a *full* factorial 2^q design would require $2^{4\cdot3} = 4096$ peptides to be synthesized and tested. With a *fractional* factorial 2^{q-r} design it is possible to obtain information concerning the main factors in the QSAR using only $2^{(4\cdot3)-8} = 16$ peptides. Table IV shows the minimum number of analogues in sets of peptides with 1–21 varied positions when using fractional factorial designs for the selection of analogues. We note that a standard factorial design assumes that each variable can be set precisely to the level corresponding to plus (+) and minus (−). In the present case this is not possible and we use the design only as a tool to find the combination of amino acids that together span the chemical property space as well as possible.

Design Example. As an example of a fractional factorial design we constructed a test series for oxytocin/vasopressin analogues varied in four positions (positions 2, 3, 4, and 8). A smaller example of a design that includes modelling as well is found in ref 33d.

The $2^{(4\cdot3)-8}$ fractional factorial design matrix, Table V, is generated according to standard rules[35] as follows.

(a) A full factorial design is constructed for the four columns with A, B, C, and D. These columns are assigned to the z_1 values of the four varied positions.

(b) Four additional columns are constructed from A, B, C, and D by multiplying the signs of three of these together and then assigning them to z_2 for positions 2, 3, 4, and 8. Thus z_2 in position 2 is ABC, z_2 in position 3 is BCD, z_2 in position 4 is ACD, and z_2 in position 8 is ABD.

(c) Four additional columns (z_3 columns) were then generated in a similar way by multiplying the signs of two z_1 columns, i.e., z_3 in position 2 is AB, z_3 in position 3 is BC, z_3 in position 4 is AC, and z_3 in position 8 is BD.

This design matrix is used for constructing the test series. For each position amino acids with the corresponding signs of the z values are chosen. However, as seen from Table II, most of the coded amino acids have small z_3 values. The only exceptions are Cys, Pro, and the amino acids with both z_1 and z_2 positive. Hence, for some combinations of signs of the three z values it is impossible to find representative coded amino acids. The amino acids Cys and Pro have unique properties that may make them less suitable to be incorporated in test series aimed to study continuous relationships. We have therefore chosen to primarily span the z_1/z_2 plane and span z_3 for the amino acids with positive z_1 and z_2. This is indicated in the design matrix (Table V) by the parentheses in the z_3 columns. For each combination of signs the amino acid with the highest absolute z values is chosen to be included in the test series.

Table VI shows the set of 16 oxytocin analogues resulting from the design. This set is one of the most informative that can be selected from the 16 000 possible peptide analogues.

Designed test series can also be used in the search for new and more selective lead compounds with, for example, D-amino acids. If the presence of a D-amino acid in a certain position leads to a dramatic change in, say, selectivity, a new test series should then be constructed, centered at this analogue.

Step 3: Synthesis and Biological Testing. The synthesis and biological testing of the analogues is, of course, crucial to a QSAR study but beyond the scope of the present paper. However, we note that information can be gained if the biological activity is measured in several ways; i.e., Y is multivariate. For example, it is customary to observe the biological tests at different dose levels. Unfortunately, most of these data are not used in the data analysis correlating structure with activity. In many cases it would be better to use the raw data from the dose–response measurements than the reported values, e.g., ED50 values, which have been estimated from dose–response models which do not always describe the test data well.

For QSAR studies it is an advantage if all the compounds in the studied set are tested in the same laboratory and approximately at the same point of time. Data collected from several different laboratories often contain some interlaboratory variation. If the biological tests are performed several years apart from each other, it is also possible that some unknown systematic difference can be introduced. Indeed this is reflected in the present peptide QSARs. The best QSARs are obtained for the examples

(34) (a) Rudinger, J. In *Drug Design*; Ariens, E. J., Ed.; Academic: New York, 1971; pp 319–419. (b) Farmer, P. S. In *Drug Design*; Ariens, E. J., Ed.; Academic: New York, 1980; pp 119.

(35) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; Wiley: New York, 1978.

Table V. Fractional Factorial Design Matrix for Peptide Analogues Varied in Four Positions

	z_1, z_2, z_3											
	position 2			position 3			position 4			position 8		
no.	A	ABC	AB	B	BCD	BC	C	ACD	AC	D	ABD	BD
1	−	−	(+)	−	−	(+)	−	−	(+)	−	−	(+)
2	+	+	−	−	−	(+)	−	+	(−)	−	+	(+)
3	−	+	(−)	+	+	−	−	−	(+)	−	+	(−)
4	+	−	(+)	+	+	−	−	+	(−)	−	−	(−)
5	−	+	(+)	−	+	(−)	+	+	−	−	−	(+)
6	+	−	(−)	−	+	(−)	+	−	(+)	−	+	(+)
7	−	−	(−)	+	−	(+)	+	+	−	−	+	(−)
8	+	+	+	+	−	(+)	+	−	(+)	−	−	(−)
9	−	−	(+)	−	+	(+)	−	+	(+)	+	+	−
10	+	+	−	−	+	(+)	−	−	(−)	+	−	(−)
11	−	+	(−)	+	−	(−)	−	+	(+)	+	−	(+)
12	+	−	(+)	+	−	(−)	−	−	(−)	+	+	+
13	−	+	(+)	−	−	(−)	+	−	(−)	+	+	−
14	+	−	(−)	−	−	(−)	+	+	+	+	−	(−)
15	−	−	(−)	+	+	+	+	−	(−)	+	−	(+)
16	+	+	+	+	+	+	+	+	+	+	+	+

Table VI. Example of a Test Series of Oxytocin Analogues Varied in Four Positions Constructed Using Fractional Factorial Design

no.	position 2	position 3	position 4	position 8
1	Ile	Ile	Ile	Ile
2	Arg	Ile	Trp	Trp
3	Trp	Arg	Ile	Trp
4	Gly	Arg	Trp	Ile
5	Trp	Trp	Arg	Ile
6	Gly	Trp	Gly	Trp
7	Ile	Gly	Arg	Trp
8	Asp	Gly	Gly	Ile
9	Ile	Trp	Trp	Arg
10	Arg	Trp	Ile	Gly
11	Trp	Gly	Trp	Gly
12	Gly	Gly	Ile	Asp
13	Trp	Ile	Gly	Arg
14	Gly	Ile	Asp	Gly
15	Ile	Asp	Gly	Gly
16	Asp	Asp	Asp	Asp

III and IV where all the compounds in each set are measured in the same laboratory; furthermore, a systematic difference is indicated between the two sets in example IV that are tested 4 years apart. For the oxytocin analogues (example I), less precise QSARs are obtained. This was expected since the biological data were collected from several different publications.

Step 4: Mathematical Modelling. The QSARs in this study have been modelled by using the PLS method.[20,21] With PLS it is possible to relate multivariate descriptor data to uni- or multivariate activity data. It is also possible to use more descriptors and activities than compounds, if this is desired. Of course, irrelevant variables introduce some noise in the models. However, PLS models are less sensitive to this noise than other regression methods.[20,21] With the PLS method one also keeps the risk for spurious correlations under control, which otherwise is a problem when many variables are used to describe a limited set of objects.[36]

The PLS Method. In the first phase of the data analysis, data from compounds with known biological activity (a training set) are used to construct a model that connects the variation in chemical structure to the variation in biological activity. In the second phase, as discussed below in "Step 5", this model is used to predict how the structure should be modified to improve the biological activity. If sufficiently many compounds are available, a test set, consisting of analogues not used in the model development, can be used to test the predictive capability of the model.

a. Scaling. Usually, if no prior knowledge is present, the variables are scaled to unit variance as to give every variable the same influence in the data analysis. Here, however, we use the z scales as descriptors that have been derived by principal components analysis of scaled measurement data. Hence we have not found it warranted to apply any scaling here, except in example III where we scaled the descriptor variables so that the z scales and the R_f values have the same variance.

b. Modelling of the Training Set. The biological test data are denoted y_{il} for the ith compound in the lth test, and the chemical descriptors x_{ik} for the ith compound and the kth descriptor. It is assumed that the biological activities are related to the chemical structure descriptors by means of latent variables u and t. This is formulated as

$$\mathbf{X} \rightarrow t \rightarrow u \rightarrow \mathbf{Y} \quad (1)$$

At the same time these latent variables, u and t, model the $\mathbf{Y}$ and $\mathbf{X}$ matrices; see eq 2 and 3.

$$y_{il} = \bar{y}_l + \sum_{a=1}^{A} u_{ia} c_{al} + f_{il}$$

$$\text{in matrix form: } \mathbf{Y} = 1\bar{y} + UC + F \quad (2)$$

$$x_{ik} = \bar{x}_k + \sum_{a=1}^{A} t_{ia} p_{ak} + e_{ik}$$

$$\text{in matrix form: } \mathbf{X} = 1\bar{x} + TP + E \quad (3)$$

The number of significant factors (A) in the models is estimated by cross validation.[21,32]

The models for $\mathbf{Y}$ and $\mathbf{X}$ resemble ordinary PC models. However, the PLS models differ in the respect that they are calculated as to simultaneously (a) minimize the residuals E and F and (b) yield latent variables u and t which are optimally correlated. In this way the predictions of $\mathbf{Y}$ by $\mathbf{X}$ are better than by PC regression.

The predictive relation between $\mathbf{Y}$ and $\mathbf{X}$ is modelled in terms of the latent variables:

$$u_{ia} = t_{ia} b_a + h_{ia}$$

$$\text{in matrix form: } \mathbf{U} = TB + H \ (B \text{ is diagonal}) \quad (4)$$

This gives the predictive relation for $\mathbf{Y}$:

$$\mathbf{Y} = TBC + F \quad (5)$$

(36) (a) Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238. (b) Wold, S.; Dunn, W. J., III *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 6.

Table VII. Summary of Results from the Data Analysis of QSARs in Four Families of Peptides

set	n^a	m^b	A^c	explained variance of biol act.
(I) oxytocins	22	4	3	OA = 88%, PA = 64%
(II) pseudopeptides	13	1	*d*	
(III) pepstatins	7	1	1	$-\log K_i$ = 80%
(IV) pentapeptides	15	5	3	log RAI = 97%

[a] n = number of analogues. [b] m = number of varied positions. [c] A = number of dimensions in the model. [d] No significant model according to cross validation.

c. Predictions of Activity for New Analogues. For predictions of activities of new analogues, the descriptor data for the new analogues are first inserted in the **X** model, which was calculated in "Step 4b". This results in t values that with eq 4 give u values. These u values inserted in eq 2 give predictions of the biological activities of the new analogue. The standard deviation of the residuals e_i (RSD), i.e., the degree of fit of the descriptor data of the new analogues to the **X** model, can be compared to the RSD of the training set analogs. If the RSD of the new analogue is considerably larger than this typical RSD, this indicates that the structure of the new analogue differs significantly from the training set and that the predicted activity values for this analogue are less reliable.

A more detailed description of the PLS method, with algorithms, is given in ref 20, 21, 32, and 37.

Step 5: The Postulation of New Analogues. A main goal for a QSAR is usually to predict the structure of new more potent and selective analogues. In our approach three scales, z_1, z_2, and z_3, are used to describe each varying position. Applying the PLS method to these data results in loadings (p_{ak}), one for each z value in each model dimension (A). Hence, for a one-component model (A = 1) we get three loadings for each varied amino acid position. The loadings reflect the influence of the different z values in their corresponding positions in the model. A high absolute value of a loading indicates that the corresponding z value contains much information related to the biological activity. The signs of the loadings give information about how the z values at the different positions are related to the biological activity. Hence, for each varied amino acid position, the magnitude and signs of the loadings indicate which of the positions that are important and how these should be modified to give a new analogue with increased (or decreased) biological activity. Quantitative predictions of the biological activities for the new peptide analogues are then calculated according to "Step 4c".

However, as for all chemical models, the more different the chemical structure of the new analogue is from the training set analogues, the less precise the prediction. This is analogous to calibration models in for example analytical chemistry, where the models are less precise far from the calibration domain.[37]

Results

We here report the result from QSAR studies on four different families of peptides. The raw data and resulting parameter tables are available as supplementary material. In the examples we have used the three z values as descriptors of the varied positions and the PLS method to model the data. A complete set of PLS parameters is given for one of the examples (IV). Peptides substituted with other than the 20 natural amino acids are not included since we presently only have z values for the coded amino acids. A summary of the results is given in Table VII. We presently study some peptide sets containing peptides with noncoded amino acids. The results will be reported when z values have been developed also for noncoded amino acids. In view of the lack of design of the investigated peptide series, we refrain from a detailed mechanistic interpretation of the resulting models.

I. Oxytocin Analogues. We have analyzed a set of oxytocin (I) analogues compiled by Sneath.[1] The biological data for this set of 22 oxytocin analogues originate from about 10 different publications dating from the late 1950s and early 1960s. Hence, it is expected that the data contain some interlaboratory variation that may complicate the modelling of the structure–activity relationship (see "Step 3").

1 2 3 4 5 6 7 8 9
Cys-*Tyr*-*Ile*-*Gln*-Asn-Cys-Pro-*Leu*-*Gly*-NH_2
I (oxytocin)

The analogues are varied in four positions, 2, 3, 4, and 8. Positions 2, 3, and 8 were described by three z values each. Position 4 was described by an indicator variable with Glu = 1 and Ser = 0, since this position only was varied with these two amino acids. Two biological activities were reported for the analogues, oxytocic activity (OA, contraction of isolated rat uterus) and pressor activity (PA, rise in blood pressure of rat). The analysis of the data for the 22 analogues resulted in a four-component PLS model describing 88% of the variance in OA and 64% of the variance in PA (Figures 1 and 2).

To investigate the predictive capability of the model, a test set of 15 peptide analogues (23–37) were collected from a compilation by Berde et al.,[38] who collected them from several different publications. It was noted by Berde et al. that the technical variations between the different laboratories could introduce some interlaboratory variation. As seen from Figures 1 and 2, the agreement between the observed and the predicted activities is good for most of the analogues. The test set analogues marked with an asterisk in Figures 1 and 2 (26, 30, 34, and 35) have very high RSD values (2.72, 2.43, 3.26, and 2.75, respectively) compared with the RSD (0.70) of the training set. This indicates that the predictions for these analogues may be less precise. Hence, the PLS model gives information about the reliability of predictions. One low active peptide (37) was predicted to have moderate oxytocic activity and pressor activity. Thus only for one out of 15 test set peptides was the activities badly predicted.

This training set has also been analysed by Simon[24] and Sneath,[1] Simon using the approach described above in "Step 1a". In the study by Sneath, four principal components, v_I–v_{IV}, were extracted from a "resemblance matrix" derived from the 20 natural amino acids and 134 noncontinuous descriptors. Sneath described the peptides by taking the sum of v_I for all positions, v_{II} for all positions, etc. Hence, in his regression analyses he had at the most four variables, v_I(tot.)–v_{IV}(tot.). His data analyses of the structure–activity relationships resulted in weak models "... predictions of the biological activity of new peptides ... would probably be better than chance, though not of high accuracy". We applied our multipositional approach to the description using Sneath's scales (v_I–v_{IV}) and the PLS data analytic method to model the peptide activities OA and PA. This resulted in a two-component model describing 77% and 44% of the variance in OA and PA,

(37) Lindberg, W.; Persson, J.-Å.; Wold, S. *Anal. Chem.* **1983**, *55*, 643.

(38) Berde, B.; Boissonnas, R. A. In *Handbuch der experimentellen Pharmakologie*; Springer: Berlin, 1968; Vol. 23, pp 802–863.

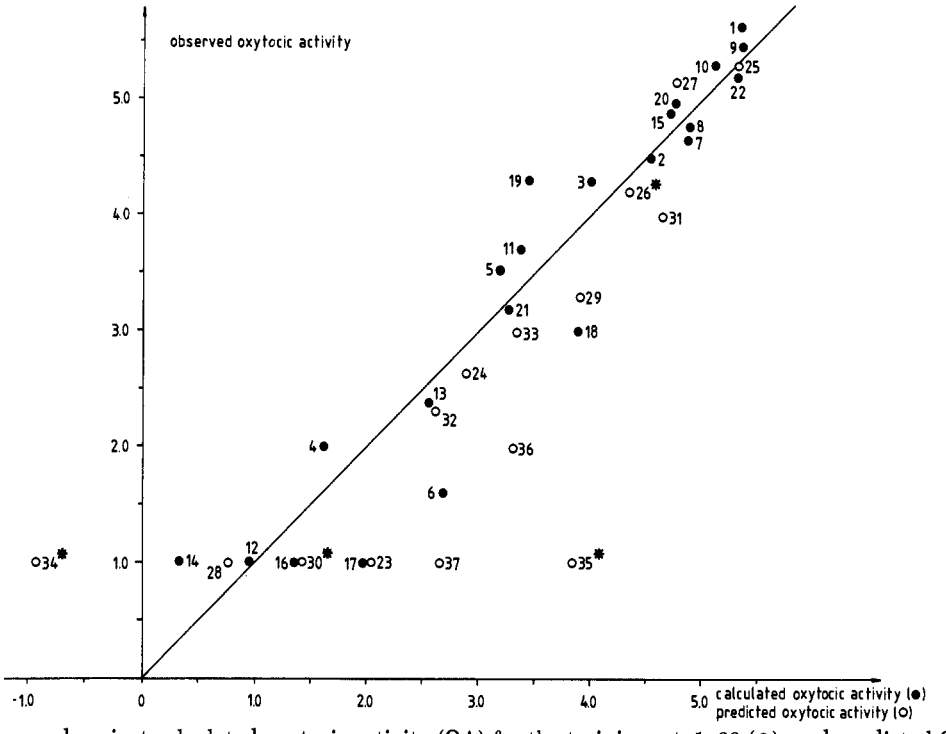

Figure 1. Plot of observed against calculated oxytocic activity (OA) for the training set, 1–22 (●), and predicted OA for the test set, 23–37 (○). The analogues are substituted in positions 2, 3, 4, and 8 as follows (peptides 1–37): YIQL, FIQL, YFQL, YYQL, FFQL, YWQL, YLQL, YVQL, YIQI, YIQV, YFQK, YYQK, FFQK, FYQK, YIQK, SIQK, YWQK, FIQK, YFQR, YIQR, YFQH, YISI, SIQL, LIQL, YISL, YIQG, YIQA, FYQL, FIQR, YSQK, YISQ, FFQR, YFSK, GIQL, YGQL, HFQL, HFQK. The peptides marked with an asterisk in the plot are those with high RSD values (see text).

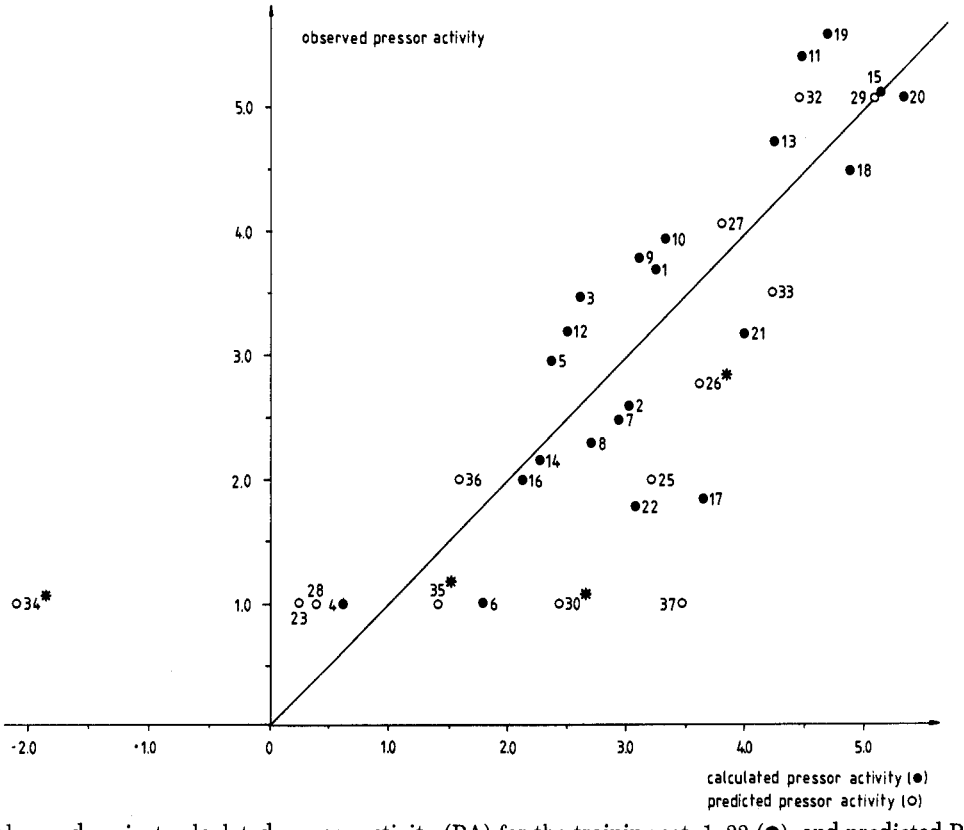

Figure 2. Plot of observed against calculated pressor activity (PA) for the training set, 1–22 (●), and predicted PA for the test set, 23, 25–30, and 32–37 (○). PA for analogues 24 and 31 were not given in the compilation by Berde et al.[38] The peptides marked with an asterisk in the plot are those with high RSD values (see text).

compared to 88% and 64% using our z_1–z_3 scales. Hence, Sneath's scales contain information, but in his data analysis this is obscured by an inefficient approach to the peptide description.

II. Pseudopeptides. For a set of 13 pseudopeptides of the general structure II, where X is one of the amino acids Gly, Ala, Ile, Leu, Phe, Pro, Asn, Met, Thr, Trp, Tyr, Asp, and Lys, Rodriques et al.[39] reported the oncostatic

activity evaluated on L1210 leukemia and acute toxicity.

$$ClCH_2CH_2N(NO)CO(X)NHCH_2CH_2Cl$$
II

The three z values were used as descriptors of change of amino acid. No successful QSAR model was obtained for this set. Hence, either the descriptor data does not contain any information that can be related to the variation of these biological activities or the biological data contain only little information that is related to the change in chemical structure of the amino acids.

III. Pepstatin Analogues. For seven analogues of pepstatin of the general structure III, the inhibition (inhibition constants k_i) of porcine pepsin was reported by Rich et al.[40] Here Iva = isovaleryl, (*S*,*S*)-Sta = 4(*S*)-

$$\overset{1}{\text{Iva}}\text{-}(\overset{2}{\text{X}})\text{-}(S,S)\text{-}\overset{3}{\text{Sta}}\text{-}(\overset{4}{\text{Y}})\text{-}\overset{5}{\text{Iaa}}$$
III

amino-3(*S*)-hydroxy-6-methylheptanoic acid, and Iaa = isoamylamide. The analogues varied in positions 2 and 4. The z values of the amino acids in the varied positions and the reported R_f values from TLC measurements on the peptide analogues were used as descriptors. Prior to the data analysis, the R_f and z values were scaled to unit variance. A two-component PLS model describing 80% of the variance in biological activity was obtained (Figure 3).

IV. Bradykinin Potentiating Pentapeptides. The bradykinin potentiating activity for 15 pentapeptides were reported by Ufkes et al.[41] in 1978, who varied the peptides in all five positions. These 15 pentapeptides were used as a training set to develop a QSAR. Schaper[26] has earlier made a QSAR study on this set of analogues using the Fujita–Ban approach.[25] However, that model cannot be used to make predictions of biological activity for peptides containing other amino acids than those already included in the set. We have previously analyzed this set with preliminary z scales.[3]

We describe the five positions using the z values, giving an **X** matrix with 15 descriptor variables. The biological activity was modelled as the logarithm of the activity index relative to peptide 1 (log RAI). The PLS analysis gives three significant components with the PLS latent variable inner regression coefficients b_a as follows: $b_1 = 0.29$, $b_2 = 0.19$, and $b_3 = 0.07$. Together, the three components describe 97% of the variance in biological activity. The parameters of the PLS model are presented in Tables VIII and IX.

After this model was developed, another set of 15 active bradykinin potentiating pentapeptides (16–30) and one inactive (RAI = 0) pentapeptide (31) reported by Ufkes et al.[42] in 1982 was used as a test set to investigate the predictive power of the PLS model. As seen from Figure 4, the prediction of the biological activity for the 15 new active peptides is satisfactory. However, there seems to be some systematic difference in peptide activity between the sets measured in 1978 and 1982 that cannot be modelled. This is seen in Figure 4 where eight peptides have lower observed activity than expected from the predictions. This may be due to a change in the biological test system. The inactive peptide (Gly-Gly-Gly-Gly-Gly) was predicted to have a RAI = 0.01; i.e., it is correctly predicted to be inactive.

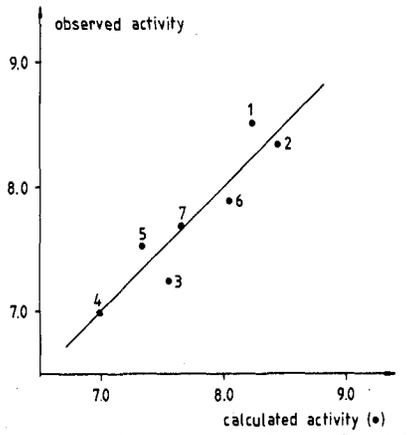

Figure 3. Plot of the observed and calculated activity for pepstatin analogues. The analogues are substituted in positions 2 and 4 as follows (peptides 1–7): VA, VL, VG, GA, AA, LA, FA.

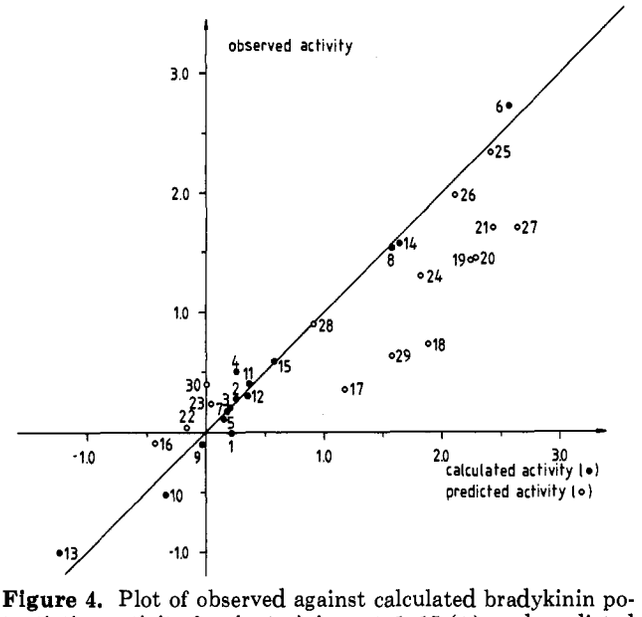

Figure 4. Plot of observed against calculated bradykinin potentiating activity for the training set, 1–15 (●), and predicted activity for the test set, 16–30 (○) of pentapeptides.

Discussion

For rigid molecules QSARs have been developed where the varied substituents in the different positions are described by substituent descriptors, e.g., σ values, π, E_s, etc.[7] Here we have demonstrated that this approach also works for small flexible peptides, albeit with other scales, z_1, z_2, and z_3.

The present approach is based on three cornerstones:

(i) A characterization of amino acids with three scales (z_1, z_2, and z_3). The scales were estimated by PCA from a matrix consisting of 29 properties for each of the twenty amino acids.

(ii) A numerical description of the peptides in terms of the three scales z_1, z_2, and z_3. Each varied position is thus described by three variables.

(iii) PLS, a multivariate data analytical method that can relate the chemical description to the biological activities even when the descriptors are collinear and numerous.[20,21,32]

In three out of four investigated families of peptides, significant QSARs were obtained (see Table VII). For two of the families it was possible to further validate (not only by cross validation of the training set) the predictive ca-

(39) Rodriguez, M.; Imbach, J.-L.; Martinez, J. *J. Med. Chem.* **1984**, *27*, 1222.
(40) Rich, D. H.; Salituro, F. G. *J. Med. Chem.* **1983**, *26*, 904.
(41) Ufkes, J. G. R.; Visser, B. J.; Heuver, G.; van der Meer, C. *Eur. J. Pharmacol.* **1978**, *50*, 119.
(42) Ufkes, J. G. R.; Visser, B. J.; Heuver, G.; Wynne, H. J.; van der Meer, C. *Eur. J. Pharmacol.* **1982**, *79*, 155.

Table VIII. Latent Variables, t_{ia}, of the PLS Model and Biological Activity for Bradykinin Potentiating Pentapeptides with 1–15 as Training Set and 16–31 as Test Set

no.	peptide	obsd[a] log RAI	calcd[c] log RAI	predicted[d] log RAI	t_{i1}	t_{i2}	t_{i3}
1	VESSK	0.00	0.22		−1.83	1.34	0.62
2	VESAK	0.28	0.26		−1.57	1.27	0.34
3	VEASK	0.20	0.21		−0.82	−0.05	0.06
4	VEAAK	0.51	0.25		−0.56	−0.12	−0.21
5	VKAAK	0.11	0.16		−0.64	−0.29	−0.63
6	VEWAK	2.73	2.57		6.56	0.58	1.47
7	VEAAP	0.18	0.19		−0.95	−0.18	0.62
8	VEHAK	1.53	1.58		0.83	3.91	2.10
9	VAAAK	−0.10	−0.03		−0.78	−0.80	−1.54
10	GEAAK	−0.52	−0.34		−1.52	−2.84	2.61
11	LEAAK	0.40	0.37		−0.28	0.68	−1.77
12	FEAAK	0.30	0.35		−0.11	1.14	−4.04
13	VEGGK	−1.00	−1.23		−5.55	−0.80	1.15
14	VEFAK	1.57	1.64		4.80	−1.18	0.26
15	VELAK	0.59	0.58		2.44	−2.66	−1.05
16	AAAAA	−0.10		−0.43	−1.44	−2.01	−1.13
17	AAYAA	0.46		1.17	2.58	−0.32	0.38
18	AAWAA	0.75		1.89	5.68	−1.31	0.55
19	VAWAA	1.43		2.24	6.08	−0.12	0.70
20	VAWAK	1.45		2.29	6.34	−0.09	0.14
21	VKWAA	1.71		2.44	6.22	0.38	1.61
22	VWAAK	0.04		−0.16	−0.88	−1.08	−2.13
23	VAAWK	0.23		0.04	0.76	−1.26	−5.56
24	EKWAP	1.30		1.82	5.42	−1.59	1.37
25	VKWAP	2.35		2.42	6.09	0.36	1.90
26	RKWAP	1.98		2.11	5.78	−0.55	1.14
27	VEWVK	1.71		2.64	6.84	0.50	1.45
28	PGFSP	0.90		0.91	3.79	−2.34	−2.83
29	FSPFR	0.64		1.57	3.98	2.01	−5.82
30	RYLPT	0.40		0.00	2.35	−4.34	−4.39
31	GGGGG	*b*		−2.11	−6.90	−4.16	3.25

[a] Reported by Ufkes et al.[40,41] [b] RAI = 0.00. [c] Biological activity for the training set calculated by the PLS model. [d] Biological activity for the test set predicted by the PLS model.

Table IX. Variable-Related Model Parameters of the PLS Model for Bradykinin Potentiating Pentapeptides[a]

variable	position	mean	p_{1k}	p_{2k}	p_{3k}	w_{1k}	w_{2k}	w_{3k}
z_1	1	−2.61	−0.06	−0.45	0.63	−0.14	−0.36	0.46
z_2		−2.36	0.03	0.35	−0.65	0.07	0.17	−0.72
z_3		−1.04	−0.02	−0.06	−0.09	−0.06	−0.16	−0.30
z_1	2	2.86	0.02	0.06	0.12	0.05	0.15	0.27
z_2		0.32	0.01	0.04	0.07	0.03	0.08	0.13
z_3		−0.26	0.02	0.02	0.04	0.03	0.06	0.14
z_1	3	−0.31	−0.72	0.69	0.14	−0.60	0.76	−0.04
z_2		−1.12	0.65	0.37	0.29	0.78	0.48	0.20
z_3		0.24	0.02	0.20	0.13	0.08	0.28	0.18
z_1	4	0.47	−0.15	0.02	0.10	−0.14	0.07	0.16
z_2		−1.96	0.17	0.08	−0.11	0.16	−0.09	−0.57
z_3		0.17	−0.02	0.01	0.01	−0.02	0.01	−0.01
z_1	5	2.57	0.03	0.02	−0.06	0.03	0.00	−0.08
z_2		1.37	0.00	0.00	−0.01	0.00	0.00	−0.01
z_3		−2.78	−0.04	−0.02	0.09	−0.05	0.00	0.10
log RAI		0.45	1.00	1.00	1.00			

[a] p_{ak} = loadings, w_{ak} = PLS weights.

pabilities of the models. This was done by separate test sets of peptides, which not had been used to establish the QSAR model. The predicted activities for the test set peptides were in good accordance with the measured biological activities except for a few cases. However, with PLS not only predictions of the activity for a test set peptide analogue is obtained but also a measure of how well the peptide fits the QSAR model. Indeed, the test set peptides with poor predictions of their activities had a poor fit to the QSAR model. This type of information cannot be obtained from a data analysis using multiple regression.

A referee has raised the critisism that in these applications there are so many variables that a good fit to the training sets is almost certain. This would be true if we used multiple regression for the data analysis. PLS, however, is a projection method similar to prinicipal components regression, where the data matrix **X** is first projected to a small number of latent variables, t_a, which then are used as independent variables in a regression model. Hence, as long as the number of latent variables is small compared to the number of compounds in the training set, there is no overfit, even if the original number of x variables is very large. This matter has been extensively covered in the literature; see, e.g., ref 20, 21, 32, 36b, and 37.

Moreover, we have used cross validiation (CV) to test the predictive significance of the PLS models. With CV, part of the training set is kept out from the model development and then later predicted by the model. Then another part of the training set is kept out, a new model developed, and the kept out data predicted. This is repeated until each training set compound has been kept out once and only once. The predictions are then, finally, compared with the actual values, and only model dimen-

sions that give predictions significantly better than chance are retained in the model.

Finally, we have shown in two of the four examples that the predicted activities for new sets of compounds that were not involved in the model development were close to the actually observed values, much closer than would be expected by chance (Figures 2 and 4).

With the current view on peptides, one would expect the variation in conformation of the peptides to have a great influence on the biological activity. The present description does not explicitly take the conformation into account. Hence, the success of the modelling of three out of four peptide families may be interpreted in either of three ways: (1) Conformation is not important in these families. (2) Conformation is important, but in some way implicitly described by the z scales. (3) Conformation is important but all peptides in each set can adopt the bioactive conformation with low energy.

We refrain from taking a strong position for one of these three possibilities and just note that in these examples the prediction of the biological activities of small flexible peptides seems to be considerably simpler than can be expected from their conformation flexibilities.

For the future development of peptide QSAR, we have proposed the estimation of improved descriptors for the amino acids and also an extension to noncoded amino acids and other fragments of interest.

Another area for improvement is the often overlooked problem of how to construct a series of peptide analogues suited for structure–activity studies. Here we propose fractional factorial designs as a possibility for constructing informative training sets. This design problem has also been discussed in a separate paper.[33d]

A designed test series can be used in different peptide families. Thus a set of designed peptide fragments (as those 16 proposed in Table VI) can be introduced as tetrapeptide units in different peptide families. Such predesigned sets of peptide fragments simplify the synthesis of multipositionally varied peptides. Furthermore, for a design with only coded amino acids, a set of codon sequences can be constructed that corresponds to a set of designed peptide fragments. The rapid development of protein engineering[43] may then make it possible to produce designed sets of mature proteins and enzymes for QSAR studies.

Acknowledgment. Grants from the Swedish Natural Science Council (NFR), the Swedish Council for Planning and Coordination of Research (FRN), and the National Swedish Board for Technical Development (STU) are gratefully acknowledged.

Registry No. A, 56-41-7; V, 72-18-4; L, 61-90-5; I, 73-32-5; P, 147-85-3; F, 63-91-2; W, 73-22-3; M, 63-68-3; K, 56-87-1; R, 74-79-3; H, 71-00-1; G, 56-40-6; S, 56-45-1; T, 72-19-5; C, 52-90-4; Y, 60-18-4; N, 70-47-3; Q, 56-85-9; D, 56-84-8; E, 56-86-0; Bradykinin, 58-82-2.

Supplementary Material Available: Tables containing the property matrix for the amino acids, PLS model parameters, and biological activities for the oxytocins, pepstatins, and pseudopeptides (8 pages). Ordering information is given on any current masthead page.

(43) Fox, J. L. *ASM News* **1985**, *51*, 566.

C3-Methylated 5-Hydroxy-2-(dipropylamino)tetralins: Conformational and Steric Parameters of Importance for Central Dopamine Receptor Activation

Anette M. Johansson,*† J. Lars G. Nilsson,† Anders Karlén,† Uli Hacksell,† Kjell Svensson,‡ Arvid Carlsson,‡ Lennart Kenne,§ and Staffan Sundell‖

Department of Organic Pharmaceutical Chemistry, Uppsala Biomedical Center, University of Uppsala, S-751 23 Uppsala, Sweden, Department of Pharmacology, University of Göteborg, S-400 33 Göteborg, Sweden, Department of Analytical Chemistry, KabiVitrum, S-112 87 Stockholm, Sweden, and Department of Structural Chemistry, University of Göteborg, S-400 33 Göteborg, Sweden. *Received November 19, 1986*

C3-Methyl-substituted derivatives of the potent dopamine (DA) receptor agonist 5-hydroxy-2-(di-*n*-propylamino)tetralin (5-OH-DPAT) have been synthesized and their conformational preferences have been studied by use of NMR spectroscopy, X-ray crystallography, and molecular mechanics calculations (MMP2). The compounds were tested for activity at central DA receptors, by use of biochemical and behavioral tests in rats. (2*R*,3*S*)-5-Hydroxy-3-methyl-2-(di-*n*-propylamino)tetralin [(−)-8] was demonstrated to be a highly potent DA receptor agonist, while the other new compounds were of low potency or inactive. Results obtained confirmed the hypothesis that the tetralin inversion angle Φ and the direction of the N-electron pair (*N*-H) τ_N are conformational parameters of critical importance for DA D_2 receptor activation in the 2-aminotetralin series. The high potency of (−)-8 allowed an extension of a previously defined "partial DA D_2 receptor excluded volume".

Previous studies of the pharmacology, stereochemistry, and conformational dynamics of the enantiomers of the potent dopamine (DA) receptor agonist 5-hydroxy-2-(di-*n*-propylamino)tetralin (5-OH-DPAT)[1] and their C1-methyl-substituted derivatives[2,3] have demonstrated that DA agonistic C5-oxygenated 2-aminotetralins have the same sense of chirality at the nitrogen-bearing carbon (C2). In addition, results obtained[2b] suggested that, in this series, only compounds that easily assume τ_N values[4] around 60° in tetralin conformations with Φ values[4] around 0° are

† University of Uppsala.
‡ Department of Pharmacology, University of Göteborg.
§ KabiVitrum.
‖ Department of Structural Chemistry, University of Göteborg.

(1) (a) McDermed, J. D.; McKenzie, G. M.; Freeman, H. S. *J. Med. Chem.* **1976**, *19*, 547. (b) Tedesco, J. L.; Seeman, P.; McDermed, J. D. *Mol. Pharmacol.* **1979**, *16*, 369. (c) Freeman, H. S.; McDermed, J. D. In *The Chemical Regulation of Biological Mechanisms*; Creighton, A. M., Turner, S., Eds.; Royal Society of Chemistry: London, 1982; p 154. (d) Seiler, M. P.; Markstein, R. *Mol. Pharmacol.* **1982**, *22*, 281. (e) Seiler, M. P.; Markstein, R. *Mol. Pharmacol.* **1984**, *26*, 452. (f) Wikström, H.; Andersson, B.; Sanchez, D.; Lindberg, P.; Arvidsson, L.-E.; Johansson, A. M.; Nilsson, J. L. G.; Svensson, K.; Hjorth, S.; Carlsson, A. *J. Med. Chem.* **1985**, *28*, 215.

0022-2623/87/1830-1135$01.50/0