37. Structure-Lipophilicity and Structure-Polarity Relationships of Amino Acids and Peptides

by **Philippe Vallat, Patrick Gaillard, Pierre-Alain Carrupt, Ruey-Shiuan Tsai,** and **Bernard Testa***

Institut de Chimie Thérapeutique, BEP, Université de Lausanne, CH-1015 Lausanne-Dorigny

(31.X.94)

The objectives of this study were to gain insights into the structure-lipophilicity relationships of peptides and to propose an improved model for estimating their lipophilicity. First, existing databases were extended to obtain the distribution coefficients of a total of 208 free or protected peptides (di- to pentapeptides). The polarity parameters (Λ) of 23 free amino acids and 19 protected amino acids ($AcNH-CHR-CONH_2$) and of their side chains were calculated from experimental distribution coefficients and computed molecular volumes. An analysis of the polarity parameters revealed that the hydrophobicity of the amino-acid side chains is largely reduced due to the polar field of the backbone. The polarity parameters of the peptides were then obtained in a similar manner and shown to be highly correlated with the sum of the polarity parameters of their side chains, *i.e.*, the lipophilicity of peptides can be calculated from their molecular volume and the sum of their side-chain polarities using the regression established for each individual series of peptides (*Fig. 1*). This last restriction is essential since the polarity and lipophilic increment of a NH–C*H–CO unit were shown to decrease with increasing length of backbone.

1. Introduction. – Endogenous peptides such as many hormones and neurotransmitters were found to modulate a wide variety of biological functions [1] [2]. The potency and specificity of these endogenous compounds make it clear why the design of synthetic peptides and of peptidomimetics is already one of the major issues in drug research. Indeed, these approaches led to the discovery of new lead compounds and drugs such as potent peptide receptor antagonists and enzyme inhibitors.

Peptides and their constituent amino acids show a wide range of physicochemical and structural properties. Amino-acid side chains contain polar, non-polar, charged, or uncharged groups and differ considerably in size and flexibility. Characterizing the structural and physicochemical properties of amino acids and peptides is an important condition in molecular biology to unravel the properties of proteins, and in molecular pharmacology to rationalize and predict the biological properties of peptide drugs. What is at stake here is a better understanding of structure-property-activity relationships of peptides. In this context, lipophilicity parametrization of amino acids and peptides is of major concern. Indeed, lipophilicity is a physicochemical property of particular significance in drug design, because it encodes a wealth of information on a solute's structure and the intermolecular interactions it elicits [3–5].

Lipophilicity can be expressed by the logarithm of the partition coefficient (*i.e.* log P, which refers to a well-defined electrical state of the solute, for example the neutral or zwitterionic state) or the logarithm of the distribution coefficient (*i.e.*, log D, which is obtained at a given pH and may thus result from the contributions of more than one electrical forms). A frequently used approach when investigating structure-property or

structure-activity relationships is the development of additive models whereby the target property or activity is factorized into contributions from molecular fragments, *e.g.* amino-acid residues in the case of peptides. Using this approach, *Fujita* and coworkers [6–8] used 124 peptides (di-, tri-, tetra-, and pentapeptides) to investigate whether their log *D* at pH 7.0 in an octanol/H_2O system could be calculated from the sum of the lipophilic increments ($\Sigma\pi$) of the amino-acid side chains (*Path 1* in *Fig. 1*). Taken alone, these increments failed to yield any satisfactory model. For an apparently good correlation to be obtained, a plethora of additional variables were needed in the form of indicators accounting for various structural features such as the possibility of β-turns and the presence of specific amino acids (see *Eqn. 1*) [8]. Despite its apparent success (*i.e.*, its good correlation coefficient), this model suffers from several shortcomings, *e.g.* the great number of independent variables, unproven assumptions about β-turns and α-helices, and too high cross correlations between some parameters. As a result, this model does not appear reliable beyond the explored property space.

$$\begin{aligned}\log D^{7.4} = {} & 0.942\,\Sigma\pi - 0.582\,I_{pep} + 0.546\,E_s^c(R_N) + 0.295\,[\Sigma E_s^c(R_M) + E_s^c(R_C)] \\ & + 0.516\,I_{turn} + 0.764\log f_{i+2} + 0.144\,I_Y + 0.378\,I_W + 0.659\,I_M \\ & + 1.581\,(I_S + I_T) - 0.807\,I_P(N) - 0.346\,I_P(\#\,pep) - 3.866 \end{aligned} \quad (1)$$

$$n = 124,\quad r^2 = 0.935,\quad s = 0.209,\quad F_{12,111} = 134$$

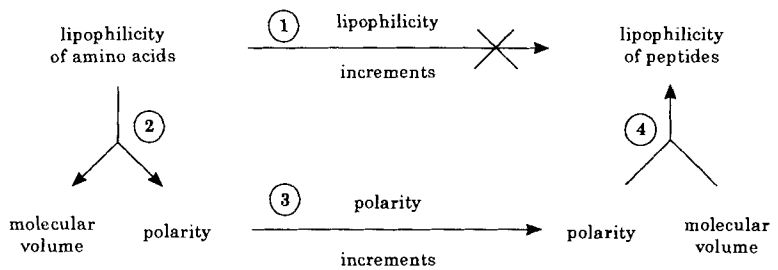

Fig. 1. *How to calculate the lipophilicity of peptides from that of amino acids? Path 1* is a traditional approach, an unsatisfactory one, while *Path 2–4* is explored in this study.

In an attempt to gain a better understanding of structure-lipophilicity relationships in peptides, we extended the database investigated by *Akamatsu* and *Fujita* [8] and used simple volume and polarity parameters. However, due mainly to the limitations of commercial availability and the range of experimentally accessible lipophilicity, the number of investigated peptides remains very low with respect to all possible combinations.

In the explored set of peptides, our results show that the lipophilicity of peptides can be calculated from their molecular volume and the polarity of their amino-acid residues (*Path 2–4* in *Fig. 1*), as long as dipeptides, tripeptides, tetrapeptides, and pentapeptides are treated separately. Indeed, with increasing length of the backbone, the polarity of each single peptide bond is shown here to decrease, while the quality of the correlations decreases somewhat, perhaps due to increasing flexibility.

2. Materials and Methods. – *Chemicals.* Anal. grade octan-1-ol and morpholinopropanesulfonic acid (= morpholine-4-propanesulfonic acid; MPS) were purchased from *Merck*, Darmstadt, Germany. The various

peptides investigated experimentally were obtained from *Bachem*, Bubendorf, Switzerland. Throughout this study, the single-letter symbols of amino acids will be used; as for peptides, the N-terminal residue will always be written on the left.

Measurement of Distribution Coefficients (log *D* values). Octanol/H_2O distribution coefficients were measured by centrifugal partition chromatography (CPC) using a horizontal flow-through, multilayer CPC instrument (*Pharma-Tech Research*, Baltimore, MD, USA). Octanol and the buffer phase (0.01M MPS, pH 7.4) were used as either mobile or stationary phase. The use of a zwitterionic buffer avoids the formation of ion pairs with ionized peptides, which would partition into the org. phase [6]. The speed of rotation was set at 1000 rpm and the applied flow rate between 0.3 and 6.5 ml/min. The detection wavelength was set at 220 nm. Further details on the equipment and procedures can be found in [9].

Measurements of Ionization Constants (pK_a values) *and Calculation of* log P *Values*. The pK_a values of several dipeptides were determined by the pH-metric method using a *Sirius-PCA101* instrument (*Sirius Analytical Instruments*, Forest Row, East Sussex, GB). The apparatus was equipped with a semi-micro combined electrode (*Orion 81035C*), a temperature probe, a stirrer, a precision dispenser, and a six-way valve for distributing reagents and titrants (0.5M HCl, 0.1M KCl, and 0.5M KOH). The weighted samples (1–3 mg) were supplied manually to the glass vial, the titrant and all other reagents being added automatically. Ar Gas was introduced into the vial during titration to exclude the dissolved CO_2, and the vial was maintained at 25 ± 1° with a temperature-controlled water bath. Once the titration was completed, the built-in *Bjerrum* plots and statistical algorithms were used to calculate precisely the pK_a values [10]. The detailed experimental procedures and data analyses were described elsewhere [11]. This novel technique is particularly well-suited for multiprotic solutes with overlapping pK_a values.

For a diprotic solute with a strong acidic and a strong basic group, the partition coefficient of the zwitterion (noted here log $P^{+/-}$) can be calculated [3] from a distribution coefficient at a known pH (log *D*) using *Eqn. 2*:

$$\log P^{+/-} = \log D + \log[1 + 10^{(pK_{a1} - pH)} + 10^{(pH - pK_{a2})}] \quad (2)$$

3. Theory and Calculations. – *Calculation of Molecular Volumes.* The various parameters used to describe molecular size (*e.g.* the molecular volume *V*, molecular surface area, *Connolly* surface, solvent-accessible surface area) are largely interrelated and carry comparable information. Deviations in their statistical intercorrelations are mainly due to molecular flexibility. In this study, we used the size descriptor that varies the least with conformation, namely the molecular volume. Preliminary calculations with a few peptides showed that this parameter varies by less than 5% over their entire conformational range.

Each investigated compound was taken in its state of maximal ionization and the geometry of a low-energy conformation optimized by the Tripos force-field including the electrostatic term (dielectric constant $\varepsilon = 2$). The molecular volume (in Å^3) was calculated with the program MOLSV (QCPE No. 509) using atomic radii determined by *Gavezotti* [12].

Statistical Calculations. Graphical-data analysis and statistical calculations were performed with the QSAR module of the SYBYL software [13] or with the TSAR program [14] using the PLS statistical method [15–17] with the leave-one-out cross-validation technique. In all equations, the 95% confidence intervals, estimated with TSAR using the jackknife approach, are given in parenthesis. Following recent rules [18] [19], the cross-validated *r*-square (q^2) is also given, together with the conventional *r*-square (r^2), the standard error (*s*), and the *F* value (*F*). In the discussion, all statistical models with a standard error *s* inferior or equal to the uncertainty of measurements were considered as significant. All softwares used were running on *Silicon Graphics* workstations (*Personal Iris 4D35*, *Power Series 4D320*, or *Indigo R4000*).

The Factorization of Lipophilicity. In previous studies, we showed that lipophilicity (log *P*) can be factorized into a hydrophobic term and a polarity term which we designated as Λ (see *Eqn. 3*) [3] [4]. In this equation, the hydrophobic term encodes all

intermolecular forces proportional to the compound size (*i.e.*, mainly hydrophobic forces [20] between the solute and the aqueous phase), while the polar term expresses *Van der Waals* forces and mainly H-bonds between the solute and both phases, *e.g. Eqn.4* [4], where π^* is the dipolarity/polarizability of solutes and β their H-bond acceptor basicity [21].

$$\log P = \text{hydrophobicity} - A = (a \cdot V + c) - A \tag{3}$$

$$A_{oct} = 0.64\ (\pm 0.12)\ \pi^* + 3.90\ (\pm 0.20)\beta + 0.19\ (\pm 0.10)$$
$$n = 168,\quad r_2 = 0.918,\quad s = 0.25 \tag{4}$$

For non-polar compounds ($A = 0$), the coefficient 'a' in *Eqn.3* is the slope of the line relating log P and V, and 'c' is the intercept. Thus, the hydrophobic term is easily determined for alkanes ($A = 0$) and was recalculated here using the literature log P values of H_2 and unbranched alkanes from methane to tetradecane (*Table 1*), yielding *Eqn.5*. Interestingly, the intercept in this equation is not zero as previously believed [4]. While *Eqn.5* cannot be extrapolated to solutes smaller than H_2, its non-zero intercept may perhaps result from discrete cavity effects which cannot be neglected and warrant further study.

$$\log P\ (\text{alkane}) = 3.087 \cdot 10^{-2}\ (\pm 0.136 \cdot 10^{-2}) \cdot V + 0.346\ (\pm 0.199)$$
$$n = 14,\quad q^2 = 0.995,\quad r^2 = 0.997,\quad s = 0.145,\quad F = 3619 \tag{5}$$

Table 1. *Lipophilicity and Molecular Volume of Linear Alkanes*

	log P[a])	V [Å^3]		log P[a])	V [Å^3]
Hydrogen	0.45	10.8	Heptane	4.50	131.7
Methane	1.09	28.7	Octane	5.15	148.6
Ethane	1.81	45.7	Nonane	5.65	166.0
Propane	2.36	63.0	Undecane	6.54	200.4
Butane	2.89	79.9	Dodecane	6.80	217.2
Pentane	3.39	97.3	Tridecane	7.56	234.7
Hexane	3.90	114.2	Tetradecane	8.00	251.5

[a]) Taken from the Pomona 1993 database [42].
[b]) Calculated according to *Section 3*.

The combination of *Eqns.3* and *5* leads to *Eqn.6* which allows the polarity parameter of any solute to be calculated. *Eqn.6* means that for any solute, the polarity parameter is the difference between the log P of a virtual alkane of identical volume (calculated with *Eqn.5*) and the log P of that solute. Using *Eqn.6*, the A parameter of the investigated amino acids and peptides was calculated from their molecular volume (see above), taking as the lipophilicity descriptor the apparent lipophilicity, better called the distribution coefficient, measured at pH 7.0 (log $D_{7.0}$).

$$A = 3.087 \cdot 10^{-2} \cdot V + 0.346 - \log P \tag{6}$$

The side-chain polarity A_{SC} of each amino acid (free or protected) was calculated simply by substracting the A of glycine (free or protected) from that of the amino acid

considered (see *Eqn. 7*). This approach allows the total polarity of a peptide to be conveniently split into two components, namely the polarity of all side chains ($\Sigma\Lambda_{SC}$) and that of the backbone (Λ_B) (see *Eqn. 8*). Hence the distribution coefficient of peptides will be factorized as shown in *Eqn. 9*.

$$\Lambda_{SC} = \Lambda - \Lambda_{glycine} \tag{7}$$

$$\Lambda = \Sigma\Lambda_{SC} + \Lambda_B \tag{8}$$

$$\log D_{7.0} = a \cdot V - \Sigma\Lambda_{SC} - \Lambda_B \tag{9}$$

Conformational Calculations. The conformational space of the two dipeptides LP and PL was explored by high-temperature molecular dynamics with the DYNAMICS module of the software SYBYL [22]. The Tripos force field [23] including an electrostatic term calculated with a dielectric constant ε of 78.0 was used to optimize the geometry of the 200 conformers randomly retained during each simulation of 100 ps at 2000 K. For LP and PL, 14 and 48 different conformers, respectively, were identified.

4. Results and Discussion. – Due to the sample of the investigated series of peptides, which contains very few ionized side chains, the hydrophobic term is largely correlated with the polarity term. This relation between independent variables does not allow *Eqn. 9* to be used directly to calculate the log *D* of peptides. To overcome the statistical difficulty of separating hydrophobic from polar contributions, we demonstrate below the interest of using a model based only on *Eqn. 8*, *i.e.*, on the analysis of polarity parameters Λ which in fact account for most of the observed variation in the experimental log *D* values.

Polarity Parameters of Free Amino Acids. The distribution coefficients of free amino acids at pH 7.0 (*Table 2*) were reported in previous studies from this laboratory [24] [25]. For amino acids without an ionizable side chain, these values are the partition coefficients of the zwitterionic forms. For the amino acids with an ionizable side chain (except

Table 2. *Distribution Coefficients, Molecular Volumes, and Polarity Parameters of Free Amino Acids*

	log D[a])	V[b])	Λ[c])	Λ_{SC}[d])		log D[a])	V[b])	Λ[c])	Λ_{SC}[d])
A	−2.77	82.1	5.7	0.3	N	−3.48	111.2	7.3	1.9
Abu[e])	−2.53	99.5	6.0	0.6	Ahx[e])	−1.54	133.8	6.0	0.6
C	−2.55	99.2	6.0	0.6	Avl[e])	−2.11	116.3	6.1	0.7
D	−3.61	103.5	7.2	1.8	P	−2.62	106.1	6.2	0.9
E	−3.51	119.9	7.6	2.2	Q	−3.11	127.9	7.4	2.0
F	−1.44	157.9	6.7	1.3	R	−3.79	161.6	9.1	3.8
G	−3.00	65.4	5.4	0.0	S	−3.00	89.1	6.1	0.7
H	−2.85	137.1	7.4	2.1	T	−2.83	105.6	6.4	1.1
I	−1.80	133.0	6.3	0.9	V	−2.29	115.9	6.2	0.8
K	−3.77	147.8	8.7	3.3	W	−1.15	187.8	7.3	1.9
L	−1.72	134.4	6.2	0.8	Y	−2.11	164.6	7.5	2.2
M	−2.10	135.9	6.6	1.3					

[a]) Distribution coefficients at pH 7.0 taken from [24] [25].

[b]) Calculated molecular volume in Å^3.

[c]) Polarity of the amino acids, calculated according to *Eqn. 6*.

[d]) Polarity of the side chains, calculated according to *Eqn. 7*.

[e]) Abu = 2-aminobutanoic acid; Ahx = norleucine = 2-aminohexanoic acid; Avl = norvaline = 2-aminopentanoic acid = Ape.

histidine), the reported log D values are the partition coefficients of the triply ionized forms. The side chain of histidine (H) is mainly unionized at pH 7.0, and hence the reported value is for the zwitterionic form with negligible contribution from the triply ionized form. In fact, the log D of histidine is *ca.* −3.3 at pH 5 and 6 and remains very close to −2.85 in the pH range 7–8 [25].

A plot of calculated molecular volumes (*Table 2*) *vs.* log D values is shown in *Fig. 2*. If in such a relationship the amino acids were to behave like alkanes, a straight line with slope $3.087 \cdot 10^{-2}$ would connect the amino acids with an unbranched alkyl side chain, namely glycine (G), alanine (A), 2-aminobutanoic acid (Abu), norvaline (Avl), and norleucine (Ahx). This is verified only for Avl and Ahx (slope 0.03), indicating that the CH_2/CH_3 groups in A, Abu, and Avl are prevented from expressing their 'normal' hydrophobicity, as also seen with other compounds [26]. Only in Ahx is the last increment fully hydrophobic [24], a fact also verified for the increment from valine (V) to leucine (L) and isoleucine (I; slope 0.03). In addition, a detailed analysis of polar side chains indicates that the polarity of the OH and amide groups is markedly decreased in amino acids relative to simple alcohols and amides (not shown). This is in agreement with other studies [27].

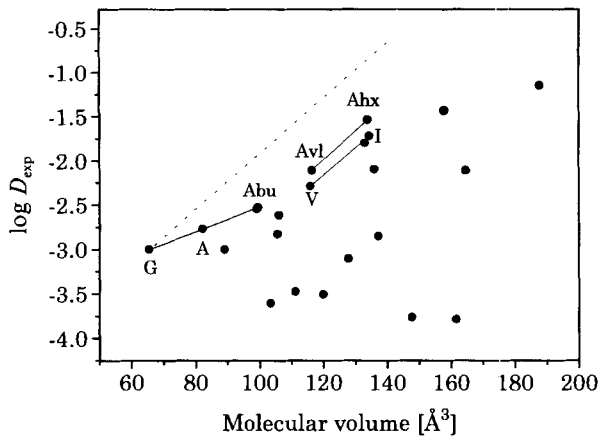

Fig. 2. *Calculated molecular volumes* (see *Table 2*) vs. *log* D *values of free amino acids*

From the experimental log D and calculated V values, the polarity parameter Λ of each amino acid is calculated using *Eqn. 6*. *Eqn. 7* then allows the side-chain polarity to be obtained (Λ_{SC}). These parameters are also listed in *Table 2* and will form the basis of the structure-property relationships presented below.

Polarity Parameters of Protected Amino Acids. To calculate the polarity parameters of protected amino acids, we used published log D data measured at pH 7.1 [28]. The compounds were the primary amides of *N*-acetylated amino acids (Ac × NH_2 = AcNH – CHR – $CONH_2$). Their distribution coefficients, calculated molecular volumes, and polarity parameters are listed in *Table 3*. The polarity of the side chains was also calculated, and is designated as Λ_{SCp} to allow differentiation with Λ_{SC}. Interestingly, the two parameters Λ_{SC} and Λ_{SCp} are highly correlated, with a slope and an

Table 3. *Distribution Coefficients, Molecular Volumes, and Polarity Parameters of Protected Amino Acids* $Ac{-}X{-}NH_2$ (AcNH–CHR–$CONH_2$)

Parent amino acid X	log D[a])	V[b])	A[c])	A_{SCp}[d])	Parent amino acid X	log D[a])	V[b])	A[c])	A_{SCp}[d])
A	−1.52	121.6	5.6	0.2	N	−2.41	150.1	7.4	2.0
D	−2.60	143.2	7.4	2.0	P	−1.34	144.6	6.2	0.7
E	−2.47	159.7	7.8	2.3	Q	−2.05	168.2	7.6	2.2
F	−0.04	197.2	6.5	1.1	R	−2.84	201.7	9.4	4.0
G	−1.83	104.8	5.4	0.0	S	−1.87	128.8	6.2	0.8
H	−1.70	176.6	7.5	2.1	T	−1.57	145.5	6.4	1.0
I	−0.03	172.7	5.7	0.3	V	−0.61	155.7	5.8	0.4
K	−2.82	188.5	9.0	3.6	W	0.42	227.1	6.9	1.5
L	−0.13	174.1	5.9	0.4	Y	−0.87	203.8	7.5	2.1
M	−0.60	175.5	6.4	1.0					

[a]) Distribution coefficients at pH 7.1 taken from [28]. The value for the cysteine derivative appears doubtful and was not considered.
[b]) Calculated molecular volume in Å^3.
[c]) Polarity of the amino acids, calculated according to *Eqn. 6*.
[d]) Polarity of the side chains, calculated according to *Eqn. 7*.

intercept close to 1 and 0, respectively (see *Eqn. 10*). *Eqn. 10* is a clear indication that the polarity of the side chains is affected by the AcNH and $CONH_2$ groups just as much as by the NH_3^+ and COO^- groups.

$$A_{SCp} = 1.14\ (\pm 0.11)\ A_{SC} - 0.30\ (\pm 0.23)$$
$$n = 10,\quad q^2 = 0.95,\quad r^2 = 0.96,\quad s = 0.22,\quad F = 459 \qquad (10)$$

Distribution Coefficients of Free Di- and Tripeptides. The studies of *Akamatsu et al.* [6] [8] contain 32 free dipeptides and 38 free tripeptides. We extended this database by measuring the log D values at pH 7.0 of 11 free dipeptides and 10 free tripeptides using centrifugal partition chromatography. In addition, the log D values of 3 free dipeptides were found in the literature. These experimental log D values are compiled in *Table 4*. The experimental error on these values is estimated to be 0.1 log D units.

We also recalculated these log D values using *Eqn. 1*. As can be seen in *Table 4*, the correspondence is excellent between some experimental and calculated values, but far from convincing in other cases (the deviations range from −0.47 to 0.95). The standard deviation is 0.39, *i.e.*, much greater than for the 32 peptides included in *Eqn. 1*. This confirms that *Eqn. 1* has only modest predictive value.

Ionization Constants of Dipeptides and Lipophilicity of Their Zwitterions. One objection that can be raised against using log D values at a fixed pH (in this study 7.0) is that depending on the pK_a values, they may express different populations of ionic forms for each solute. To assess the uncertainty resulting from a fixed pH, we measured the pK_a values of 21 dipeptides using the *Sirius* titrator (*Table 5*). These values in turn allow the calculation of the log P values of the zwitterionic forms (using *Eqn. 2*). The correlation between log P and log D values in *Table 5* is $r^2 = 0.98$, with a standard deviation of 0.1. We thus estimate that the uncertainty in lipophilicity due to ionization and ionic populations is *ca.* 0.2 logarithmic units.

Table 4. *Distribution Coefficients of Free Di- and Tripeptides*

	log D[a])	log D[d])		log D[a])	log D[d])
A F	−2.21	−2.10	W G	−1.98	−2.15
A L	−2.46	−2.18	W S	−2.20	−2.11
A W	−2.21[b])	−1.74	A F A	−2.11	−2.44
F G	−2.31	−2.52	A L A	−2.88	−2.73
F S	−2.59	−2.48	G F G	−2.74	−2.87
G F	−2.30	−2.29	G W G	−2.71	−2.75
G G	−2.92[b])	−3.87	L F L	−0.40	−0.68
G V	−2.98[c])	−3.05	L H L	−1.59	–
G W	−2.17	−1.93	L V L	−1.17	−1.24
L H	−2.74	–	P P P	−3.13	–
S F	−2.54	−2.38	V Y V	−2.22	−2.28
V G	−2.74	−3.37	Y Y Y	−2.13	−1.65

[a]) Distribution coefficients at pH 7.0 as measured by CPC, except when indicated otherwise.
[b]) Data from the Pomona 1993 database [42].
[c]) Data from *Fujita et al.* [43].
[d]) log D Values calculated by *Eqn. 1*.

Table 5. *p*K_a *and log* P *Values of Dipeptides*

	pK_{a1}[a])	pK_{a2}[a])	log $P^{+/-}$[b])	Δ[c])		pK_{a1}[a])	pK_{a2}[a])	log $P^{+/-}$[b])	Δ[c])
A F	7.91	3.08	−2.13	0.08	L F	7.70	3.25	−0.95	0.20
A I	8.01	3.34	−2.51	0.09	L H[d])	7.78	2.76	−2.53	0.21
A L	8.02	3.35	−2.40	0.06	L Y	7.69	3.32	−1.77	0.17
F F	7.17	2.98	−0.58	0.27	M L	7.31	3.39	−1.66	0.18
F G	7.38	3.60	−2.16	0.15	S F	7.23	2.93	−2.17	0.37
F L	7.20	3.41	−0.83	0.34	S L	7.30	3.35	−2.37	0.12
F S	7.48	3.02	−2.35	0.24	V G	8.00	3.21	−2.75	−0.01
F Y	7.14	3.19	−1.38	0.30	V Y	7.78	3.23	−2.27	0.25
G F	8.12	2.93	−2.25	0.05	W F	7.30	3.20	−0.17	0.30
G G	8.08	3.10	−2.89	0.03	W G	7.76	3.12	−1.86	0.12
G W	8.06	3.14	−2.12	0.05					

[a]) pK_a Values of the NH_2 and COOH groups, respectively, as measured by the *Sirius* titrator at 25°.
[b]) log P Values of the zwitterionic forms, as calculated from the log D values at pH 7.0 using *Eqn. 2*.
[c]) log $P^{+/-}$ − log D.
[d]) The pK_a of the histidinyl residue is 6.69.

Polarity of Free Dipeptides. The database now contains 46 free dipeptides whose experimental log D values, molecular volumes, molecular polarity, and side-chain polarities were determined[1]). Based on the above discussions, the experimental uncertainty is *ca.* 0.3 logarithmic units. A good correlation is shown to exist between Λ and $\Sigma\Lambda_{SC}$ (see *Eqn. 11*). This is in fact the application of *Eqn. 8* to dipeptides. *Eqn. 11* is of significance since it demonstrates that the log D of dipeptides can be calculated with high prediction power (note the excellent cross-validated correlation coefficient $q^2 = 0.91$) from a single property of the constitutive residues, namely the polarity of their side chains.

$$\Lambda = 1.08\ (\pm 0.16)\ \Sigma\Lambda_{SC} + 7.55\ (\pm 0.38)$$
$$n = 45,\quad q^2 = 0.91,\quad r^2 = 0.92,\quad s = 0.28,\quad F = 493 \qquad (11)$$

[1]) All previously published experimental values and the calculated parameters for the di-, tri-, tetra-, and pentapeptides are available from the authors upon request.

The good correlation between Λ obtained from log D and Λ calculated from *Eqn. 11* is shown in *Fig. 3*. Four dipeptides emerge as notable outliers, namely GG, FP, IP, and LP, which all are more lipophilic than calculated. The outlying behavior of GG is difficult to explain but could be due to a higher flexibility of the peptide backbone in the absence of α-substituents. The case of the three proline-containing peptides is investigated in the next section, being shown to possibly result from a *trans-cis* isomerization. When the four outliers are deleted from *Eqn. 11*, an increase of the predictive power of the correlation is obtained (see *Eqn. 12*).

$$\Lambda = 1.00\ (\pm 0.11)\ \Sigma\Lambda_{SC} + 7.79\ (\pm 0.24)$$
$$n = 41,\quad q^2 = 0.95,\quad r^2 = 0.96,\quad s = 0.17,\quad F = 938 \tag{12}$$

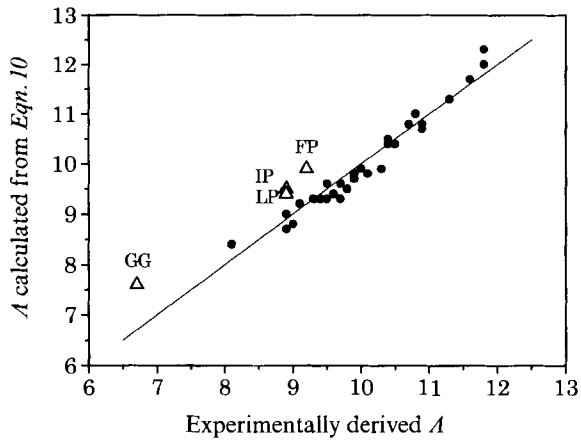

Fig. 3. *Total polarity of free dipeptides, as determined from their log* D *values,* vs. *the same parameter calculated by* Eqn. 11. The line in this figure represents the perfect prediction (slope = 1, intercept = 0).

trans-cis-*Isomerization of Peptides with a C-Terminal Proline.* The higher-than-expected lipophilictiy of peptides with a terminal proline had already been noted by *Akamatsu et al.* [8] who postulated steric effects and a perturbed solvation of carboxylate groups. We believe that such deviations may result from changes in geometry caused by a C-terminal proline, since it is well-known that the peptidic bond of a C-terminal proline may isomerize from the *trans*- to the *cis*-form [29–33]. The stability of the *cis*-form relative to the *trans*-form depends on the peptide itself and its environment [34–39]. For dipeptides, this isomerization would bring closer together the COO^- and NH_3^+ groups, thereby decreasing their polarity and producing a higher-than-expected lipophilicity. To examine this hypothesis, we explored the conformational hyperspace of LP and PL using high-temperature molecular dynamics. The results (not shown) demonstrate that LP has two families of low-energy conformers with C–N distances of 3.3–3.5 and 4.5–5.5 Å, respectively, while the low-energy conformers of PL all have a C–N distance of 4.5–6.2 Å. Such conformational effects cannot be expressed in the Λ_{SC} parameter of proline and may explain the deviant behavior of FP, IP, and LP in *Eqn. 11*.

Polarity of Protected Dipeptides. The parameters of 31 protected dipeptides [7] (with N-terminus and C-terminus being AcNH and $CONH_2$, resp.) were also calculated. As for

free dipeptides, a significant correlation is found between the polarity of the peptides and the sum of side-chain polarities (see *Eqn. 13*). This equation quite logically uses the side-chain polarities of protected amino acids (*Table 3*). Interestingly, using the side-chain polarities of free amino acids (*Table 2*) yields *Eqn. 14* which is only marginally interferior to *Eqn. 13*. However, any comparison between *Eqns. 13* and *14* would be informative only if based on a much larger set of dipeptides.

$$\Lambda = 0.99\ (\pm 0.13)\ \Sigma\Lambda_{SCp} + 7.58\ (\pm 0.29) \qquad (13)$$
$$n = 31,\quad q^2 = 0.92,\quad r^2 = 0.91,\quad s = 0.27,\quad F = 372$$

$$\Lambda = 1.10\ (\pm 0.17)\ \Sigma\Lambda_{SC} + 6.92\ (\pm 0.42) \qquad (14)$$
$$n = 31,\quad q^2 = 0.89,\quad r^2 = 0.91,\quad s = 0.17,\quad F = 286$$

Polarity of Cyclodipeptides. A small set of cyclodipeptides is also investigated here using lipophilicity data measured in this study. The distribution coefficients, molecular volumes, and polarity parameters are listed in *Table 6*. As with the dipeptides above, this set of special compounds is amenable to our approach, the polarity of the cyclodipeptides being well predicted form the side-chain polarities of protected amino acids (see *Eqn. 15*). Replacing $\Sigma\Lambda_{SCp}$ with $\Sigma\Lambda_{SC}$ yields an equation of slightly smaller statistical quality ($r^2 = 0.91$, eqn. not shown). Although the small number of compounds investigated does not allow a detailed analysis, it is interesting to note the deviant behavior of cSS. Its experimental polarity is higher than predicted by both equations, suggesting that the decrease in polarity of residue side chains is less marked in the rigid environment of *cis*-peptidic bonds [27].

$$\Lambda = 1.06\ (\pm 0.16)\ \Sigma\Lambda_{SCp} + 5.12\ (\pm 0.36) \qquad (15)$$
$$n = 11,\quad q^2 = 0.93,\quad r^2 = 0.92,\quad s = 0.23,\quad F = 145$$

Table 6. *Parameters of Cyclodipeptides*

	log D[a])	V[b])	Λ[c])	$\Sigma\Lambda_{SC}$[d])	$\Sigma\Lambda_{SCp}$[e])		log D[a])	V[b])	Λ[c])	$\Sigma\Lambda_{SC}$[d])	$\Sigma\Lambda_{SCp}$[e])
cA G	−1.49	110.8	5.3	0.3	0.2	cG S	−2.24	118.4	6.2	0.7	0.8
cA H	−1.57	180.4	7.5	2.4	2.3	cG Y	−0.69	193.1	7.0	2.2	2.1
cF F	1.59	279.0	7.4	2.6	2.2	cS S	−2.48	142.2	7.2	1.4	1.6
cF S	−0.45	209.7	7.3	2.0	1.9	cS Y	−1.09	217.1	8.1	2.9	2.9
cG F	0.05	184.5	6.0	1.3	1.1	cW Y	1.05	316.3	9.1	4.1	3.6
cG H	−1.82	164.0	7.2	2.1	2.1						

[a]) log D Values at pH 7.0 measured in this study.
[b])[c])[d]) See *Table 2*.
[e]) See *Table 3*.

Polarity of Free and Protected Tripeptides. A large set of 48 free tripeptides from the literature [8] and from our measurements (*Table 4*) and a smaller set of 22 protected tripeptides [7] (with N-terminus and C-terminus being AcNH and $CONH_2$, resp.) were analyzed. The estimated uncertainty of the lipophilicity measurement is *ca.* 0.4. Thus, for the free tripeptides, a highly significant statistical model is obtained (see *Eqn. 16*). Using for the free tripeptides the $\Sigma\Lambda_{SCp}$ parameters yields a less satisfactory correlation ($r^2 = 0.79$, eqn. not shown).

$$\Lambda = 0.97\ (\pm 0.06)\ \Sigma\Lambda_{SC} + 9.92\ (\pm 0.22)$$
$$n = 48,\quad q^2 = 0.94,\quad r^2 = 0.94,\quad s = 0.24,\quad F = 728 \tag{16}$$

For the protected tripeptides, the correlation between their polarity parameter and the increments of protected amino acids is still good (see *Eqn. 17*), but somewhat less than *Eqn. 16*. Surprisingly, a better correlation is obtained with $\Sigma\Lambda_{SC}$ ($r^2 = 0.98$, eqn. not shown). We believe this effect to be an artifact caused by a biased sampling, glycine-containing tripeptides being over-represented in this set.

$$\Lambda = 0.87\ (\pm 0.18)\ \Sigma\Lambda_{SCp} + 9.89\ (\pm 0.30)$$
$$n = 22,\quad q^2 = 0.87,\quad r^2 = 0.90,\quad s = 0.23,\quad F = 170 \tag{17}$$

Polarity of Tetra- and Pentapeptides. The parameters for the 33 variable free tetrapeptides gives a calculated correlation between peptide polarity and side-chain polarity of decreased quality compared to shorter peptides (*Eqn. 18*). Nevertheless, the correlation is statistically significant and demonstrates that $\Sigma\Lambda_{SC}$ accounts for more than 75% of the variance of Λ. Furthermore, the standard deviation of the model ($s = 0.41$) is of the same order as the experimental precision. The two most outlying peptides are YPGI and IAAI, which are more and less polar, respectively, than predicted by the model.

$$\Lambda = 0.93\ (\pm 0.19)\ \Sigma\Lambda_{SC} + 11.85\ (\pm 0.69)$$
$$n = 33,\quad q^2 = 0.74,\quad r^2 = 0.77,\quad s = 0.41,\quad F = 105 \tag{18}$$

For the 23 free pentapeptides, the model of *Eqn. 19* is obtained. This model is again poorer than those for di- and tripeptides, and it is of a statistical quality practically identical to that of *Eqn. 18*, accounting for almost 80% of the variance of Λ, with a standard deviation close to the experimental uncertainty. The peptide GGGGG had to be removed from the analysis not so much because it is the most outlying solute (its predicted polarity is much greater than measured, probably due to a considerable flexibility), but because it introduces a statistical bias, being too far removed from the cloud of all other compounds in a Λ *vs.* $\Sigma\Lambda_{SC}$ plot. The statistical problem generated by the peptide GGGGG is clearly demonstrated by the very large 95% confidence intervals in *Eqn. 19* with respect of those of *Eqn. 20*, which does not include GGGGG.

$$\Lambda = 1.22\ (\pm 0.53)\ \Sigma\Lambda_{SC} + 12.40\ (\pm 2.04)$$
$$n = 23,\quad q^2 = 0.80,\quad r^2 = 0.88,\quad s = 0.52,\quad F = 153 \tag{19}$$

$$\Lambda = 0.96\ (\pm 0.25)\ \Sigma\Lambda_{SC} + 13.39\ (\pm 0.99)$$
$$n = 22,\quad q^2 = 0.74,\quad r^2 = 0.79,\quad s = 0.43,\quad F = 0.73 \tag{20}$$

For free tetra- and pentapeptides, the use of $\Sigma\Lambda_{SCp}$ instead of $\Sigma\Lambda_{SC}$ leads to equations devoid of statistical significance ($r^2 = 0.45$ and 0.41, resp.).

The preliminary conclusion to emerge at this stage is that di- and tripeptides yield very well indeed to the proposed analysis, while tetra- and pentapeptides demonstrate additional structural effects not taken into account by the polarity model and accounting for *ca.* 20–25% of the variance. Such effects might be of conformational nature. However, the inclusion of the structural parameters proposed by *Akamatsu* and *Fujita* [8] as conformational descriptors for peptides (*e.g.*, I_{turn} or $\log f_{i+2}$) enhance the r-squared

coefficient of *Eqns. 18* and *20* by *ca.* 10% without changing significantly the standard errors. These errors remain close to the experimental precision suggesting a statistical origin for the improvement of r^2. Thus, considering the low number of tetra- and pentapeptides selected, no definitive indication for the influence of conformational effects on their lipophilicity can be derived from our analyses.

Polarity of the Peptide Backbone. Eqns. 12, 13, 15–18, and *20* represent applications of the general *Eqn. 8* to specific classes of peptides. In other words, the intercept in these equations must represent Λ_B, the polarity of the backbone, if the slope is one. *Table 7* lists the slopes and intercepts of these equations, showing indeed that the former are never statistically different from one. Hence, the polarity of the backbone can be approximated to 5.1 for cyclodipeptides, and as shown in *Table 8* for the other investigated peptides. These polarity values (strictly fragmental polarity constants) can then be used to calculate fragmental lipophilicity constants for the backbones. The corresponding polarity and lipophilicity increments of –C*H–CONH– (or –CONH–C*H–) fragments are also compiled in *Table 8*.

Table 7. *A Comparison of Slopes and Intercepts in* Eqns. 12, 13, 15–18, *and* 20

	Slope	Λ_B	*Eqn.*
Dipeptides	1.00 (±0.11)	7.79 (±0.24)	*12*
Protected dipeptides	0.99 (±0.13)	7.58 (±0.29)	*13*
Cyclodipeptides	1.06 (±0.16)	5.12 (±0.36)	*15*
Tripeptides	0.97 (±0.06)	9.92 (±0.22)	*16*
Protected tripeptides	0.87 (±0.18)	9.89 (±0.30)	*17*
Tetrapeptides	0.93 (±0.19)	11.85 (±0.69)	*18*
Pentapeptides	0.96 (±0.25)	13.39 (±0.99)	*20*

Table 8. *Polarity and Lipophilicity Increments of Free and Protected Peptides*

	Λ_B[a])	Polarity increment	Lipophilic constant[b])	Lipophilic increment
Glycine[c])	5,4		–3.1	
		2.4		–1.0
Dipeptides[c])	7.8		–4.1	
		2.1		–0.8
Tripeptides[c])	9.9		–4.9	
		1.9		–0.4
Tetrapeptides[c])	11.8		–5.3	
		1.6		–0.3
Pentapeptides[c])	13.4		–5.6	
Protected glycine[d])	5.4		–1.9	
		2.2		–0.8
Protected dipeptides[d])	7.6		–2.7	
		2.3		–0.8
Protected tripeptides[d])	9.9		–3.5	

[a]) Rounded values.
[b]) Calculated according to *Eqn. 6* using the molecular volume calculated for each fragment.
[c]) $H_3N^+(C^*HCONH)_nC^*HCOO^-$ (C* = C-atom with a free valence).
[d]) $AcNH(C^*HCONH)_nC^*HCONH_2$.

Validation of the Polarity Approach for the Calculation of Peptide Lipophilicity. The above arguments indicate clearly that mixing various classes of peptides can only result in a loss of information. To determine the performance of our model for the prediction of log D_{exp} for all investigated peptides, we estimated log D values (*i.e.* log D_{est}) using molecular volumes and A values calculated by *Eqns. 12*, *13*, *15–19*, and *20*. *Eqn. 21* summarizes the estimation of log D_{est} where A_{calc} is determined by *Eqn. 12* for free dipeptides, *Eqn. 13* for protected dipeptides, *Eqn. 15* for cyclodipeptides, *Eqn. 16* for free tripeptides, *Eqn. 17* for protected tripeptides, *Eqn. 18* for free tetrapeptides, and *Eqn. 20* for free pentapeptides.

$$\log D_{est} = 3.087 \cdot 10^{-2} \cdot V + 0.346 - A_{calc} \tag{21}$$

The relationship between estimated and experimental log D values is expressed by *Eqn. 22* and is also shown in *Fig. 4* demonstrating that the proposed model (*Fig. 1*) is able to estimate log D values for the investigated di- to pentapeptides within the limits of experimental precision.

$$\log D_{exp} = 0.94\ (\pm 0.04) \log D_{est} - 0.08\ (\pm 0.05)$$
$$n = 208,\quad q^2 = 0.92,\quad r^2 = 0.92,\quad s = 0.28,\quad F = 2331 \tag{22}$$

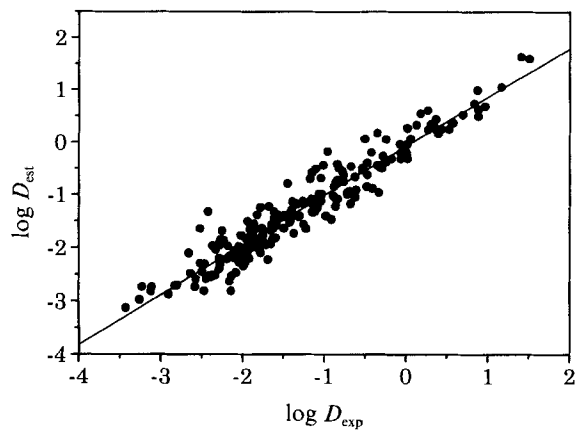

Fig. 4. *Experimental log* D *values* (log D_{exp}) vs. *estimated log* D *values* (log D_{est}) *for all peptides* (208) *in this study.* The straight line corresponds to *Eqn. 21.*

5. Conclusion. – A rigorous statistical model is established here for the first time to analyze the structural information encoded in the lipophilicity of di- to pentapeptides. While the predictability of this model is good within the explored space and offers a new method to calculate the lipophilicity of peptides from the parameters of their constitutive amino acids, its extrapolative capacity is still uncertain due to the poor representativity of the investigated peptides. In addition, for tetra- and pentapeptides, the probability of outliers must increase relative to di- and tripeptides. Furthermore, the lipophilic behavior of peptides with six or more residues is largely unexplored. Another restriction concerns configuration, since the model is valid for peptides containing only L-amino acids. Indeed, the introduction of one or more D-amino acids would result in diastereoiso-

merism also known to influence lipophilicity [40] [41]. Yet despite these limitations, the polarity scales of amino acids appear as a valuable tool in exploring structure-lipophilicity relations of peptides. Significantly, the polarity of amino acids and peptides reveals that the hydrophobicity of side chains is not fully expressed due to the polar influence of the backbone, a phenomenon of major potential importance in molecular biology.

The authors thank Prof. *Lemont B. Kier* and Dr *Nabil El Tayar* for their interest and advice. *B. T.* and *P. A. C.* are indebted to the *Swiss National Science Foundation* and *Herbette Foundation*, University of Lausanne, for support.

REFERENCES

[1] 'Peptides: Chemistry, Structure, and Biology, Proceedings of the 11th American Peptide Symposium', Eds. J. E. Rivier and G. R. Marshall, ESCOM Science Publishers B. V., Leiden, 1990.
[2] M. Mutter, S. Vuilleumier, *Angew. Chem. Int. Ed.* **1989,** *28,* 535.
[3] H. van de Waterbeemd, B. Testa, in 'Advances in Drug Research', Ed. B. Testa, Academic Press, London, 1987, Vol. 16, pp. 87–227.
[4] N. El Tayar, B. Testa, P. A. Carrupt, *J. Phys. Chem.* **1992,** *96,* 1455.
[5] B. Testa, L. B. Kier, *Med. Res. Rev.* **1991,** *11,* 35.
[6] M. Akamatsu, Y. Yoshida, H. Nakamura, M. Asao, H. Iwamura, T. Fujita, *Quant. Struct.-Act. Relat.* **1989,** *8,* 195.
[7] M. Akamatsu, S. Okutani, K. Nakao, N. J. Hong, T. Fujita, *Quant. Struct-Act. Relat.* **1990,** *9,* 189.
[8] M. Akamatsu, T. Fujita, *J. Pharm. Sci.* **1992,** *81,* 164.
[9] N. El Tayar, R. S. Tsai, P. Vallat, C. Altomare, B. Testa, *J. Chromatogr.* **1991,** *556,* 181.
[10] A. Avdeef, D. L. Kearney, J. A. Brown, A. R. Chemotti, Jr., *Anal. Chem.* **1982,** *54,* 2322.
[11] A. Avdeef, *Quant. Struct.-Act. Relat.* **1992,** *11,* 510.
[12] A. Gavezzotti, *J. Am. Chem. Soc.* **1983,** *105,* 5220.
[13] 'SYBYL 5.41, 5.55, 6.0', Tripos Associates, Inc., St-Louis, Missouri, 1993.
[14] 'Tsar 2.1', Oxford Molecular Ltd., Oxford, UK, 1993.
[15] W. J. Dunn III, S. Wold, U. Edlund, S. Hellberg, J. Gasteiger, *Quant. Struct.-Act. Relat.* **1984,** *3,* 131.
[16] R. D. Cramer III, J. D. Bunce, D. E. Patterson, I. E. Frank, *Quant. Struct.-Act. Relat.* **1988,** *7,* 18.
[17] S. Wold, E. Johansson, M. Cocchi, in '3D QSAR in Drug Design. Theory Methods and Applications', Ed. H. Kubinyi, ESCOM Science Publishers, Leiden, 1993, pp. 523–550.
[18] U. Thiaut, G. Folkers, H. Kubinyi, A. Merz, D. Rognan, in '3D QSAR in Drug Design. Theory Methods and Applications', Ed. H. Kubinyi, ESCOM Science Publishers, Leiden, 1993, pp. 711–716.
[19] U. Thibaut, G. Folkers, G. Klebe, H. Kubinyi, A. Merz, D. Rognan, *Quant. Struct.-Act. Relat.* **1994** *13,* 1.
[20] W. Bokzijl, J. B. F. N. Engberts, *Angew. Chem. Int. Ed.* **1993,** *32,* 1545.
[21] M. J. Kamlet, R. W. Taft, *J. Am. Chem. Soc.* **1976,** *98,* 377.
[22] C. Altomare, S. Cellamare, A. Carotti, G. Casini, M. Ferappi, E. Gavuzzo, F. Mazza, P. A. Carrupt, P. Gaillard, B. Testa, *J. Med. Chem.* **1995,** *38,* 170.
[23] M. Clark, R. D. Cramer III, N. Van Opdenbosch, *J. Comput. Chem.* **1989,** *10,* 982.
[24] R. S. Tsai, B. Testa, N. El Tayar, P. A. Carrupt, *J. Chem. Soc., Perkin Trans. 2* **1991,** 1797.
[25] N. El Tayar, R. S. Tsai, P. A. Carrupt, B. Testa, *J. Chem. Soc., Perkin Trans. 2* **1992,** 79.
[26] A. J. Dallas, P. W. Carr, *J. Phys. Chem.* **1994,** *98,* 4927.
[27] M. A. Roseman, *J. Mol. Biol.* **1988,** *200,* 513.
[28] J. L. Fauchère, V. Pliska, *Eur. J. Med. Chem.* **1983,** *18,* 369.
[29] R. N. Hunston, J. P. Gerothanassis, J. Lauterwein, *J. Am. Chem. Soc.* **1985,** *107,* 2654.
[30] J. Lauterwein, I. P. Gerothanassis, R. N. Hunston, *J. Chem. Soc., Chem. Commun.* **1984,** 367.
[31] D. S. Clark, J. J. Dechter, L. Mandelkern, *Macromolecules* **1979,** *12,* 626.
[32] W. E. Hull, H. R. Kricheldorf, *Biopolymers* **1980,** *19,* 1103.
[33] R. Nagaraj, Y. V. Venkatachalapathi, P. Balaram, *Int. J. Pept. Protein. Res.* **1980,** *16,* 291.
[34] C. Grathwohl, K. Wuethrich, *Biopolymers* **1981,** *20,* 2623.
[35] M. Juy, H. Lam-Thanh, K. Lintner, S. Fermandjian, *Int. J. Pept. Protein Res.* **1983,** *22,* 437.
[36] J. Bremer, G. L. Mendz, *Aust. J. Chem.* **1989,** *42,* 1011.

[37] W. J. Chazin, J. Koerdel, T. Drakenberg, E. Thulin, P. Brodin, T. Grundstroem, S. Forsen, *Proc. Natl. Acad. Sci. U.S.A.* **1989,** *86,* 2195.
[38] J. Koerdel, S. Forsen, T. Drakenberg, W. J. Chazin, *Biochemistry* **1990,** *29,* 4400.
[39] R. E. Loomis, M. Gonzalez, P. M. Loomis, *Int. J. Pept. Protein Res.* **1991,** *38,* 428.
[40] R. S. Tsai, P. A. Carrupt, B. Testa, N. El Tayar, G. L. Grunewald, A. F. Casy, *J. Chem. Res. (M)* **1993,** 1901.
[41] S. E. Evans, A. Galdes, M. Gall, *Pharmacol. Biochem. Behavior* **1991,** *40,* 1033.
[42] 'DAYLIGHT Software 4.32', Daylight Chemical Information System, Inc., Irvine, California, 1993.
[43] M. Asao, H. Iwamura, M. Akamatsu, T. Fujita, *J. Med. Chem.* **1987,** *30,* 1873.