# Analysis of $\chi_1$ Rotamer Populations from NMR Data by the CUPID Method

Željko Džakula,[†] Arthur S. Edison,[‡] William M. Westler, and John L. Markley*,[‡]

*Contribution from the Biochemistry Department, College of Agricultural and Life Sciences, University of Wisconsin—Madison, 420 Henry Mall, Madison, Wisconsin 53706. Received September 30, 1991*

**Abstract:** The theoretical foundations of CUPID, a new method for determining the rotamer probability distributions from NMR spin–spin coupling constants and NOE data, has been presented (Džakula, Ž.; Westler, W. M.; Edison, A. S.; Markley, J. L. *J. Am. Chem. Soc.*, accompanying paper in this issue). Here we apply the method to find the distribution of rotamers about the $\chi_1$ angle in L-leucine (cation and anion) on the basis of six measured coupling constants [Fischman, A. J.; Wyssbrod, H. R.; Agosta, W. C.; Cowburn, D. *J. Am. Chem. Soc.* 1978, *100*, 54–58] between various pairs of spins across the $C^\alpha–C^\beta$ bond. We also test the ability of the method to reproduce simulated rotamer probability distributions from calculated spin–spin coupling and proton–proton cross-relaxation parameters with and without simulated errors. In all cases CUPID reproduced error-free distributions to the precision of the numerical procedures used. Data with simulated errors were reproduced quite well, indicating that given sufficient quality and quantity of experimental data, CUPID is able to construct accurate rotamer probabilities from experimental data. Guidelines for CUPID's implementation and its accuracy as a function of experimental errors are presented.

## 1. Introduction

The preceding paper[1] presents the theoretical foundations of a new method for the analysis of nuclear magnetic spin–spin couplings and cross-relaxation data: ContinUous ProbabIlity Distribution analysis of rotamers (CUPID). The aim of CUPID is to reconstruct the continuous probability distribution of rotamers by calculating Fourier coefficients of the probability distribution from NMR data. In this paper, we illustrate the implementation of the method, explore its range of applicability, and demonstrate the wide variety of probability distributions that can be successfully determined. We also discuss some practical aspects of CUPID including its sensitivity to experimental errors.

We used the method to determine the rotamer populations and angles about the $C^\alpha–C^\beta$ bond in L-leucine (cation and anion) from experimental multinuclear coupling constants obtained by Fischman et al.[2] The resulting continuous probability distribution can be compared to the populations of three discrete staggered conformations calculated from the same set of coupling constants.[2]

Only the first- and second-order Fourier coefficients of the probability distribution can be derived directly from coupling constants. To evaluate third-order Fourier coefficients directly from experimental data and thus to improve the quality of the calculated distribution, coupling information must be supplemented with NOEs. In macromolecular systems, such as small proteins, where extensive NOE data have been measured, the exact probability distribution of rotamers in solution is uncertain. In addition, the actual NOE data can be used only after spin diffusion and strong coupling between $\beta$-methylene protons,[3] as well as exchange of amide protons[4] and their fast relaxation,[3] are taken into account. Finally, few appropriate heteronuclear coupling constants have been determined for proteins, although the technology for their measurement is under rapid development.[5–7] Therefore, before obtaining experimental input for CUPID analysis of a macromolecule, we decided to test whether such effort would be justified. The test was to reproduce a variety of simulated rotamer distributions from the corresponding rotationally averaged NOEs and coupling constants. We chose to analyze rotations about the $C^\alpha–C^\beta$ bond ($\chi_1$ angle) in an amino acid located in a rigid, isotropically-reorienting $\alpha$-helix. This model was chosen because the geometry gives rise to a known pattern of NOEs[8] and because the $\alpha$-helix is a well-defined structure of biological im-

portance. However, extensions to other structures are straightforward.

First, average spin–spin couplings and NOEs were calculated by numerical integration from a random sampling of simulated probability distributions. From these average data, Fourier coefficients of the probability distribution were determined analytically by using CUPID (equations described in the preceding paper).[1] Real data, however, contain a wealth of possible errors, such as poor signal-to-noise ratio, poor digitization, signal overlap, deformations of the baseline (or baseplane), and J-peaks in NOESY spectra. In addition to errors in measurement, complications such as the approximate nature of Karplus relations, uncertainties in bond lengths and angles, spin diffusion, and different correlation times for internal motion and overall rotation introduce errors in data analysis. We have systematically introduced errors of both kinds into our simulations to test the practical application and limits of the method.

A typical full data set used in the simulations consisted of six vicinal coupling constants (homo- and heteronuclear) across the $C^\alpha–C^\beta$ bond and eight NOE peaks from individual $\beta$-protons to neighboring backbone protons, with and without noise as described above. However, we also considered smaller sets of simulated data, including a completely homonuclear case with only two couplings. The effects of data size and data type on the probability distribution are presented in the Conclusion section.

## 2. Methodology

**2.1. Rotamers of L-Leucine.** Extensive coupling data, but no NOE values, are available for L-leucine,[2] so the number of directly evaluated Fourier coefficients of the probability distribution of $\chi_1$ rotamers is limited to four ($\rho_1$, $\sigma_1$, $\rho_2$, and $\sigma_2$). Therefore, only two Gaussians were used to reconstruct the distributions (see section 2.2.1. of the preceding paper[1]). Four independent Gaussian parameters (two positions, $\varphi_1$ and $\varphi_2$, one probability, $p_1$, and one

(1) Džakula, Ž.; Westler, W. M.; Edison, A. S.; Markley, J. L. *J. Am. Chem. Soc.*, preceding paper in this issue.

(2) Fischman, A. J.; Wyssbrod, H. R.; Agosta, W. C.; Cowburn, D. *J. Am. Chem. Soc.* 1978, *100*, 54–58.

(3) Neuhaus, D.; Williamson, M. P. *The Nuclear Overhauser Effect in Structural and Conformational Analysis*; VCH: Weinheim, 1989.

(4) Delepierre, M.; Larvor, M.-P.; Baleux, F.; Goldberg, M. E. *Eur. J. Biochem.* 1991, *201*, 681–693.

(5) Montelione, G. T.; Winkler, M. E.; Rauenbuehler, P.; Wagner, G. *J. Magn. Reson.* 1989, *82*, 198–204.

(6) Edison, A. S.; Westler, W. M.; Markley, J. L. *J. Magn. Reson.* 1991, *92*, 434–438.

(7) Schmieder, P.; Kurz, M.; Kessler, H. *J. Biomolec. NMR* 1991, *1*, 403–420.

(8) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; John Wiley: New York, 1986; pp 120–163.

†Permanent address: Institute of Physiology and Biochemistry, Faculty of Biology, School of Sciences, and Biophysical Laboratory, Center for Multidisciplinary Studies, University of Belgrade, Studentski trg 3, 11000 Belgrade, Serbia, Yugoslavia.
*To whom correspondence should be addressed.
‡Graduate Biophysics Program, University of Wisconsin—Madison.

width, $\Delta$, which was assumed to be the same for the two Gaussians) were adjusted so that the values of $\rho_1$, $\sigma_1$, $\rho_2$, and $\sigma_2$, calculated as

$$\rho_n = \frac{1}{\pi}e^{-(n\Delta/2)^2}\sum_{i=1}^{2}p_i\cos(n\varphi_i)$$

and

$$\sigma_n = \frac{1}{\pi}e^{-(n\Delta/2)^2}\sum_{i=1}^{2}p_i\sin(n\varphi_i), \quad n = 1, 2, ...,$$

exactly matched the values obtained from the experimental couplings by using linear regression [eqs 15–27 of the preceding paper[1]]. As an alternative to the two Gaussians, coupling data for L-leucine were supplemented by the nonnegativity condition in order to estimate the Fourier coefficients $\rho_3$ and $\sigma_3$, as described in section 2.2.2. of the preceding paper.[1]

**2.2. Construction of Simulated Probability Distributions in a Model $\alpha$-Helix.** Each initial probability distribution $\rho_{\text{initial}}(\chi_1)$ was constructed as follows. Three discrete values of the torsion angle $\varphi_i$ ($i = 1, 2, 3$) and the corresponding probabilities $p_i$ (which ranged from 0 to 1) were assumed. Gaussian curves of widths $\Delta_i$ were then centered at angles $\varphi_i$ and $\varphi_i \pm 2\pi$ and weighted by $p_i$

$$\rho_{\text{initial}}(\chi_1) = \sum_{i=1}^{3}p_i\rho_i(\chi_1) \tag{1}$$

where

$$\rho_i(\chi_1) = \frac{1}{\sqrt{\pi}}\sum_{n=-1}^{1}\frac{1}{\Delta_i}\exp\{-[\chi_1 + 2n\pi - \varphi_i]^2\Delta_i^{-2}\} \tag{2}$$

and

$$\sum_{i=1}^{3}p_i = 1 \tag{3}$$

Note that the angular distribution generated in this way satisfies the conditions of periodicity and nonnegativity. The normalization condition

$$\int_{-\pi}^{\pi}\rho_{\text{initial}}(\chi_1)\,d\chi_1 = 1 \tag{4}$$

is also satisfied to a very good approximation if $\Delta_i \ll 2\pi$ (e.g., $\Delta_i < 60°$). The peak widths in most of the initial distributions ranged from $\Delta_i = 25°$ to $\Delta_i = 40°$. This range of widths was chosen on the basis of the molecular dynamics study by McCammon and Karplus,[9] which showed that the $\chi_1$ angle in an amino acid from a small protein (residue 35 in the bovine pancreatic trypsin inhibitor) fluctuates within $\pm30°$.

**2.3. Karplus Equations.** The vicinal couplings across the $C^\alpha$–$C^\beta$ bond included in these calculations were $^3J(H^\alpha\text{–}H^{\beta_2})$, $^3J(H^\alpha\text{–}H^{\beta_3})$, $^3J(N\text{–}H^{\beta_2})$, $^3J(N\text{–}H^{\beta_3})$, $^3J(C'\text{–}H^{\beta_2})$, and $^3J(C'\text{–}H^{\beta_3})$. In some cases only the homonuclear coupling constants $^3J(H^\alpha\text{–}H^{\beta_2})$ and $^3J(H^\alpha\text{–}H^{\beta_3})$ were considered in order to simulate systems where only proton spectra are available. The angular dependence of each coupling constant was computed according to the appropriate Karplus equation[10]

$$^3J(\text{H–H}) =$$
$$(9.4 \pm 0.3)\cos^2\theta - (1.4 \pm 0.2)\cos\theta + (1.6 \pm 0.3)\text{ Hz} \tag{5}$$

$$^3J(\text{N–H}) =$$
$$(-3.75 \pm 0.01)\cos^2\theta + 0.26\cos\theta - (0.54 \pm 0.50)\text{ Hz} \tag{6}$$

$$^3J(\text{C'–H}) = 7.2\cos^2\theta - 2.04\cos\theta + (0.6 \pm 0.11)\text{ Hz} \tag{7}$$

where the dihedral angle $\theta$ between the coupled nuclei can be expressed in terms of the torsion angle $\chi_1$ as $\theta = \chi_1 + \Omega$. The phase factors for the atom pairs involved are $\Omega(H^\alpha\text{–}H^{\beta_2}) = \Omega(N\text{–}H^{\beta_3}) = -120°$, $\Omega(H^\alpha\text{–}H^{\beta_3}) = \Omega(C'\text{–}H^{\beta_2}) = 0°$, and $\Omega(N\text{–}H^{\beta_2})$
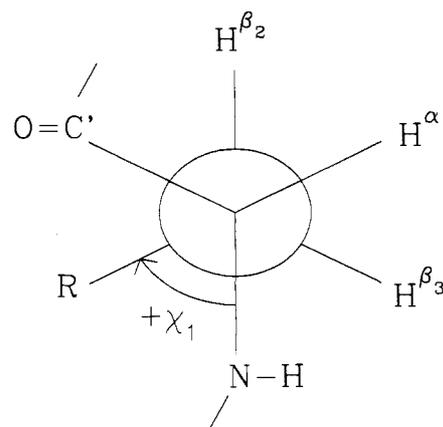


**Figure 1.** Definition of the torsion angle $\chi_1$ in an L-amino acid side chain. The phase angle, $\Omega$, from eq 6 of the preceding article[1] relates $\chi_1$ to the dihedral angle between two coupled nuclei. For example, $\Omega$ for the coupling between $H^\alpha$ and $H^{\beta_2}$ is $0°$; $\Omega$ for the coupling between $H^\alpha$ and $H^{\beta_3}$ is $-120°$.

**Table I.** Cartesian Coordinates of the Fixed Backbone Protons[a] That Give Rise to NOEs with $H^{\beta_2}_i$ and $H^{\beta_3}_i$

| atom | $x$ (Å) | $y$ (Å) | $z$ (Å) |
|---|---|---|---|
| $H^\alpha_{i-3}$ | -3.22 | -1.59 | 1.85 |
| $H^N_i$ | -2.20 | -0.03 | -1.36 |
| $H^\alpha_i$ | 0.53 | 0.88 | -1.89 |
| $H^N_{i+1}$ | -1.05 | -2.47 | -1.91 |

[a] The source of the atomic coordinates and the definition of the coordinate system are given in the text.

$= \Omega(C'\text{–}H^{\beta_3}) = +120°$ (see Figure 1). By introducing errors we have attempted to account for both experimental and theoretical uncertainties. Experimental errors were modeled by using Gaussian probability distributions with standard deviations amounting to $\Delta J(\text{H–H}) = 0.3$ Hz, $\Delta J(\text{N–H}) = 0.5$ Hz, and $\Delta J(\text{C'–H}) = 0.11$ Hz (these are the largest values reported as uncertainties for the Karplus parameters).[10] Introduction of larger errors ($\pm1$ Hz) did not change the reconstructed probability distributions significantly. Actual experimental errors in relevant biological systems can be $\leq0.5$ Hz.[7]

**2.4. Rotational Averaging of Nuclear Overhauser Enhancements.** In calculating average NOEs, it was assumed that the backbone of the $\alpha$-helix is rigid. The $i$th position in the primary structure of the helix was assigned to the amino acid whose $\chi_1$ rotamers were examined. The Cartesian coordinates of fixed atoms $H^\alpha_{i-3}$, $H^N_i$, $H^\alpha_i$, and $H^N_{i+1}$ were obtained from the 1.8-Å resolution X-ray structure of the Asn[36]-Phe[37]-Cys[38]-Asn[39]-Ala[40] pentapeptide from the $\alpha$-helical segment of the third domain of turkey ovomucoid.[11] Asn[39] was treated as the "$i$th" residue, whose internal rotation was simulated. The $H^\alpha$ of Asn[36] was considered as $H^\alpha_{i-3}$. Coordinates of the amide proton of Ala[40] were used for $H^N_{i+1}$. The origin of the coordinate system was placed at the $C^\beta_i$ carbon. The negative $z$ axis passed through $C^\alpha_i$, and $N_i$ was placed in the $xz$ plane.[12] In this coordinate system, the distance between any of the fixed atoms ($H^\alpha_{i-3}$, $H^N_i$, $H^\alpha_i$, or $H^N_{i+1}$) and one of the $\beta$-protons in the $i$th residue ($H^{\beta_2}_i$ or $H^{\beta_3}_i$) is

$$r(\chi_1) = \{[x - R_{\text{CH}}\cos(\chi_1 + \omega_j)\sin(\zeta_{\text{CCH}})]^2 +$$
$$[y - R_{\text{CH}}\sin(\chi_1 + \omega_j)\sin(\zeta_{\text{CCH}})]^2 +$$
$$[z - R_{\text{CH}}\cos(\zeta_{\text{CCH}})]^2\}^{1/2} \tag{8}$$

where $x$, $y$, and $z$ are the Cartesian coordinates of the fixed atom (Table I), $R_{\text{CH}}$ is the $C^\beta$–$H^\beta$ bond length (1.1 Å),[11] $\zeta_{\text{CCH}}$ is the $C^\alpha$–$C^\beta$–$H^\beta$ valence angle (109.4°),[11] and the phase factor $\omega_j$ ($j = 2$ or 3) is defined in the same way as described for coupling between the N and $\beta$-protons, where $\omega_j = \Omega(\text{N–H}^{\beta_j})$.

(9) McCammon, J. A.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1979**, *76*, 3585–3589.

(10) Fischman, A. J.; Live, D. H.; Wyssbrod, H. R.; Agosta, W. C.; Cowburn, D. *J. Am. Chem. Soc.* **1980**, *102*, 2533–2539.

(11) Fujinaga, M.; Sielecki, A. R.; Read, R. J.; Ardelt, W.; Laskowski, M., Jr.; James, M. N. G. *J. Mol. Biol.* **1987**, *195*, 397–418.

(12) The torsion angle $\chi_1$ is measured from the $x$ axis (from the plane $N_i$–$C^\alpha_i$–$C^\beta_i$ to the plane $C^\alpha_i$–$C^\beta_i$–$C^\gamma_i$ in the positive mathematical direction).

**Table II.** Fourier Coefficients[a] $C_i$ and $S_i$ of the Functions $r(\chi_1)^{-3}$ for the Eight Pairs of Atoms That Give Observable NOEs in an $\alpha$-Helix

| NOE | $C_0$ (Å$^{-3}$) | $C_1$ (Å$^{-3}$) | $S_1$ (Å$^{-3}$) | $C_2$ (Å$^{-3}$) | $S_2$ (Å$^{-3}$) | $C_3$ (Å$^{-3}$) | $S_3$ (Å$^{-3}$) |
|---|---|---|---|---|---|---|---|
| $H^\alpha_{i-3}$–$H^{\beta 2}_i$ | 0.0193 | 0.0009 | 0.0139 | −0.0042 | 0.0006 | −0.0002 | −0.0012 |
| $H^\alpha_{i-3}$–$H^{\beta 3}_i$ | 0.0193 | 0.0116 | −0.0077 | 0.0016 | −0.0039 | −0.0002 | −0.0012 |
| $H^N_i$–$H^{\beta 2}_i$ | 0.0500 | 0.0197 | 0.0353 | −0.0073 | 0.0118 | −0.0045 | −0.0002 |
| $H^N_i$–$H^{\beta 3}_i$ | 0.0500 | 0.0208 | −0.0347 | −0.0066 | −0.0122 | −0.0045 | −0.0002 |
| $H^\alpha_i$–$H^{\beta 2}_i$ | 0.0559 | 0.0122 | −0.0221 | −0.0025 | −0.0040 | −0.0008 | 0 |
| $H^\alpha_i$–$H^{\beta 3}_i$ | 0.0559 | −0.0252 | 0.0005 | 0.0048 | −0.0002 | −0.0008 | 0 |
| $H^N_{i+1}$–$H^{\beta 2}_i$ | 0.0242 | −0.0094 | 0.0124 | −0.0012 | −0.0041 | 0.0010 | 0.0004 |
| $H^N_{i+1}$–$H^{\beta 3}_i$ | 0.0242 | 0.0154 | 0.0019 | 0.0041 | 0.0010 | 0.0010 | 0.0004 |

[a] See eq 11a of the preceding article.[1]

**Table III.** Fourier Coefficients of the Functions $r(\chi_1)^{-6}$ for the Eight Pairs of Atoms That Give Observable NOEs in an $\alpha$-Helix[a]

| NOE | $c_0$ (Å$^{-6}$) | $c_1$ (Å$^{-6}$) | $s_1$ (Å$^{-6}$) | $c_2$ (Å$^{-6}$) | $s_2$ (Å$^{-6}$) | $c_3$ (Å$^{-6}$) | $s_3$ (Å$^{-6}$) |
|---|---|---|---|---|---|---|---|
| $H^\alpha_{i-3}$–$H^{\beta 2}_i$ | 0.0478 | 0.0040 | 0.0600 | −0.0277 | 0.0037 | −0.0022 | −0.0109 |
| $H^\alpha_{i-3}$–$H^{\beta 3}_i$ | 0.0478 | 0.0500 | −0.0334 | 0.0107 | −0.0258 | −0.0022 | −0.0109 |
| $H^N_i$–$H^{\beta 2}_i$ | 0.3423 | 0.2275 | 0.4082 | −0.1268 | 0.2049 | −0.1072 | −0.0049 |
| $H^N_i$–$H^{\beta 3}_i$ | 0.3423 | 0.2398 | −0.4011 | −0.1141 | −0.2122 | −0.1072 | −0.0049 |
| $H^\alpha_i$–$H^{\beta 2}_i$ | 0.3458 | 0.1426 | −0.2578 | −0.0464 | −0.0740 | −0.0218 | 0.0012 |
| $H^\alpha_i$–$H^{\beta 3}_i$ | 0.3458 | −0.2945 | 0.0054 | 0.0873 | −0.0032 | −0.0218 | 0.0012 |
| $H^N_{i+1}$–$H^{\beta 2}_i$ | 0.0714 | −0.0495 | 0.0657 | −0.0094 | −0.0330 | 0.0114 | 0.0044 |
| $H^N_{i+1}$–$H^{\beta 3}_i$ | 0.0714 | 0.0816 | 0.0100 | 0.0333 | 0.0083 | 0.0114 | 0.0044 |

[a] See eq 11b of the preceding article.[1]

The functions $r(\chi_1)^{-3}$ and $r(\chi_1)^{-6}$ for the eight pairs of protons that give observable NOEs in an $\alpha$-helix were expanded into a Fourier series for rotation about $\chi_1$; the calculated Fourier coefficients are listed in Tables II and III. The NOEs were represented as percentages of $\text{NOE}_{\text{calibr}}$, which corresponds to the $H^{\beta 2}$–$H^{\beta 3}$ pair. The value of $r_{\text{calibr}} = r(H^{\beta 2}–H^{\beta 3}) = 1.75$ Å was assumed.[13]

The experimental errors for NOEs were simulated by generating random numbers according to a Gaussian distribution with standard deviations ranging from 0.3% to 0.5% $\text{NOE}_{\text{calibr}}$ for $\langle r^{-3}\rangle^2$ averaging and from 0.7% to 1% $\text{NOE}_{\text{calibr}}$ for $\langle r^{-6}\rangle$ averaging. The NOEs that are relevant to our study range from < 1% to 30% $\text{NOE}_{\text{calibr}}$; these are equivalent to errors of 3% for strong NOEs or 100% or more for weak NOEs. A common way of measuring distances using NOEs is to collect several spectra with different mixing times. Each spectrum gives rise to a point on a curve (the "build-up curve") whose initial slope defines a distance. Here, a series (typically seven) of data points were generated that corresponded to NOE measurements with different mixing times (different points on the NOE build-up curve).

**2.5. Gaussian Parameters for Artificial Probability Distributions.** After applying CUPID to the simulated data to obtain the trigonometric parameters $\rho_1$, $\sigma_1$, $\rho_2$, $\sigma_2$, $\rho_3$, and $\sigma_3$, the Gaussian parameters (three peak positions, two independent probabilities, and widths, which were constrained to be equal for all three peaks) were fitted to the trigonometric parameters as described in the preceding paper.[1]

Measured $J(H^\alpha$–$H^{\beta 2})$ and $J(H^\alpha$–$H^{\beta 3})$ vicinal couplings in amino acid residues, when plotted on a correlation diagram,[14,15] can indicate that the actual probability distribution consists of a single peak (or one predominant peak) if the data point is located close to the border of the diagram. In this case, the number of fitted Gaussian curves can be reduced from three to one without risk of generating an artificial average structure devoid of physical meaning (such as the virtual average conformers critically examined by Jardetzky[16]). As a consequence, only two parameters were evaluated for single-peak distributions: the position and the width of the Gaussian bell.

**3. Results**

**3.1. Torsion Rotation about the $\chi_1$ Dihedral Angle in L-Leucine.** For the analysis of $\chi_1$ rotamers of L-leucine (cation and anion), we chose the Karplus coefficients used by Fischman et al.[10] so

(13) Braun, W.; Bösch, C.; Brown, L. R.; Gō, N.; Wüthrich, K. *Biochim. Biophys. Acta* **1981**, *667*, 377–396.
(14) Nagayama, K.; Wüthrich, K. *Eur. J. Biochem.* **1981**, *115*, 653–657.
(15) Wagner, G.; Wüthrich, K. *Methods Enzymol.* **1986**, *131*, 307–326.
(16) Jardetzky, O. *Biochim. Biophys. Acta* **1980**, *621*, 227–232.

**Table IV.** Average Vicinal Couplings (Hz) across the $C^\alpha$–$C^\beta$ Bond in L-Leucine

| coupled atoms | exptl (cation)[a] | theoret (cation)[b] | exptl (anion)[a] | theoret (anion)[b] | theoret (anion)[c] |
|---|---|---|---|---|---|
| $H^\alpha$–$H^{\beta 2}$ | 8.42 | 8.30 | 8.76 | 8.57 | 8.58 |
| $H^\alpha$–$H^{\beta 3}$ | 5.98 | 5.66 | 5.94 | 5.78 | 5.90 |
| N–$H^{\beta 2}$ | −2.47 | −1.97 | −2.15 | −1.90 | −1.79 |
| N–$H^{\beta 3}$ | −3.47 | −3.16 | −3.15 | −3.21 | −3.21 |
| $C'$–$H^{\beta 2}$ | 3.75 | 3.85 | 4.06 | 4.04 | 3.90 |
| $C'$–$H^{\beta 3}$ | 2.92 | 2.86 | 2.36 | 2.27 | 2.37 |

[a] Fischman et al.[2]  [b] Probability distribution as the sum of two Gaussians.  [c] Probability distribution obtained by optimizing simultaneously $\rho_1$, $\sigma_1$, $\rho_2$, $\sigma_2$, $\rho_3$, and $\sigma_3$ with the nonnegativity condition incorporated in the fitting procedure.
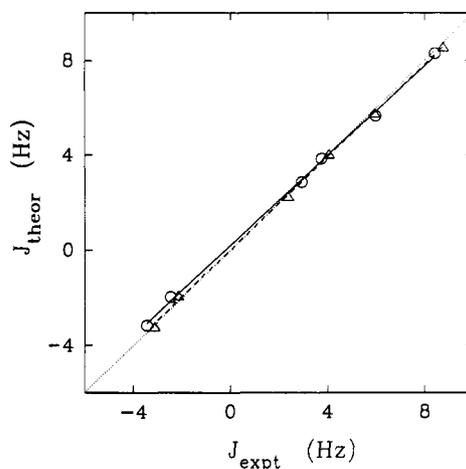


**Figure 2.** Correlation between the theoretical and experimental coupling constants across the $C^\alpha$–$C^\beta$ bond in L-leucine (Table IV). The solid line corresponds to a linear regression analysis of the cation data (circles: $J_{\text{theor}} = J_{\text{expt}} \times 0.9441 + 0.2085$, $r^2 = 0.9989$). The dashed line corresponds to a linear regression analysis of the anion data (triangles: $J_{\text{theor}} = J_{\text{expt}} \times 0.9760 + 0.0174$, $r^2 = 0.9994$). The dotted line is drawn through the origin at 45° from the $x$-axis for reference. Experimental data are from ref 2.

that our results could be compared directly to theirs. Experimental values for six vicinal couplings across the $C^\alpha$–$C^\beta$ bond in L-leucine[2] were used as input for this analysis. These values are given in Table IV, along with theoretically derived values based on CUPID analysis. The correlation between theoretical and experimental coupling values is illustrated in Figure 2. Points on this plot are aligned along straight lines with slopes equal to 0.9441 (cation:

**Table V.** Fourier Coefficients $\rho_i$ and $\sigma_i$ ($i$ = 1, 2, 3) of the Probability Distributions of Rotamers about the $C^\alpha$–$C^\beta$ Bond in L-Leucine and Corresponding Regression Coefficients $r^2$ for Coupling Constants

|  | $\rho_1$ | $\sigma_1$ | $\rho_2$ | $\sigma_2$ | $\rho_3$ | $\sigma_3$ | $r^2$ |
|---|---|---|---|---|---|---|---|
| cation[a] | −0.0446 | −0.0884 | −0.0566 | −0.1048 | | | 0.9989 |
| anion[a] | −0.0785 | −0.1811 | −0.0584 | −0.1030 | | | 0.9994 |
| anion[b] | −0.0036 | −0.1479 | −0.0283 | −0.1186 | −0.0825 | 0.0395 | 0.9989 |

[a] Fourier coefficients were obtained by linear regression; the probability distribution was then reconstructed as a sum of two Gaussians. [b] Probability distribution was obtained by optimizing simultaneously $\rho_1$, $\sigma_1$, $\rho_2$, $\sigma_2$, $\rho_3$, and $\sigma_3$ with the nonnegativity condition incorporated in the fitting procedure.

**Table VI.** Gaussian Coefficients of the Angular Probability Distributions for Internal Rotation about the $C^\alpha$–$C^\beta$ Bond in L-Leucine

|  | $\varphi_1$ (deg) | $\varphi_2$ (deg) | $p_1$ | $p_2$ | $\Delta$ (deg) |
|---|---|---|---|---|---|
| cation | −216 | −74.2 | 0.406 | 0.594 | 49.5 |
| anion | −185 | −77.6 | 0.382 | 0.618 | 5.55 |
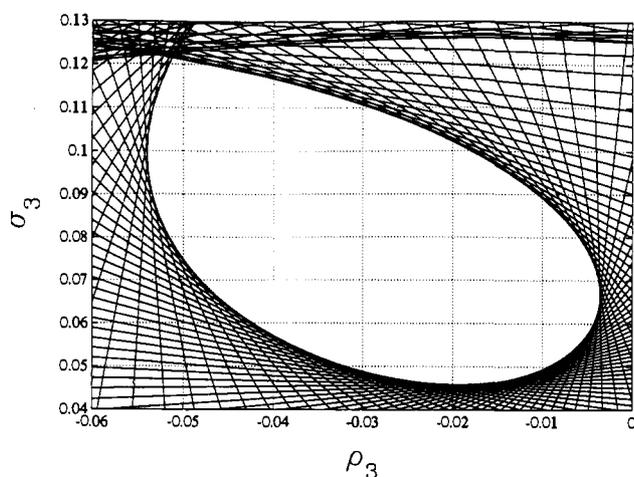


**Figure 3.** The region in the $\rho_3$–$\sigma_3$ plane which corresponds to positive $\rho(\chi_1)$ for any value of $\chi_1$ in L-leucine (cation), according to eq 32 of the preceding paper;[1] values of $\rho_1$, $\sigma_1$, $\rho_2$, and $\sigma_2$ are given in Table V.

circles, solid line) and 0.9760 (anion: triangles, broken line). With perfect correlation, the slope would be unity. The regression coefficients $r^2$ listed in Table V show a high degree of correlation between the theoretical and experimental values. The coefficients $\rho_1$, $\sigma_1$, $\rho_2$, and $\sigma_2$ of the Fourier series for $\rho(\chi_1)$, obtained using linear regression (eq 25 from the preceding paper[1]) are listed in Table V.

The ranges for values of coefficients $\rho_3$ and $\sigma_3$ were estimated for the cation data under the assumption that fourth- and higher-order terms are negligible. A diagram of $\rho_3$ and $\sigma_3$ values that do not violate nonnegativity of $\rho(\chi_1)$ was constructed according to eq 32 of the preceding paper[1] and is shown in Figure 3. By sampling the central (nonshaded) region of Figure 3, we obtained the family of curves $\rho(\chi_1)$ shown in Figure 4A (solid lines). All the curves in this family are qualitatively similar. This confirms that the approach used to compute $\rho_3$ and $\sigma_3$ values gave consistent results for the L-leucine cation.

The coefficients $\rho_1$, $\sigma_1$, $\rho_2$, and $\sigma_2$ from cation data (Table V) were also used to find the Gaussian parameters (Table VI) and to construct the probability distribution $\rho(\chi_1)$ as a sum of two Gaussian functions (Figure 4A, dashed curve). As illustrated by Figure 4A, the Gaussian curve falls within the family of curves obtained by neglecting the fourth- and higher-order terms (Figure 4A, solid lines); thus the assumption that these terms can be neglected in the case of the cation appears justified.

In the case of the anion, it was not possible to use the nonnegativity condition to find a restricted range of third-order coefficients $\rho_3$ and $\sigma_3$. Instead, the first- and second-order coefficients were used to determine Gaussian parameters (Table VI). The sum of the Gaussian curves corresponding to these parameters is shown in Figure 4B. The peaks in Figure 4B are much narrower than those of the Gaussian distribution for the cation (Figure 4A). This indicates that higher-order terms cannot be neglected.
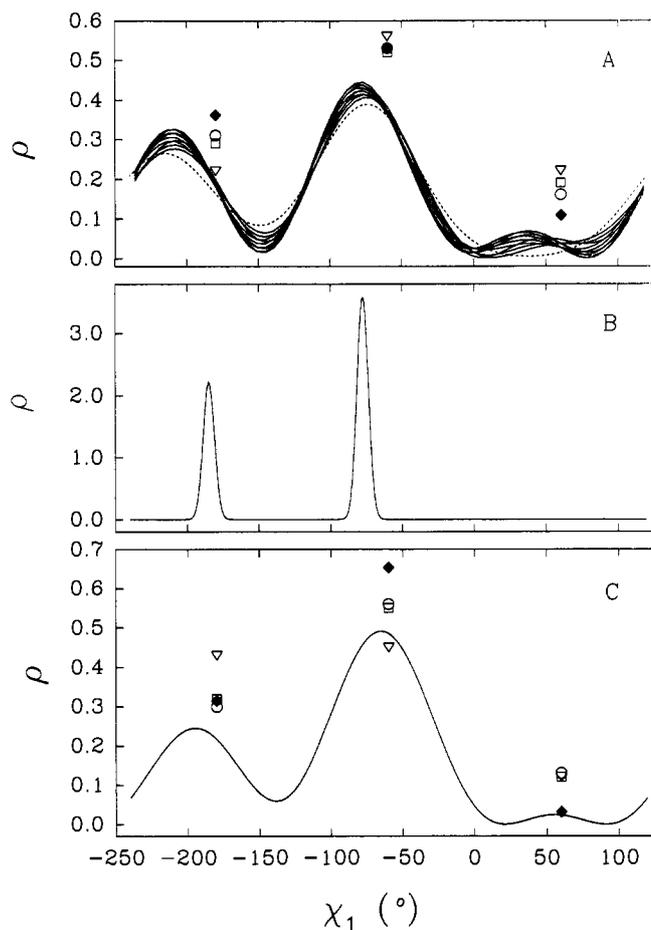


**Figure 4.** (A) Comparison of probability distributions across the $C^\alpha$–$C^\beta$ bond in L-leucine (cation) at 19 °C. The dashed line is a sum of two Gaussians whose positions, probabilities, and widths are obtained as described in section 2.5 and tabulated in Table VI. The solid lines correspond to the family of curves obtained from the non-negativity-based method.[1] The coefficients $\rho_1$, $\sigma_1$, $\rho_2$, and $\sigma_2$ are the same for the family of solid curves and for the sum of Gaussians, and their values are given in Table V. Values of $\rho_3$ and $\sigma_3$ that satisfy the condition of nonnegativity correspond to points on the grid inside the central (nonshaded) region of Figure 3. Each ($\rho_3$, $\sigma_3$) pair gives rise to a single curve. Fourth- and higher-order terms are set to zero in the case of the family of solid curves. The hollow circles, triangles, and squares correspond to discrete populations obtained from $H^\alpha$–$H^\beta$, $N$–$H^\beta$, and $C'$–$H^\beta$ couplings, respectively.[2] The filled diamonds correspond to integrals of the continuous distribution derived from the spin–spin couplings, as described in the text. (B) Probability distribution $\rho(\chi_1)$ in L-leucine (anion) obtained as a sum of two Gaussians. The positions, probabilities, and widths of the Gaussians are given in Table VI. (C) Comparison of probability distributions across the $C^\alpha$–$C^\beta$ bond in L-leucine (anion). The solid curve $\rho(\chi_1)$ was obtained by fitting the first three orders of Fourier coefficients with the nonnegativity condition incorporated in the fitting procedure. The best-fit trigonometric coefficients are tabulated in Table V. The hollow circles, triangles, squares, and the filled diamonds have the same meaning as in Figure 4A.

The nonlinear fitting procedure, which incorporates the nonnegativity condition (eq 32 of the preceding paper[1]) was employed as an alternative way to find the angular probability distribution for the anion. Its results (Figure 4C) are unique in the sense that

they do not depend on the initial guess. The convergence was rapid over a wide range of initial values of the fitted Fourier coefficients, and all final values were the same as those given in Table V. The positions and relative heights of the peaks of that probability distribution are similar to those of the sum of Gaussians for the anion (Figure 4B), but the peaks of the trigonometric distribution for the anion are considerably wider than the corresponding Gaussian peaks. This is a consequence of truncating the Fourier expansion at the third order while simultaneously requiring that the truncated Fourier series remain nonnegative. Table V shows that the relative values of the Fourier coefficients $\rho_1$, $\sigma_1$, $\rho_2$, and $\sigma_2$ obtained by nonlinear fitting of the anion coupling data roughly parallel those obtained from the same data by linear regression. Since the nonlinear fitting procedure scales down the absolute values of the Fourier coefficients, by forcing the truncated Fourier series to satisfy the nonnegativity condition, it causes peak broadening.

**3.2. Simulated Data.** Four general categories of initial simulated distributions were considered: (a) two or three populations of varied size at the normal staggered rotamer positions ($t = -180°$, $g^+ = 60°$, $g^- = -60°$; Figure 5, insets 1–9); (b) two or three populations of varied size at nonstaggered angles with 120° spacing between the centers of the populations (Figure 5, insets 10–15); (c) two or three populations of varied size with random angles (Figure 6, insets 16–26); and (d) a single population at a fixed angle with different widths (Figure 6, insets 27–30).

With error-free input data, CUPID reproduced the parameters of the initial distributions in all cases examined. Thus the solid lines in Figures 5 and 6 show the initial distributions and the (identical) distributions reproduced by Fourier analysis of error-free data. In the absence of experimental errors (including parameterization of the Karplus equation), the only factors found to limit the precision of CUPID were the tolerance of the numerical procedure for solving the set of nonlinear equations and the accuracy of the numerical integrations. These factors led to such a small difference between initial and reproduced parameters that they can be neglected. Moreover, with error-free data the reconstructed probability distributions are the same for both $\langle r^{-3}\rangle^2$ and $\langle r^{-6}\rangle$ averaging of the NOEs.

Probability distributions reconstructed from data containing simulated errors are represented (Figures 5 and 6) by dashed lines ($\langle r^{-3}\rangle^2$ averaging of NOEs) and dotted lines ($\langle r^{-6}\rangle$ averaging). Approximately two-thirds of the distributions shown in Figures 5 and 6 were reproduced to a quantitative accuracy of about ±20° in position and to within about ±10% accuracy in width and probability for both slow and fast internal motions. In about half of these, the accuracy was much higher. Although the remaining distributions reproduced qualitative features, they deviated quantitatively from the initial simulated distributions in one or more of the following ways: (1) peak doubling (Figures 5 and 6: 3, 9, 15, 16, 17, and 23); (2) peak narrowing (Figures 5 and 6: 3, 9, 13, 14, 15, 16, 17, 23, and 30); (3) shifts in position (Figures 5 and 6: 14, 22, 24, 25); (4) regions of negative probability distribution when parameters $p_i$ were not forced to be positive (Figures 5 and 6: 2, 22); (5) poorly defined peaks (Figure 6: 21, 22, 24). These deviations are discussed below.

## 4. Discussion

**4.1. $\chi_1$ Rotamers of L-Leucine.** The family of continuous curves obtained by CUPID for the cation (Figure 4A) agree qualitatively with the population distribution derived earlier from the discrete model.[2] The angles at which the continuous curves show maxima (Figure 4A) approximate the staggered conformations assumed in the discrete model. The largest peak in the probability distribution for the cation ($\chi_1 \approx -70°$) is only 10° away from the corresponding staggered angle ($\chi_1 \approx -60°$). However, the second largest peak in the probability distribution for the cation ($\chi_1 \approx -210°$) is 30° from the value assumed in the discrete model ($\chi_1 \approx -180°$), and the smallest peak ($\chi_1 = +40°$) is 20° from the staggered conformation ($\chi_1 \approx +60°$).

Can the continuous probability distribution $\rho(\chi_1)$ be compared directly with the discrete probability distribution? If the two are

to be consistent, integrals of $\rho(\chi_1)$ over the intervals (−240°, −120°), (−120°, 0°), and (0°, 120°) (Figure 4, A and C, filled diamonds) should reproduce the calculated values of the discrete probabilities (Figure 4, A and C, open symbols). The agreement is quite good for both cation and anion. Note that the values of the integrals do not depend on the coefficients $\rho_3$ and $\sigma_3$. Thus the whole family of continuous curves $\rho(\chi_1)$ for the cation yields a unique set of three areas (Figure 4A, filled diamonds).

The method described in section 2.1 gives a family of distributions for the cation rather than one single curve. Despite this, the uncertainty in the probability at a given angle is similar to that of the discrete model. The spread of discrete values corresponding to different couplings is comparable with the vertical width of the family of continuous curves (or even larger as in the region around $\chi_1 \approx -180°$).

The smallest peak in all probability distributions, located near $\chi_1 = +60°$, corresponds to a conformation in which the bulky side chain of leucine has closest contact with both the N and C' groups. Strong repulsions are expected to lead to a smaller probability for this rotamer than for the other two. Both the discrete and continuous model reproduce this feature of L-leucine (both for the cation and anion).

The agreement found for leucine between the continuous and discrete methods is not unexpected, since $\chi_1$ for a free amino acid probably has "ethane-like" rotational barriers. Such agreement, however, would be less likely for more complex molecules such as peptides and proteins that exhibit hindered rotations.

**4.2. Simulated Distributions.** The ultimate value of any analytical method depends upon its accuracy. In this section we develop a set of empirical guidelines from which a "confidence level" can be defined. To establish these guidelines, we must examine the deviations from the simulated distributions described above. These deviations all arise as a result of error. As noted above, a given amount of experimental error will produce a larger deviation in the fast internal motion case than in the slow internal motion case. This is because the higher-order coefficients of the Fourier expansion have smaller values with $\langle r^{-3}\rangle^2$ averaging[17] than with $\langle r^{-6}\rangle$.

In all cases where the full, error-free, data set (six coupling constants and eight NOEs) was considered, CUPID reproduced the initial distribution exactly. Although the error-free case is not realistic experimentally, it is important because it establishes the credibility of the CUPID approach. When errors were included, at least qualitatively correct distributions were obtained in all cases. It is important to note that many of the cases examined in this study would be "invisible" to discrete analysis. This is best seen in examples 21, 22, and 24 (Figure 6); these distributions were classified above as "poorly defined peaks". However, each one of the cases reproduced a distribution of two or three closely spaced peaks. Relatively large deviations were observed between initial and calculated distributions, in cases that had one peak of low probability. These results suggest that caution should be exercised in interpreting distributions that contain small peaks. The results show that peak doubling can occur in situations where no peaks of low probability are present. Inspection of Figures 5 and 6 shows that most distributions with doubled peaks had unusually narrow peaks (<20°). In addition, all of the doubled peaks occurred within about 60–90° of one another (with the correct population-weighted average in the center). From this, we conclude that calculated distributions with narrow (i.e., less than 20°) peaks and/or two closely spaced peaks must be treated qualitatively.

The CUPID results improved significantly when larger numbers of coupling constants and when NOE data from multiple mixing times were used as input. The reliability of the final probability distribution (derived from data with relatively large experimental errors) was found to increase dramatically when the number of mixing times (points on the buildup curve) increased only slightly (from five to seven). The improvement results from the averaging

---

(17) Figure 5 does not show this effect because smaller errors were used in generating the $\langle r^{-3}\rangle^2$ data than the $\langle r^{-6}\rangle$ data.
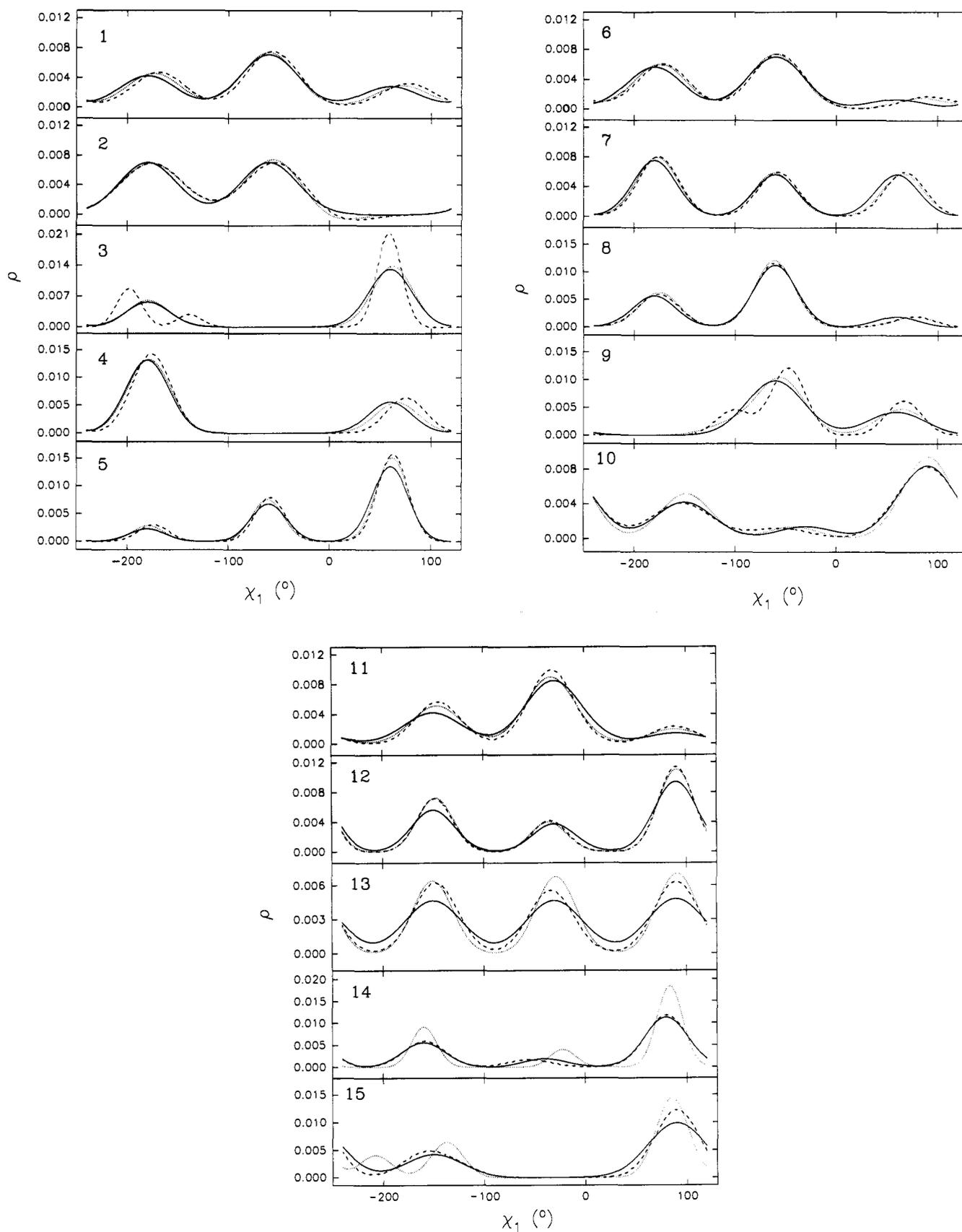
**Figure 5.** Assumed initial distributions and distributions calculated by CUPID analysis of NOEs and coupling constants: insets 1–9, probability distributions with peaks at staggered angles; insets 10–15, distributions with 120° spacings between peaks. The initial probability distributions are indicated by continuous curves. Dashed lines represent distributions reproduced under the assumption that the internal rotation is fast compared to overall tumbling (averaging of NOEs occurs as $\langle r^{-3} \rangle^2$). Dotted lines represent distributions reconstructed for slow internal motion ($\langle r^{-6} \rangle$ averaging of NOEs).
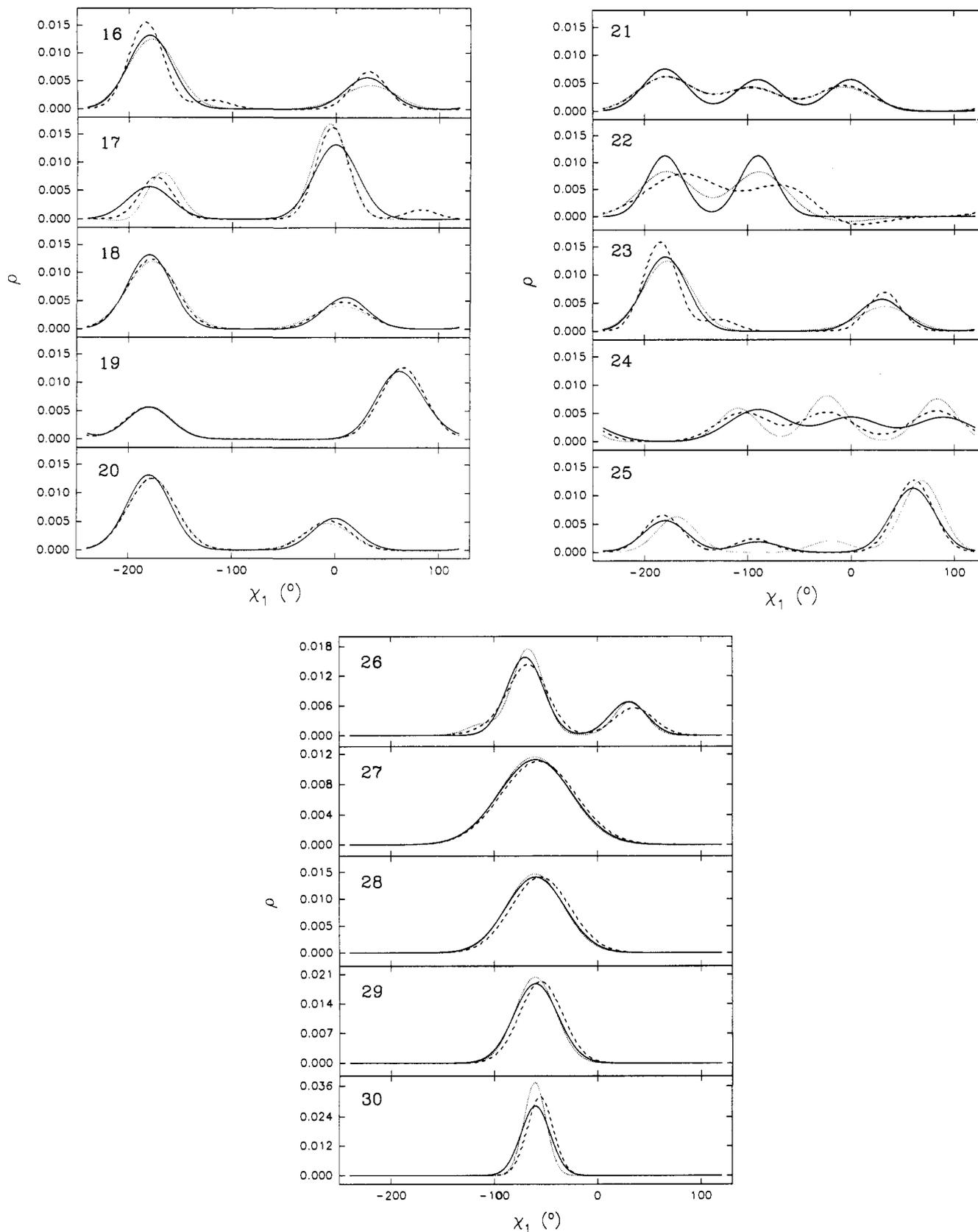
**Figure 6.** Assumed initial distributions and distributions calculated by CUPID analysis of NOEs and coupling constants: Insets 16–26, irregular distributions; insets 27–30, single-peak distributions. The initial probability distributions are indicated by continuous curves. Dashed lines represent distributions reproduced under the assumption that the internal rotation is fast compared to overall tumbling (averaging of NOEs occurs as $\langle r^{-3}\rangle^{2}$). Dotted lines represent distributions reconstructed for slow internal motion ($\langle r^{-6}\rangle$ averaging of NOEs).

of random errors in individual measurements.

The third-order Fourier coefficients of the functions $r(\chi_1)^{-3}$ and $r(\chi_1)^{-6}$ ($C_3$, $S_3$, $c_3$, and $s_3$; see eqs 11a,b of the preceding paper[1]) are smaller than the first- and second-order terms (see Tables II

and III). Therefore, $\rho_3$ and $\sigma_3$ are less well defined by the experimental data than coefficients $\rho_1$, $\sigma_1$, $\rho_2$, and $\sigma_2$. This problem is particularly prominent when the experimental error is large. It is even more pronounced when averaging of NOEs occurs as

$\langle r^{-3}\rangle^2$, since the angular dependence of $r^{-6}$ is sharper and the corresponding $c_3$ and $s_3$ terms are larger. Moreover, in the specific example studied here (rotation about $\chi_1$ in an $\alpha$-helical environment), the $S_3$ terms are even smaller than the $C_3$ terms (particularly in the case of distances between $H^\alpha_i$ and both $H^\beta_i$ protons; see Tables II and III), and $\sigma_3$ is extremely sensitive to experimental errors. In practice, if the data contain errors comparable to $S_3$, one can first find the coefficients $\rho_1$, $\sigma_1$, $\rho_2$, $\sigma_2$, and $\rho_3$, and then estimate $\sigma_3$ from the nonnegativity condition (eq 32 of the preceding paper[1]). If the experimental errors are larger than $C_3$, then both $\rho_3$ and $\sigma_3$ must be evaluated in this way, and the resulting distribution must be considered qualitative at best.

There may be situations, such as that presented by the leucine anion, in which to evaluate $\rho_3$ and $\sigma_3$ one has to resort either (1) to fitting $\rho_1$, $\sigma_1$, $\rho_2$, $\sigma_2$, $\rho_3$, and $\sigma_3$, with the nonnegativity condition incorporated into the nonlinear fitting procedure, or (2) to fitting a sum of two Gaussian functions. Such situations arise when the experimental data are such that the $\rho_3$ and $\sigma_3$ terms are not defined (for example, when only coupling constants are known and only $\rho_1$, $\sigma_1$, $\rho_2$, and $\sigma_2$ can be derived from the measured data), and when the distribution based on these values does not satisfy the nonnegativity condition for any pair of $\rho_3$ and $\sigma_3$ values.

A two-step procedure was applied in this work. The first step, a linear regression, yields the values of the trigonometric coefficients $\rho_1$, $\sigma_1$, $\rho_2$, $\sigma_2$, $\rho_3$, and $\sigma_3$. The second step replaces the truncated trigonometric series by a sum of Gaussians. One could argue that our aim, at least in principle, could have been achieved in a single step by fitting a sum of Gaussians to the data. This conceptually simpler approach was used previously by Shirmer et al.[18] and by Hart and Davis.[19] However, our experience shows that the two-step procedure is superior to the direct one, since the latter often leads to artifacts such as large, narrow "ghost" peaks at false positions.[20] By breaking the process into two steps, the linear regression ensures that the best possible values are found for the trigonometric parameters. The problem of nonlinearity is transferred completely to the final stage, in which a nonlinear fit of Gaussian parameters is performed. This nonlinear fit is aided by the initial estimate, which is suggested by the shape of the truncated trigonometric series. The direct approach, on the other hand, can use only randomly selected initial values for the Gaussian parameters. Another drawback of the direct approach is that, since it does not take advantage of the linear regression, it does not provide a general way to verify whether the fit is really the best one. The approach based on the trigonometric series, on the other hand, always gives *the best fit*.

Interestingly, experimental couplings and nuclear Overhauser enhancements are not reproduced well when prochiral resonance assignments are inverted. This suggests a method for stereospecific assignment of prochiral substituents: the correct assignment is most probably the one that results in better agreement between calculated and experimental data.

## 5. Conclusion

CUPID has advantages over the discrete model used by other authors,[2,16,21-25] because it does not rely upon guesses of the torsional angles. Rather, the angles and probabilities are both found from the calculated distribution function. CUPID can identify nonstaggered rotamers, including those with totally irregular patterns of probability peak positions. We note that the energy of a hydrogen bond is of the same order of magnitude as many internal torsional barriers. Thus, it is likely that staggered conformations are sometimes violated for $\chi_1$ angles in proteins.

The application of CUPID to a variety of systems, such as peptides, proteins, sugars, and nucleic acids, should be straightforward. The basic procedure is to establish a coordinate system rotating with the molecule and then to define a set of internal coordinates across the dihedral angle of interest which describes the internal rotation. With this choice of coordinates, any internuclear distance can be determined geometrically and expanded into a Fourier series. The Karplus equations for spin–spin couplings are already equivalent to a truncated Fourier series. The CUPID equations, which are based on these facts, show that rotationally averaged distances and couplings are linear functions of the Fourier coefficients of the probability distribution. These coefficients can be determined from a linear combination of several measured spin–spin couplings and NOEs. In order to reproduce six Fourier coefficients of the probability distribution ($\rho_1$, $\sigma_1$, $\rho_2$, $\sigma_2$, $\rho_3$, and $\sigma_3$), it is necessary to have at least six measured rotationally averaged parameters. Further, in order to be able to determine $\rho_3$ and $\sigma_3$ directly from experimental data, at least two out of these six experimental values must be NOEs other than those from vicinal protons. The information obtained from spin–spin couplings is limited only to first- and second-order coefficients since Karplus equations do not contain third-order terms. NOEs between vicinal protons can give $\rho_3$, but fail to yield reliable $\sigma_3$ values since the sin $(3\chi_1)$ component of the $^1$H–C–C–$^1$H distance is small. In principle, all the experimental values can be NOEs. However, it is better to combine NOEs with at least two homonuclear proton–proton couplings, since CUPID is less vulnerable to experimental errors in couplings than in NOEs.

These are the minimal requirements for CUPID analysis, but none of the available data should be ignored. With more than six experimental values, one can take advantage of the linear regression. On the other hand, if the minimal requirements are not satisfied, one has to resort to a "lower level" CUPID analysis, in which the nonnegativity condition is exploited in order to predict possible values of the coefficients $\rho_3$ and $\sigma_3$ (see section 2.2.2 of the preceding article[1]). As demonstrated by our analysis of the $\chi_1$ rotamers in L-leucine, CUPID can give reasonable results even in this case.

The authors realize that the amount of data and effort needed for CUPID analysis are greater than those required for a discrete distribution model. This extra effort, however, provides results that are unbiased by assumptions concerning structure and motions. To our knowledge no previous method has been capable of simultaneously detecting nonstaggered conformations and rotational averaging. Deviations from staggered conformations are interesting and might be expected to play important roles in reactivity or molecular recognition. It is our hope that CUPID will lead to better use of data generated by the powerful array of experimental NMR techniques currently available.

(18) Schirmer, R. E.; Davis, J. P.; Noggle, J. H.; Hart, P. A. *J. Am. Chem. Soc.* **1972**, *94*, 2561–2572.

(19) Hart, P. A.; Davis, J. P. *J. Am. Chem. Soc.* **1972**, *94*, 2572–2577.

(20) A large and narrow peak similar to artifacts appearing in our own results (not shown here) is noticeable in Figure 7 of ref 18.

(21) Pople, J. A. *Mol. Phys.* **1958**, *1*, 3–8.

(22) Bystrov, V. L.; Portnova, S. L.; Balashova, T. A.; Tsetlin, V. I.; Ivanov, V. T.; Kostetzky, P. V.; Ovchinnikov, Yu. A. *Tetrahedron Lett.* **1969**, *59*, 5225–5228.

(23) Colucci, W. J.; Gandour, R. D.; Mooberry, E. S. *J. Am. Chem. Soc.* **1986**, *108*, 7141–7147.

(24) Cowburn, D.; Live, D. H.; Fischman, A. J.; Agosta, W. C. In *Intramolecular Dynamics*; Jortner, J., Pullman, B., Eds.; D. Reidel: New York, 1982; pp 473–480.

(25) Cowburn, D.; Live, D. H.; Fischman, A. J.; Agosta, W. C. *J. Am. Chem. Soc.* **1983**, *105*, 7435–7442.