The definition of a standard set of reference random coil chemical shift values is a key component in many applications of protein NMR spectroscopy. The comparison of measured chemical shifts with their random coil counterparts is commonly used to identify secondary structure elements in folded proteins and to reveal the presence of regions with residual structure in unfolded states. The importance of measuring backbone chemical shifts in unfolded states has recently been further increased with the recognition that proteins containing natively unfolded regions may represent up to one-third of eukaryotic proteomes and play a variety of essential biological roles; furthermore, it has also been realized that several amyloidogenic proteins associated with neurodegenerative diseases are natively unfolded.

Several methods for associating random coil chemical shift values to amino acid sequences of proteins based on experimental measurements of chemical shifts from model peptides that mimic the random coil state0–3 or derived by analysis of protein databases have been proposed. In this work, we present an approach called CamCoil, in which we map the relationship between amino acid sequences and chemical shifts using the flexible loop regions in native states as a model of the random coil state (Figure 1a). This strategy enables us to discriminate the dependence of the chemical shifts on the primary structure of proteins from the effects associated with the secondary and tertiary structures. The parameters were derived by statistical analysis of a recently constructed database of 1772 proteins for which structures and chemical shifts are known4 [see the Supporting Information (SI)]. From this database, we extracted for analysis fragments classified by STRIDE5 as loops (Figure 1a), i.e., not as α-, β-, or 310-helices, β-sheets, turns or bends; we further selected only flexible loops by including only residues with an RCI index6 smaller than 0.5 (Figure S1 in the SI). We first considered tripeptide fragments, since we expect the dominant sequence-dependent effects on the chemical shifts in a given amino acid to be due to the identities of its nearest neighbors. We thus can express the random coil (RC) chemical shift $\delta_{\text{RC}}^i$ of an atom of type $i$ in amino acid of type $A$ as

$$\delta_{\text{RC}}^i = \delta_{\text{A}}^0 + \alpha^- \delta_{\text{A}}^i + \alpha^+ \delta_{\text{AC}}^i$$

In this formula, the term $\delta_{\text{A}}^0$ represents the contribution due to the identity of the amino acid in which atom $i$ is present. The list of values for $\delta_{\text{A}}^0$ is provided in the form of residue-specific scales of chemical shifts for the nuclei $^{13}$C, $^{13}$C, $^{13}$C, $^{15}$N, $^1$H, and $^1$H (Table S1 in the SI). Nearest-neighbor effects are included through the $\delta_{\text{AC}}^i$ terms in eq 1; the $\delta_{\text{A}}^i$ and $\delta_{\text{AC}}^i$ terms represent the contributions from the flanking residues (of types $B$ and $C$, respectively).

The weights of these contributions are given by the parameters $\alpha^-$ and $\alpha^+$ (Table S2), which were optimized by applying a calibration procedure on five experimental data sets of random coil chemical shifts measured under conditions minimizing the presence of residual structure (see the SI). We found consistent results for weights calculated using independent data sets (Figure S2), thereby enabling a global optimization procedure (Figure 2). Thus, the hybrid parametrization that we carried out takes advantage of a large database of flexible native loops to obtain the main set of parameters and correction factors and employs data from unstructured proteins to calibrate the balance between these terms, with the aim of improving the predictions of the chemical shifts in random coil states.

The residue-specific $\delta_{\text{A}}^0$ values are already in good agreement with the experimental data for the five experimental random coil data sets that we considered (Figure S4). This correlation is comparable with that obtained using the method by Schwarzinger et al., although some differences exist between the two sets of values (Figure S5). When the sequence-specific correction factors are applied, the quality of the method increases significantly (Figure 1b and Figure S6). In all cases, the CamCoil root-mean-square (RMS) distances are smaller than the overall variability of the random coil chemical shift values in the data sets that we considered in this work (blue bars in Figure 1b). The analysis of the RMS
distance surface projected on the (α^–, α^+) space reveals that the
use of unitary weights for neighbor corrections is not the optimal
solution (Figure 2). To better account for sequence-dependent
effects on chemical shifts, in principle, we could use amino acid
triplets (or quintuplets and so on); however, larger databases would
be required to derive the corresponding parameters in these cases.
Here, in order to at least partially take into account next-nearest-
eighbor effects,^2 we considered two additional pairwise terms (eq
S1 in the SI).

The approach that we have presented, in which random coil
chemical shifts are determined by analyzing the amino acid
sequences in the loop regions in a database of known structures,^1
employed two sets of 400 correction factors in eq 1 and
determined, which is shown in Figure S7). Moreover, since
removing biases associated with specific experimental conditions
is important since the comparison of experimental chemical shifts
are determined by analyzing the amino acid
sequences in the loop regions in a database of known structures.

Another useful application of the CamCoil approach is the
prediction of chemical shifts of loops in native-state proteins (Figure
S9 and Table S3). In this case, the parameters α^– and α^+ in eq 1
are determined by optimizing the agreement between experimental
and predicted chemical shifts in native states of proteins (see the
SI).

The close agreement that we have presented between the
CamCoil random coil chemical shifts and the chemical shifts
measured experimentally for unfolded proteins (Figure 1 and Figure
S3) provides further support for the idea that it is possible to
describe fairly accurately random coil states by analyzing the loop
regions in folded structures.\textsuperscript{18}

In conclusion, we suggest that increasingly accurate random coil
chemical shift scales will be obtained through approaches of the
type that we have presented here by exploiting the continuous
growth of databases of protein structures and chemical shifts, which
will enable progressively more sophisticated functions to be
parametrized.

Acknowledgment. We thank G. G. Tartaglia for useful comments and discussions. We acknowledge support by EMBO (A.D.S.
and M.V.), Netherlands Ramsay (S.-T.D.H.), the Human Frontier
Science Program (S.-T.D.H.), NSC Taiwan ROC (S.-T.D.H.), the
EU FP6 Extend-NMR Grant (18988) (M.V.), the Leverhulme Trust
(M.V.), and the Royal Society (M.V.).

Supporting Information Available: Materials and methods, Tables
S1–S3, and Figures S1–S9. This material is available free of charge
via the Internet at http://pubs.acs.org.

References
(2) Schwarzinger, S.; Kroon, G. J. A.; Foss, T. R.; Chung, J.; Wright, P. E.;
(12) Hsu, S. T. D.; Cabrita, L. D.; Fucini, P.; Dobson, C. M.; Christodoulou, J.
(13) Hsu, S. T. D.; Cabrita, L. D.; Christodoulou, J.; Dobson, C. M. Biomol.
NMR Assign. 2009, 5, 29.
(16) Kumar, D.; Kumar, A.; Mitha, J. R.; Chugii, J.; Sharma, S.; Horsa, R. V.
2008, 130, 16914.
JA904937A 164