J|A|C|S COMMUNICATIONS

Accurate Random Coil Chemical Shifts from an Analysis of Loop Regions in Native States of Proteins

Alfonso De Simone,† Andrea Cavalli,† Shang-Te Danny Hsu,† Wim Vranken,‡ and Michele Vendruscolo*,†

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K., and European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, U.K.

Received June 16, 2009; E-mail: mv245@cam.ac.uk

The definition of a standard set of reference random coil chemical shift values is a key component in many applications of protein NMR spectroscopy.[1−3] 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.[3,4] 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;[5] furthermore, it has also been realized that several amyloidogenic proteins associated with neurodegenerative diseases are natively unfolded.[6]

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 state[1−3] or derived by analysis of protein databases have been proposed.[7,8] 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 known[9] [see the Supporting Information (SI)]. From this database, we extracted for analysis fragments classified by STRIDE[10] as loops (Figure 1a), i.e., not as α-, π-, or 3_{10}-helices, β-sheets, turns or bends; we further selected only flexible loops by including only residues with an RCI index[11] corresponding to an S^2 order parameter 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 δ_{iA}^{RC} of an atom of type i in amino acid of type A as

$$\delta_{iA}^{RC} = \delta_{iA}^{0} + \alpha_i^{-}\delta_{iBA}^{1} + \alpha_i^{+}\delta_{iAC}^{1} \quad (1)$$

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

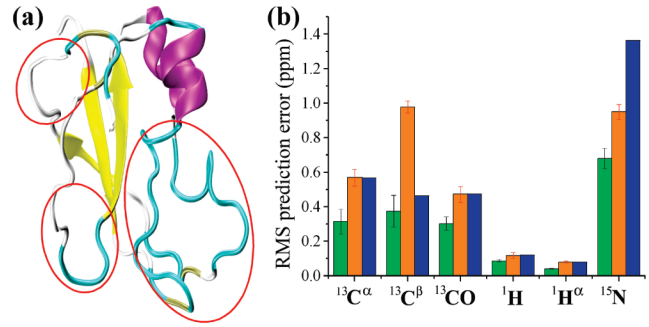

Figure 1. (a) CamCoil random coil chemical shift values are obtained by analyzing the amino acid sequences in the loop regions in a recently compiled database of native structures and corresponding chemical shifts.[9] (b) Average values of the RMS distances (in ppm) for five experimental sets of chemical shifts (a leave-one-out procedure was adopted); green bars refer to the CamCoil values, orange bars to the values of Schwarzinger et al.,[2] and blue bars to the standard deviations of chemical shift values in the database. The five experimental chemical shift data sets are: ddFLN5[12,13] (Figure S3), GED of dynamin in 9.7 M urea,[14] GED of dynamin in 6 M GuHCl,[15] SUMO from *Drosophila melanogaster* in 8 M urea,[16] and *Azotobacter vinelandii* apoflavodoxin in 6 M GuHCl.[17] A web server for the CamCoil method is available at http://www-vendruscolo.ch.cam.ac.uk/camcoil.php.

The weights of these contributions are given by the parameters α_i^{-} and α_i^{+} (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 δ_{iA}^{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.,[2] 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

† University of Cambridge.
‡ European Bioinformatics Institute.

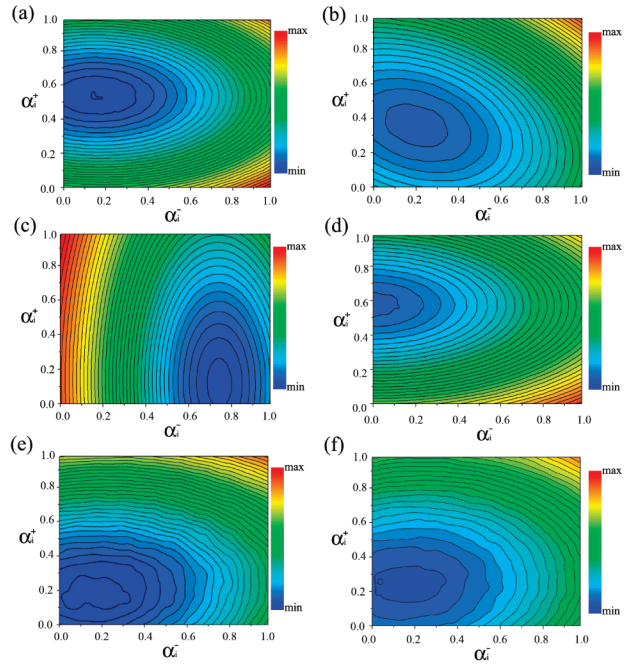

Figure 2. RMS distance surfaces as a function of the parameters α_i^- and α_i^+ in eq 1. These plots were calculated through a global optimization of the five sets of experimental random coil chemical shifts analyzed here (see the Figure S2 caption for more details). Each panel refers to a different atom type: (a) $^{13}C^{\alpha}$; (b) $^{13}C^{\beta}$; (c) ^{13}CO; (d) ^{15}N; (e) ^{1}H; (f)$^{1}H^{\alpha}$.

distance surface projected on the (α_i^-, α_i^+) 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 quintuples 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-neighbor 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 and corresponding experimentally measured chemical shifts, enables a variety of experimental conditions to be averaged out, thus removing biases associated with specific experimental conditions (e.g., the range of pH values at which the structures in the database were determined, which is shown in Figure S7). Moreover, since this approach is based on the analysis of a very large data set, we were able to employ two sets of 400 correction factors in eq 1 and four sets in eq S1, thereby achieving a high accuracy in defining the random coil chemical shift values. The database also enables us to discriminate between oxidized and reduced cysteine residues and between *cis*- and *trans*-proline residues.

The CamCoil method can be also used to obtain pH-specific random coil chemical shift scales. The random coil chemical shift values that we report refer to pH 6.1 (Figure S7). It is possible, however, to define random coil chemical shifts at other pH values by recalibrating the chemical shifts of the side chains of residues D, E, and H in the experimental data sets (Figure S8). This feature is important since the comparison of experimental chemical shifts measured at a given pH with random coil chemical shifts defined at a different pH can generate a significant bias in the interpretation of the results.

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_i^- and α_i^+ 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.[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

(1) Braun, D.; Wider, G.; Wüthrich, K. *J. Am. Chem. Soc.* **1994**, *116*, 8466.
(2) Schwarzinger, S.; Kroon, G. J. A.; Foss, T. R.; Chung, J.; Wright, P. E.; Dyson, H. J. *J. Am. Chem. Soc.* **2001**, *123*, 2970.
(3) Wishart, D. S.; Sykes, B. D.; Richards, F. M. *J. Mol. Biol.* **1991**, *222*, 311.
(4) Spera, S.; Bax, A. *J. Am. Chem. Soc.* **1991**, *113*, 5490.
(5) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197.
(6) Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.* **2006**, *75*, 333.
(7) Wang, L.; Eghbalnia, H. R.; Markley, J. L. *J. Biomol. NMR* **2007**, *39*, 247.
(8) Wang, Y. J.; Jardetzky, O. *J. Am. Chem. Soc.* **2002**, *124*, 14075.
(9) Vranken, W. F.; Rieping, W. *BMC Struct. Biol.* **2009**, 9.
(10) Heinig, M.; Frishman, D. *Nucleic Acids Res.* **2004**, *32*, W500.
(11) Berjanskii, M. V.; Wishart, D. S. *Nucleic Acids Res.* **2007**, *35*, W531.
(12) Hsu, S. T. D.; Cabrita, L. D.; Fucini, P.; Dobson, C. M.; Christodoulou, J. *J. Mol. Biol.* **2009**, *388*, 865.
(13) Hsu, S. T. D.; Cabrita, L. D.; Christodoulou, J.; Dobson, C. M. *Biomol. NMR Assign.* **2009**, *3*, 29.
(14) Chugh, J.; Sharma, S.; Hosur, R. V. *Arch. Biochem. Biophys.* **2009**, *481*, 169.
(15) Chugh, J.; Sharma, S.; Hosur, R. V. *Biochemistry* **2007**, *46*, 11819.
(16) Kumar, D.; Kumar, A.; Misra, J. R.; Chugh, J.; Sharma, S.; Hosur, R. V. *Biomol. NMR Assign.* **2008**, *2*, 13.
(17) Nabuurs, S. M.; Westphal, A. H.; van Mierlo, C. P. M. *J. Am. Chem. Soc.* **2008**, *130*, 16914.
(18) Jha, A. K.; Colubri, A.; Freed, K. F.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099.

JA904937A