The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts

Mark V. Berjanskii[1] and David S. Wishart[1,2,3,]*

[1]Department of Computing Science and [2]Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada T6G 2E8, and [3]National Institute for Nanotechnology (NINT), 11421 Saskatchewan Drive, Edmonton, AB, T6G 2M9

ABSTRACT

Protein motions play important roles in numerous biological processes such as enzyme catalysis, muscle contractions, antigen–antibody interactions, gene regulation and virus assembly. Knowledge of protein flexibility is also important in rational drug design, protein docking and protein engineering. However, the experimental measurement of protein motions is often difficult, requiring sophisticated experiments, complex data analysis and detailed information about the protein's tertiary structure. As a result, there is a considerable interest in developing simpler, more effective ways of quantifying protein flexibility. Recently, we described a method, called the random coil index (RCI), which is able to quantitatively estimate backbone root mean square fluctuations (RMSFs) of structural ensembles and order parameters using only chemical shifts. The RCI method is very fast ($<5\,s$) and exceedingly robust. It also offers an excellent alternative to traditional methods of measuring protein flexibility. We have recently extended the RCI concept and implemented it as a web server. This server allows facile, accurate and fully automated predictions of MD RMSF values, NMR RMSF values and model-free order parameters (S^2) directly from chemical shift assignments. It also performs automatic chemical shift re-referencing to ensure consistency and reproducibility. On average, the correlation between RCI predictions and experimentally obtained motional amplitudes is within the range from 0.77 to 0.82. The server is available at http://wishart.biology.ualberta.ca/rci.

INTRODUCTION

NMR spectroscopy occupies a unique place among experimental methods for investigating protein dynamics. This is because it can provide site-specific information about protein motions over a large range of time scales. Over the past decade, ^{15}N NMR relaxation experiments employing model-free analysis (1,2) have become the *de facto* standard used to characterize protein motions on a picosecond to nanosecond time scale. However, as with any scientific method, this approach has certain limitations (3,4). Perhaps the most obvious difficulty lies in the fact that ^{15}N relaxation measurements are inherently time-consuming and tedious, often requiring many hours of data collection, processing and spectral analysis. A second problem is that relaxation measurements are often seriously compromised by peak overlap, poor signal intensity or peak broadening. A third problem is that the precision and accuracy of relaxation-derived measurements tends to deteriorate rapidly as the frequency of internal nanosecond motions approach that of the protein's overall tumbling rate (5,6). This is because ^{15}N relaxation rates become insensitive to internal fluctuations that are much slower than overall tumbling. Since calculations of relaxation rates are based on measuring peak intensities, their accuracy can be severely compromised by low signal-to-noise ratios. This is especially true when dealing with larger (>150 residues) proteins or proteins undergoing μs–ms conformational exchange. Another complication to the model-free formalism lies in the fact that its proper application often requires information about anisotropy of protein overall diffusion and, as a result, the method cannot be used when the 3D structure is not known or when it is largely disordered. These limitations with traditional relaxation measurements prompted us to develop a new, chemical-shift-based technique to characterize protein mobility from NMR data. Specifically, we wanted to develop an easy-to-use, robust approach that would not be affected by protein tumbling rates, uncertainties in peak intensities or lack of knowledge about the protein's 3D structure. This method is called the random coil index or RCI (7).

As described in our previous publications (7,19), the RCI method exploits the fact that there is a remarkable amount of dynamic information intrinsic to NMR

*To whom correspondence should be addressed. Tel: +1 780 492 0383; Fax: +1 780 492 1071; Email: david.wishart@ualberta.ca

chemical shifts. The connection between chemical shifts, especially random coil chemical shifts, and protein flexibility has been known for quite some time (8–11). Random coil chemical shifts can be defined as shifts that result from a fast exchange among energy-weighted populations of all theoretically possible conformations of an unfolded polypeptide chain (12,13). The difference between an observed chemical shift for a given amino acid in a given protein, and its corresponding random coil value is called the secondary chemical shift. Secondary chemical shifts have been used for many years to qualitatively estimate the level of protein structural disorder (14–18). However, until recently, no quantitative relationships between secondary chemical shifts and protein dynamic parameters had been derived. The RCI is able to combine the chemical shift data from six different nuclei ($^{13}C\alpha$, $^{13}C\beta$, ^{13}CO, ^{15}N, ^{1}HN and $^{1}H\alpha$—or any combinations thereof) into a single parameter that closely correlates with amplitudes of backbone protein motions such as order parameters (S^2) and root mean square fluctuations (RMSFs) of structural ensembles.

Previous descriptions of the RCI method focused on explaining the algorithm, rationalizing its utility and assessing its accuracy. As a result, only a modest effort went into making the RCI-based software user-friendly and flexible. In an attempt to make the RCI approach more accessible to the NMR community and much more robust, we decided to create an RCI web server and to optimize the RCI protocol for NMR assignments with different degrees of completeness and mis-referencing. The RCI server is a unique server, designed to support rapid (5–10 s), residue-specific determination of protein flexibility and protein mobility using only chemical shift assignment data as input. It accepts almost any combination of NMR chemical shift assignments as its input and it outputs the expected values of the RMFs of MD and NMR ensembles as well as model-free order parameters (1,2). To improve the accuracy and reproducibility of the calculations, the RCI server also supports automated chemical shift re-referencing. Assessments of the RCI server performance show an agreement between the values it calculates and experimentally obtained amplitudes of motions range from $R = 0.77-0.82$ depending on the type of experimental method. The RCI web server can be accessed at: http://wishart.biology.ualberta.ca/rci.

SERVER AND PROGRAM DESCRIPTION

The RCI server is composed of two parts, a front-end web-interface (written in Python and HTML) and a back-end consisting of several programs including RCI (7), CSI (11), REFCOR [based on (19)], as well as several parsing and conversion utilities for handling different input files. The CSI program is written in ANSI standard C, while RCI, REFCOR, the input parsing and the conversion utilities are written in Python. The source code for the basic algorithm is available from the authors upon request. The RCI server accepts protein chemical shift assignments in standard BMRB NMR-STAR (20) and SHIFTY (21) formats as input. Users can either upload an

input file into the web server (via a browse button) or paste the data in a standard text box (Figure 1). Users are also offered several options to adjust program operations to suit their specific needs. Additional details about these options and what they can offer are described later.

A flow chart describing the basic RCI server operations is shown in Figure 2. It is also described in more detail here. To begin: (1) Experimental chemical shifts are first uploaded by the user. (2) The input chemical shifts are then re-referenced (if necessary) by REFCOR. (3) The protein sequence is extracted from the NMR assignments and used to (4) determine the appropriate random coil chemical shifts (22) and (5) determine the neighboring residue correction factors for the $i \pm 1$ and $i \pm 2$ residues (23). The correction factors are applied to the random coil chemical shifts to obtain reference chemical shifts. (6) Reference chemical shifts are then subtracted from the corrected experimental chemical shifts to obtain the secondary chemical shifts for the $^{13}C\alpha$, $^{13}C\beta$, ^{13}CO, ^{15}N, ^{1}HN and $^{1}H\alpha$ nuclei. (7) Optionally, gaps in the chemical shift assignments, if any, are filled in by averaging the chemical shifts of neighboring residues. (8) Secondary chemical shifts are smoothed by three-point moving averaging. (9) Secondary chemical shifts of are scaled to account for differences in their resonance frequencies, and, if below a certain 'floor limit' (currently 0.5 p.p.m.), replaced with this floor value. (10) Initial RCI values are calculated using the following expression.

$$\text{RCI} = \left(< A \left|\Delta\delta_{C\alpha}\right| + B \left|\Delta\delta_{co}\right| + C \left|\Delta\delta_{C\beta}\right| + D \left|\Delta\delta_{N}\right| \right.$$
$$\left. + E \left|\Delta\delta_{NH}\right| + F \left|\Delta\delta_{H\alpha}\right| > \right)^{-1} \qquad \textbf{1}$$

where $\left|\Delta\delta_{C\alpha}\right|$, $\left|\Delta\delta_{CO}\right|$, $\left|\Delta\delta_{C\beta}\right|$, $\left|\Delta\delta_{N}\right|$, $\left|\Delta\delta_{NH}\right|$ and $\left|\Delta\delta_{H\alpha}\right|$ are the absolute values of the secondary chemical shifts (in p.p.m.) of Cα, CO, Cβ, N, NH and Hα, respectively. A, B, C, D, E and F are weighting coefficients (Table 1 in the RCI Online Help). Left angle and right angle brackets ($< \cdot >$) indicate that the average is being calculated. (11) End-effect corrections are optionally applied (see later). (12) If the RCI values are above a certain 'ceiling limit' (currently 0.6), they are replaced with this ceiling value. (13) The final RCI values are obtained after a second smoothing by three-point averaging. (14) Finally, in the last step the expected values of model-free order parameters (S^2), RMFs of MD and NMR ensembles are calculated using the following empirical expressions.

$$S^2 = 1 - 0.5 \ln(1 + \text{RCI}^*10.0) \qquad \textbf{2}$$

$$\text{RMSF (MD)} = \text{RCI}^*23.6\text{Å} \qquad \textbf{3}$$

$$\text{RMSF (NMR)} = \text{RCI}^*12.7\text{Å} \qquad \textbf{4}$$

A more detailed description of the aforementioned steps has been published elsewhere (19).

By default, the RCI web server uses random coil reference chemical shifts and neighboring residue correction values originally published by Schwarzinger and co-authors (22,23). This is the only set of random coil values and neighboring residue correction for residues $i \pm 1$ and $i \pm 2$ that were obtained under similar experimental conditions and, therefore, are expected to be quite

Figure 1. Sample screenshots of the RCI web server's input and output pages.
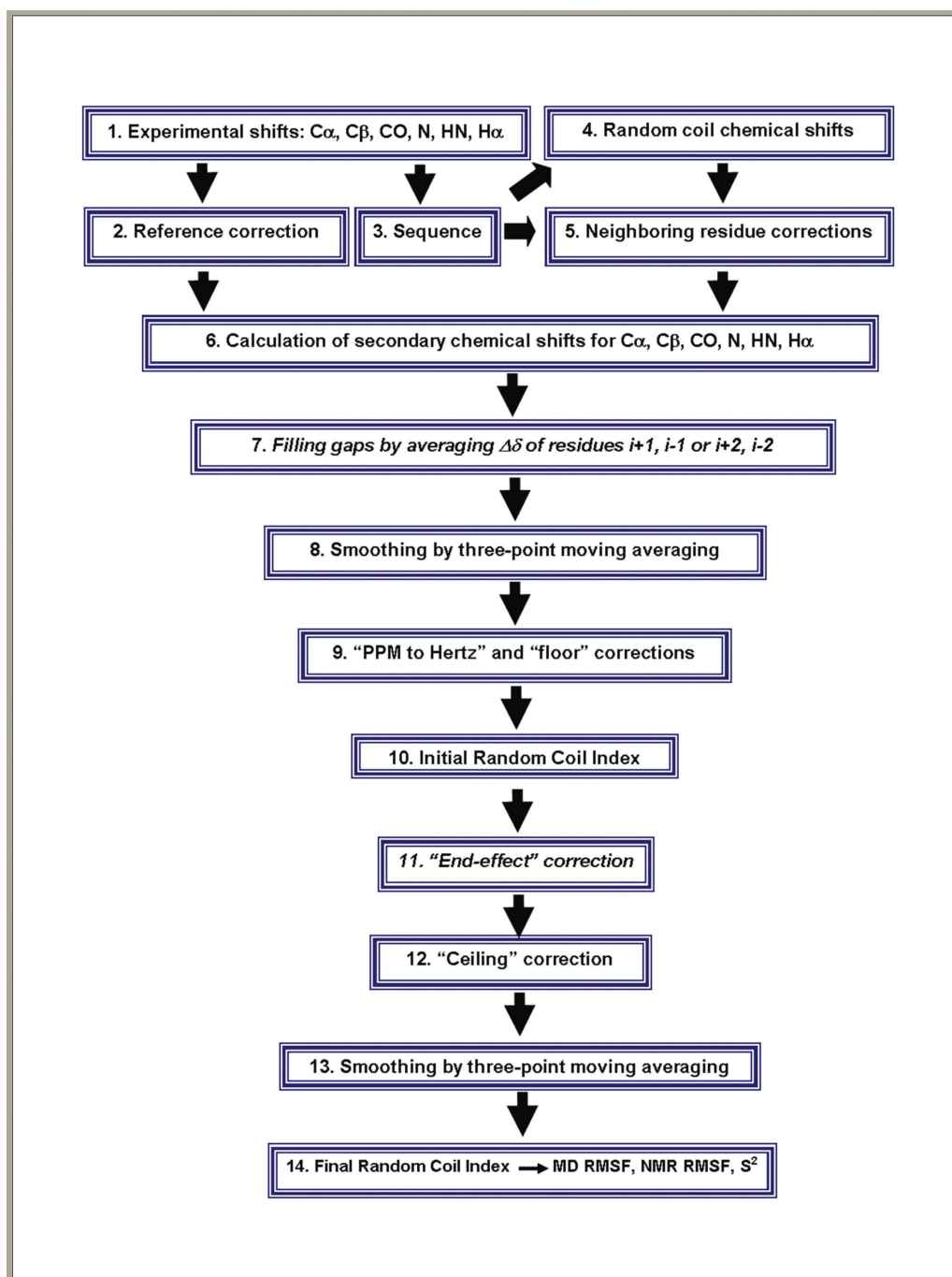
Figure 2. A flow chart describing the RCI protocol.

consistent with each other. However, users can also select different sets of random coil chemical shifts, including those values published by Wang and Jardetzky (24), Wishart *et al.* (25) and Lukin *et al.* (26) using radio-buttons labeled 'Wang', 'Wishart' and 'Lukin', respectively. Users are also offered an option to select $i \pm 1$ neighboring residue correction values published by Wang and Jardetzky (24) instead of the default values. While these options do affect RCI-based flexibility profiles, the differences in mean correlation coefficients for different combinations of random coil values and neighboring

residue corrections (Table 2 of the RCI Online Help) are relatively small and may be considered statistically insignificant.

By default, the RCI web server does not produce predictions for unassigned residues. However, if unassigned residues are located in the middle of a well-defined secondary structure region (e.g. α-helix, β-sheet), it is not unreasonable to expect that their flexibility will be similar to that of neighboring assigned resides. In these cases, the RCI web server allows users to predict the flexibility of unassigned residues based dynamic properties of adjacent

regions. This option is generally not recommended when more than two consecutive residues are unassigned.

The RCI web server also has a somewhat similar option to fill small gaps in the per-residue distributions of secondary chemical shifts. Such gaps happen more often than a lack of assignments for all the six nuclei and should be dealt with separately. The use of incomplete assignments generally reduces the quality of RCI-based flexibility predictions (see Table 1 of the RCI Online Help). By default, the RCI web server fills small gaps in per-residue distributions of secondary chemical shifts by averaging secondary chemical shifts of residues $i+1$ and $i-1$, and, if not available, residues $i+2$ and $i-2$. Users are offered an option to turn this feature off.

RCI values can also be affected by so-called 'end effects'. These are poorly understood phenomena that alter the chemical shifts of the N- and C-termini of proteins in unpredictable ways. The RCI web server offers an option to apply an end-effect correction to the RCI values. This end-effect correction is turned on by default. A detailed protocol describing how the end-effect correction is handled was published elsewhere (27). Users should be aware that the improvement of RCI predictions due to the end-effect correction is not related to the dynamic averaging of chemical shifts. Rather, it originates from the indirect correlation between the severity of the end effects and protein flexibility due to their shared dependence on the proximity to a terminal residue. Therefore, if the correction is applied, the RCI profile at the terminal regions should be interpreted as a commonly observed trend of protein flexibility at termini of proteins. The RCI server offers users the option to disable the end effect correction or to exclude the first and the last three residues from predictions.

A particularly useful and important feature of the RCI web server is its capacity to correct mis-referenced chemical shifts. This automated reference correction is done using a local implementation of the REFCOR program. About 20% of newly deposited assignments in the BMRB database were found mis-referenced by a recent survey (28). Given the relatively high level of mis-referencing in biomolecular NMR and given the fact that mis-referenced shifts can substantially reduce the performance of chemical-shift-based methods such as RCI (7), we believed that implementing this reference correction protocol was absolutely vital to maintaining the server's performance.

The REFCOR reference correction protocol is based on the secondary structure predictions from the chemical shift index (CSI) (29) and was published elsewhere (27). Briefly, REFCOR uses CSI-based predictions of secondary structure along with typical chemical shift values observed in different secondary structures are used to calculate the necessary reference correction for each nucleus. Since CSI predictions are also affected by shift mis-referencing, the protocol is repeated several times using newly generated re-referenced shifts until the CSI predictions become stable. The reference correction option is always turned on by default although it can be switched off, if necessary.

An average RCI run takes between 5 CPU s (without chemical shift reference correction) and ∼20 s (with reference correction). An example of the web server output is shown in Figure 1. As might be expected, the RCI web server displays the name of the input file, the options selected and a plot of the per-residue RCI distribution. Users may download per-residue distributions of the RCI, predicted model-free order parameters, predicted RMSFs of MD and NMR ensembles as both text files and graphical images. These files may also be obtained individually or as a single linked file containing all the results. Re-referenced chemical shifts and CSI-based predictions of secondary structure are also available for download or direct viewing on the website. In addition to its extensive data output, the RCI web server also offers a comprehensive list of help pages to assist users in preparing their input files, in understanding the RCI method and in interpreting the web server output. This information is provided to make the RCI protocol as transparent as possible and to facilitate protocol trouble-shooting if required.

PERFORMANCE AND VALIDATION

The RCI web server was optimized and evaluated on a set of 18 proteins ranging in length from 56 to 283 residues (1585 residues in total), for which complete or nearly complete ^1H, ^{13}C and ^{15}N chemical shifts were known (Table 3 of the RCI Online Help). A subset of 14 proteins was used to generate molecular dynamic (MD) ensembles and optimize weighting coefficients in the RCI expression [Equation (1)]. This was done by using a simple grid search to maximize RCI correlation with RMSF values determined by molecular dynamics. The weighting coefficients were obtained for all 63 possible assignment scenarios (Table 1 of the RCI Online Help). To evaluate the performance of the algorithm for the training set, we used a leave-one-out procedure by removing the query protein from RCI training set prior to running the program. The average correlation coefficients for the full training set and for all leave-one-out runs were identical ($R = 0.82$). To ensure that the RCI algorithm had not been over-trained, four additional proteins, which were not previously included in the training set, were tested. The average correlation coefficient between the RCI determined MD RMSF and the calculated MD RMSF for these proteins was identical ($R = 0.82$) to that of the training set (see Table 3 of the RCI Online Help for information about the tested proteins).

To validate the relationship between RCI predictions and the amplitudes of protein motions, the correlation of RCI with model-free order parameters [observed and predicted from the structure (30)] and per-residue RMSF values of NMR ensembles were calculated. The average correlation coefficients for the eighteen sets of order parameters and sixteen sets of NMR RMSFs were 0.77 and 0.81, respectively. The performance of the conversion expressions shown in Equations (2–4) were assessed by calculating the average error between RCI-predicted motional amplitudes and the corresponding

experimentally and theoretically obtained parameters. The average errors for the order parameters (18 proteins), the NMR RMSFs (16 proteins) and the MD RMSFs (18 proteins) were 0.05, 0.44 Å and 0.50 Å, respectively. Figure 1 in the RCI Online Help shows examples of the good correlations obtained between RCI and other measures of protein motional amplitudes (e.g. NMR RMSF, MD RMSF, S^2) for such pharmaceutically important proteins as interleukin-4 and the HIV-1 Gag protein. Additional information about optimizing and testing the RCI method (i.e. details of MD simulations, PDB IDs and BMRB codes of proteins, etc.) can be found in the original papers on the RCI method (7) as well as on the RCI Online Help pages.

CONCLUSION

In summary, we have described a web server that is capable of rapidly and accurately predicting protein flexibility using only chemical shift assignments as the input. Comparisons suggest that these predictions correlate well with other measures of protein mobility, such as model-free order parameters and root-mean square fluctuations (RMSFs) of NMR and MD ensembles. The approach is generally applicable to proteins of any size for which ^{1}H, ^{13}C and ^{15}N shift assignments are available. The web server is unique in its ability to extract dynamic information from NMR data without the prior knowledge of tertiary structure, without the need for favorable rates of rotational diffusion and without the need for exceptionally good spectral sensitivity. We have already used the RCI approach to monitor a number of interesting dynamic processes in proteins, such as the pH-induced conversion of the prion protein (PrP[c]) into its scrapie form (PrP[sc]) and the dynamic response of picornaviral protease active sites to inhibitor binding (manuscripts in preparation). We believe the RCI method may lead to important changes in the ways backbone protein flexibility is measured and reported by the scientific community.

ACKNOWLEDGEMENTS

Conflict of interest statement. None declared.

REFERENCES

1. Lipari,G. and Szabo,A. (1982) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.*, **104**, 4546–4559.
2. Clore,G.M., Szabo,A., Bax,A., Kay,L.E., Driscoll,P.C. and Gronenborn,A.M. (1990) Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *J. Am. Chem. Soc.*, **112**, 4989–4991.
3. Korzhnev,D.M., Orekhov,V.Y. and Arseniev,A.S. (1997) Model-free approach beyond the borders of its applicability. *J. Magn. Reson.*, **127**, 184–191.
4. Case,D.A. (2002) Molecular dynamics and NMR spin relaxation in proteins. *Acc. Chem. Res.*, **35**, 325–331.
5. Jin,D.Q., Andrec,M., Montelione,G.T. and Levy,R.M. (1998) Propagation of experimental uncertainties using the Lipari-Szabo model-free analysis of protein dynamics. *J. Biomol. NMR*, **12**, 471–492.
6. Chen,J., Brooks,C.L.III and Wright,P.E. (2004) Model-free analysis of protein dynamics: assessment of accuracy and model selection protocols based on molecular dynamics simulation. *J. Biomol. NMR*, **29**, 243–257.
7. Berjanskii,M.V. and Wishart,D.S. (2005) A simple method to predict protein flexibility using secondary chemical shifts. *J. Am. Chem. Soc.*, **127**, 14970–14971.
8. Grathwoh,C and Wuthrich,K. (1974) C-13 Nmr of protected tetrapeptides Tfa-Gly-Gly-L-X-L-Ala-Och3, where X Stands for 20 common amino-acids. *J. Magn. Reson.*, **13**, 217–225.
9. Bundi,A., Grathwohl,C., Hochmann,J., Keller,R.M., Wagner,G. and Wuthrich,K. (1975) Proton Nmr of protected tetrapeptides Tfa-Gly-Gly-L-X-L-Ala-Och3, where X Stands for one of 20 common amino-acids. *J. Magn. Reson.*, **18**, 191–198.
10. Wishart,D.S., Sykes,B.D. and Richards,F.M. (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.*, **222**, 311–333.
11. Wishart,D.S. and Sykes,B.D. (1994) Chemical shifts as a tool for structure determination. *Meth. Enzymol.*, **239**, 363–392.
12. Bundi,A. and Wuthrich,K. (1979) H-1-Nmr parameters of the common amino-acid residues measured in aqueous-solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-Oh. *Biopolymers*, **18**, 285–297.
13. Vila,J.A., Ripoll,D.R., Baldoni,H.A. and Scheraga,H.A. (2002) Unblocked statistical-coil tetrapeptides and pentapeptides in aqueous solution: a theoretical study. *J. Biomol. NMR*, **24**, 245–262.
14. Fiaux,J., Bertelsen,E.B., Horwich,A.L. and Wuthrich,K. (2002) NMR analysis of a 900K GroEL GroES complex. *Nature*, **418**, 207–211.
15. Chou,Y.T., Swain,J.F. and Gierasch,L.M. (2002) Functionally significant mobile regions of Escherichia coli SecA ATPase identified by NMR. *J. Biol. Chem.*, **277**, 50985–50990.
16. Lecroisey,A., Martineau,P., Hofnung,M. and Delepierre,M. (1997) NMR studies on the flexibility of the poliovirus C3 linear epitope inserted into different sites of the maltose-binding protein. *J. Biol. Chem.*, **272**, 362–368.
17. Pufall,M.A., Lee,G.M., Nelson,M.L., Kang,H.S., Velyvis,A., Kay,L.E., McIntosh,L.P. and Graves,B.J. (2005) Variable control of Ets-1 DNA binding by multiple phosphates in an unstructured region. *Science*, **309**, 142–145.
18. Turner,C.F. and Moore,P.B. (2004) The solution structure of ribosomal protein L18 from Bacillus stearothermophilus. *J. Mol. Biol.*, **335**, 679–684.
19. Berjanskii,M. and Wishart,D.S. (2006) NMR: prediction of protein flexibility. *Nature Protoc.*, **1**, 683–688.
20. Seavey,B.R., Farr,E.A., Westler,W.M. and Markley,J.L. (1991) A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, **1**, 217–236.
21. Wishart,D.S., Watson,M.S., Boyko,R.F. and Sykes,B.D. (1997) Automated 1H and 13C chemical shift prediction using the BioMagResBank. *J. Biomol. NMR*, **10**, 329–336.
22. Schwarzinger,S., Kroon,G.J., Foss,T.R., Wright,P.E. and Dyson,H.J. (2000) Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRview. *J. Biomol. NMR*, **18**, 43–48.
23. Schwarzinger,S., Kroon,G.J., Foss,T.R., Chung,J., Wright,P.E. and Dyson,H.J. (2001) Sequence-dependent correction of random coil NMR chemical shifts. *J. Am. Chem. Soc.*, **123**, 2970–2978.
24. Wang,Y. and Jardetzky,O. (2002) Investigation of the neighboring residue effects on protein chemical shifts. *J. Am. Chem. Soc.*, **124**, 14075–14084.

25. Wishart,D.S., Bigam,C.G., Holm,A., Hodges,R.S. and Sykes,B.D. (1995) 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *J. Biomol. NMR*, **5**, 67–81.

26. Lukin,J.A., Gove,A.P., Talukdar,S.N. and Ho,C. (1997) Automated probabilistic method for assigning backbone resonances of (C-13,N-15)-labeled proteins. *J. Biomol. NMR*, **9**, 151–166.

27. Berjanskii,M.V., Neal,S. and Wishart,D.S. (2006) PREDITOR: a web server for predicting protein torsion angle restraints. *Nucleic Acids Res.*, **34**, W63–W69.

28. Zhang,H., Neal,S. and Wishart,D.S. (2003) RefDB: a database of uniformly referenced protein chemical shifts. *J. Biomol. NMR*, **25**, 173–195.

29. Wishart,D.S. and Sykes,B.D. (1994) The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. *J. Biomol. NMR*, **4**, 171–180.

30. Zhang,F. and Bruschweiler,R. (2002) Contact model for the prediction of NMR N-H order parameters in globular proteins. *J. Am. Chem. Soc.*, **124**, 12654–12655.