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

Mark V. Berjanskii¹ and David S. Wishart¹,²,³,*

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

Received February 1, 2007; Revised March 25, 2007; Accepted April 17, 2007

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 (S²) 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, ¹⁵N 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 ¹⁵N 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 ¹⁵N 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...
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$_{a}$, $^{13}$CB, $^{13}$CO, $^{15}$N, $^{1}$HN and $^{1}$Hz—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 RMSFs 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$_{a}$, $^{13}$CB, $^{13}$CO, $^{15}$N, $^{1}$HN and $^{1}$Hz 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.

$$RCI = (A |\Delta \delta_{C_{a}}| + B |\Delta \delta_{CO}| + C |\Delta \delta_{CB}| + D |\Delta \delta_{N}| + E |\Delta \delta_{NH}| + F |\Delta \delta_{Hz}|)^{-1}$$

where $|\Delta \delta_{C_{a}}|$, $|\Delta \delta_{CO}|$, $|\Delta \delta_{CB}|$, $|\Delta \delta_{N}|$, $|\Delta \delta_{NH}|$ and $|\Delta \delta_{Hz}|$ are the absolute values of the secondary chemical shifts (in p.p.m.) of C$_{a}$, CO, CB, N, NH and Hz, respectively. A, B, C, D, E and F are weighting coefficients (Table 1 in the RCI Online Help). Left angle and right angle brackets (<> ) 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 ($S^2$), RMSFs of MD and NMR ensembles are calculated using the following empirical expressions.

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

$$\text{RMSF (MD)} = RCI^{23.6} \AA$$

$$\text{RMSF (NMR)} = RCI^{12.7} \AA$$

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.
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. \( \alpha \)-helix, \( \beta \)-sheet), it is not unreasonable to expect that their flexibility will be similar to that of neighboring assigned residues. 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 assign-
ments generally reduces the quality of RCI-based flex-
ibility 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 correc-
tion 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 perfor-
mance 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 second-
ary 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 \(\sim 20\) 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 distribu-
tions 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 \(^1\)H, \(^13\)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 coeffi-
cients 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 deter-
mined 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 A and 0.50 A, 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 ^1H, ^13C and ^15N 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

Funding for this project was provided by PrioNet Canada, Alberta Prion Research Institute, the Protein Engineering Network of Centres of Excellence (PENCE), the Natural Sciences and Engineering Research Council (NSERC), the National Research Council (NINT) and Genome Alberta (a division of Genome Canada). Funding to pay the Open Access publication charges for this article was provided by Genome Alberta.

Conflict of interest statement. None declared.

REFERENCES


