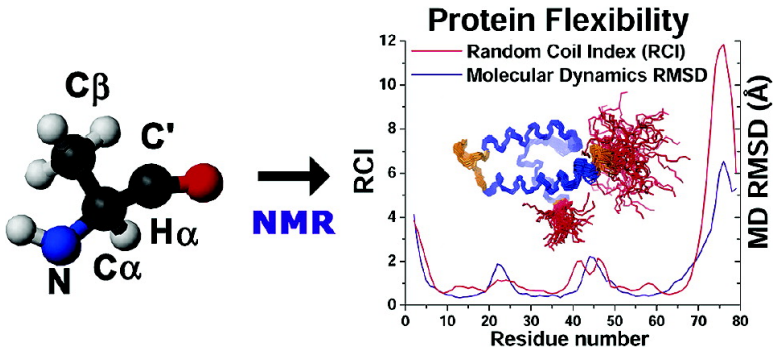
Communication

A Simple Method To Predict Protein Flexibility Using Secondary Chemical Shifts

Mark V. Berjanskii, and David S. Wishart

J. Am. Chem. Soc., **2005**, 127 (43), 14970-14971• DOI: 10.1021/ja054842f • Publication Date (Web): 05 October 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 5 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

J|A|C|S
COMMUNICATIONS

Published on Web 10/05/2005

A Simple Method To Predict Protein Flexibility Using Secondary Chemical Shifts

Mark V. Berjanskii and David S. Wishart*

Departments of Computing Science and Biological Sciences, University of Alberta, Edmonton, AB, Canada T6G 2E8

Received July 19, 2005; E-mail: david.wishart@ualberta.ca

Protein motions play a critical role in many biological processes, such as enzyme catalysis, allosteric regulation, antigen−antibody interactions, and protein−DNA binding. NMR spectroscopy occupies a unique place among methods for investigating protein dynamics due to its ability to provide site-specific information about protein motions over a large range of time scales. However, most NMR methods require a detailed knowledge of the 3D structure and/or the collection of additional experimental data (NOEs, T_1, T_2, etc.) to accurately measure protein dynamics.[1] Here we present a simple method based on chemical shifts that allows accurate, quantitative, site-specific mapping of protein backbone mobility without the need of a 3D structure or the collection and analysis of NMR relaxation experiments. Further, we show that this chemical shift method is able to quantitatively predict per-residue RMSD values (from both MD simulations and NMR structural ensembles) as well as model-free backbone order parameters.[2,3]

Our approach is based on the observation that random coil chemical shifts are often characteristic of highly flexible regions, while nonrandom coil shifts are commonly found in rigid areas of protein structures. Indeed, a qualitative relationship between Hα shifts and X-ray β-factors had been noted as early as 1994.[4] On the basis of this work, we hypothesized that, by including more backbone shifts and by fitting the chemical shift data to lengthy MD simulations, a more quantitative relationship might be revealed. Specifically, we expected that a weighted sum of absolute secondary backbone shifts (i.e., the difference between observed and reference random coil shifts) would be inversely proportional to the calculated amplitudes of backbone motions.

To refine this relationship, 14 well-resolved proteins with complete ^{1}H, ^{13}C, and ^{15}N backbone assignments were selected (Table 1). This training/testing set, consisting of 1585 residues, was chosen to span a range of sizes (56−283 residues) and protein fold classes (all α, all β, mixed α/β) with both ordered and disordered regions. To obtain detailed, residue-specific information on the backbone mobility of these proteins, we calculated a set of 4 ns MD trajectories for each protein using Gromacs 3.2.1.[5] Our MD simulations employed the GROMOS96 43a1 force field,[6] with explicit solvent (SPC water model[7]), PME treatment of electrostatic interactions,[8] Berendsen thermostats for the protein and solvent,[9,10] Berendsen pressure coupling,[11] and a 2 fs time-step. The full set of MD simulations required more than 3700 CPU hours to complete and generated more than 40 Gigabytes of data. Residue-specific backbone RMSD calculations were performed on each protein trajectory as a proxy measure of that protein's backbone mobility. The inverse weighted sum of the observed Cα, CO, Cβ, N, and Hα secondary chemical shifts, henceforth called the Random Coil Index (RCI), was then fit to the residue-specific RMSD of the backbone amide nitrogens as determined from our MD simulations (MD RMSD). The weighting coefficients were optimized via a simple grid search. This grid search resulted in the following expression for RCI.

Table 1. Correlation of RCI with Model-Free Order Parameter S^2, Per-Residue RMSD of MD, and NMR Ensembles

protein	BMRB[c]	PDB	MD[rmsd]	$1-S^2$	NMR[rmsd]
cadherin − N[a]	4380	1SUH	0.86	0.60[d]	0.77
disulfide isomerase[a]	4156	2BJX	0.94	0.83[d]	0.97
Ets-1, 1-110[a]	4205	1BQV	0.74	0.83[d]	0.74
FimC[a]	4070	1BF8	0.78	0.70[d]	0.84
GEF of hEF-1β	4117	1B64	0.8	0.78	0.72
Hngal[a]	4267	1DFV	0.87	0.73[12]	n/a[e]
HSF[a]	4046	1HKT	0.74	0.76[d]	0.85
Kh of Hnrnp K	4405	1KHM	0.79	0.83[13]	0.89
OMTKY3[a]	5473	1OMT	0.77	0.71[14]	0.73
PyJ[a]	4403	1FAF	0.93	0.87[15]	0.83
S4 Δ41[a]	4577	1C06	0.87	0.77[d]	0.89
syntaxin 1A[a]	4198	1BR0	0.75	0.74[d]	0.71
SV40 T DBD[a]	4127	2TBD	0.82	0.81[d]	0.90
ubiquitin[a]	5387	1D3Z 1XQQ[f]	0.81	0.80[16]	0.81[f]
Eh1 of Eps15[b]	*4140*	*1QJT*	*0.74*	*0.71*[d]	*0.62*
Foxo4[b]	*4675*	*1E17$_W$*	*0.81*	*0.74*[d]	*0.77*
HIV-1 Gag[b]	*5316*	*1L6N*	*0.77*	*0.81*[d]	*n/a*[g]
interleukin-4[b]	*4094*	*1BCN*	*0.96*	*0.77*[17]	*0.88*

[a] Proteins used in the grid search. [b] Proteins not included in the grid search. [c] BMRB accession number. [d] S^2 was predicted in our lab using the contact model method.[18] [e] NMR ensemble of Hngal is not available. [f] The ubiquitin model with PDB ID 1XQQ was used to determine the RMSD of the ubiquitin NMR ensemble. [g] Reliable determination of RMSD was impossible due to domain reorientation in NMR ensemble (1L6N). Subscript W = water-refined model from DRESS database.[19]

$$\mathrm{RCI} = (4.80|\Delta\delta_{C\alpha}| + 4.80\,|\Delta\delta_{CO}| + |\Delta\delta_{C\beta}| + 3.93\,|\Delta\delta_{N}| + 5.69\,|\Delta\delta_{H\alpha}|)^{-1} \quad (1)$$

where $|\Delta\delta_{C\alpha}|$, $|\Delta\delta_{CO}|$, $|\Delta\delta_{C\beta}|$, $|\Delta\delta_{N}|$, and $|\Delta\delta_{H\alpha}|$ are the absolute values of the secondary chemical shifts (in ppm) of Cα, CO, Cβ, N, and Hα, respectively. The RCI, itself, is a unitless index.

The actual calculation of RCI involves several steps. First, neighboring residue corrections[20] for residues $i \pm 1$ and $i \pm 2$ are applied to the reference random coil values[21] of all chemical shifts. Second, corrected random coil values are subtracted from re-referenced[22] experimental chemical shifts to obtain secondary chemical shifts ($\Delta\delta$). Third, small gaps in per-residue distribution of $\Delta\delta$ due to missing assignments are filled in by averaging $\Delta\delta$ of $i \pm 1$ residues or, if not available, $i \pm 2$ residues. Next, ^{13}C, ^{15}N, and ^{1}H secondary chemical shifts are scaled by 2.5, 1, and 10, respectively, to account for the characteristic resonance frequencies of these nuclei. Fifth, the floor value for the scaled $\Delta\delta$ is set to 0.8 to avoid infinitely large RCI values when secondary chemical shifts approach zero. Once these corrections are made, the RCI is

10.1021/ja054842f CCC: $30.25

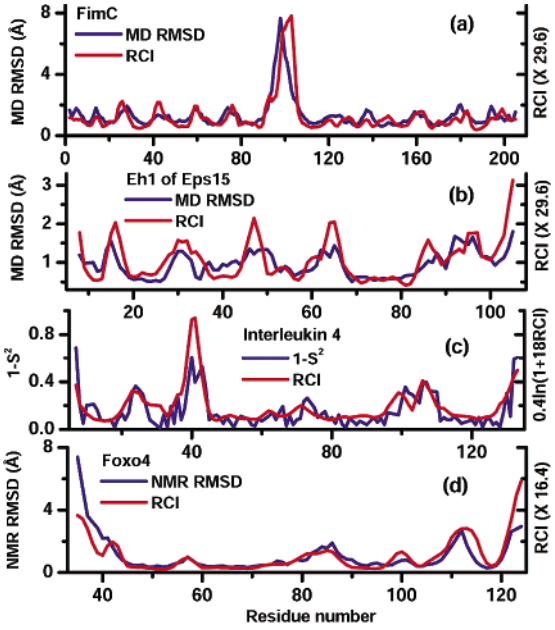


Figure 1. Correlation of RCI with MD RMSD (a and b), experimental S^2 (c), and NMR RMSD (d).

calculated using eq 1. Smoothing by three-point averaging is applied to all shifts before the calculation of RCI as well as to the RCI values themselves. Double smoothing appears to improve the correlation (~4%) by mitigating the contributions of missing assignments, imperfect random coil shifts, ring currents, end-effects, and local environmental processes.

A leave-one-out strategy was employed to test the performance of this method in predicting protein flexibility. Specifically, weighting coefficients for each protein were optimized without including its own chemical shift data in the grid search. The average coefficient of correlation between the RCI and MD RMSD was 0.82 (identical to that obtained using the whole data set). To ensure that the good correlation was not a result of over-fitting, another four proteins, not included in the grid search, were analyzed (Table 1, italics). The average correlation between RCI and the MD RMSD values of these proteins was also 0.82.

We also investigated the generality of the relationship between our RCI values and the amplitudes of protein motions by determining the correlation of RCI with model-free[2,3] order parameters (observed and calculated[18]) as well as with the per-residue (amide nitrogen) RMSD values of NMR ensembles. As seen in Table 1, the RCI values correlate well with these parameters. The average correlation coefficients between RCI versus S^2 and RCI versus NMR RMSD are 0.77 and 0.81, respectively. Figure 1 demonstrates the correlation of RCI with MD RMSD, model-free S^2, and per-residue RMSD from NMR ensembles.

It was also found that RCI values could be used to obtain quantitative estimates of order parameters and RMSD values for all residues excluding the first three N-terminal and the last three C-terminal residues. The scaling relationships are as follows:

$$S^2 = 1 - 0.4 \ln(1 + 17.7\ \mathrm{RCI}) \quad (2)$$

$$\mathrm{RMSD\ (MD)} = \mathrm{RCI} \times 29.6\ \text{Å} \quad (3)$$

$$\mathrm{RMSD\ (NMR)} = \mathrm{RCI} \times 16.4\ \text{Å} \quad (4)$$

Average absolute errors for predicted S^2, MD RMSD, and NMR RMSD from RCI values are 0.05, 0.50, and 0.44 Å, respectively.

Since both model-free order parameters and MD RMSD characterize protein dynamics on a picosecond−nanosecond time scale, it is tempting to speculate that motions identified by secondary chemical shifts also occur on the same time scale.

In summary, by carefully fitting backbone chemical shift data to an extensive set of protein MD simulations, we have developed a simple, chemical shift-based method for detecting picosecond−nanosecond motions. Furthermore, we have also shown that this new method allows quantitative determination of model-free order parameters as well as RMSD (MD and NMR) values. The RCI approach has certain advantages over the commonly used model-free analysis[2,3] of ^{15}N NMR relaxation data in that the method does not rely on a model of overall rotation, it does not need prior knowledge of the protein's tertiary structure, nor does it require additional NMR measurements beyond standard experiments for backbone assignments. The good correlation of the RCI values with S^2 and MD RMSD values for the 32 kDa HIV-1 Gag protein (Table 1) indicates that this method should be especially beneficial for large proteins, for which the collection of relaxation data can be difficult due to spectral overlap and low signal intensity. A Python program that performs all of the RCI calculations described here is freely available at http://wishart.biology.ualberta.ca/download/rci.

Acknowledgment. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) and by the Protein Engineering Network of Centres of Excellence (PENCE).

Supporting Information Available: Details of MD simulations and figures that show correlation of RCI with S^2, MD RMSD, and NMR RMSD for five other proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Palmer, A. G. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 129−155.
(2) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546−4559.
(3) Clore, G. M.; Szabo, A.; Bax, A.; Kay, L. E.; Driscoll, P. C.; Gronenborn, A. M. *J. Am. Chem. Soc.* **1990**, *112*, 4989−4991.
(4) Wishart, D. S.; Sykes, B. D. *Methods Enzymol.* **1994**, *239*, 363−392.
(5) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306−317.
(6) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596−3607.
(7) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel Publishing Co: Dordrecht, The Netherlands, 1981; pp 331−342.
(8) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(9) Nose, S. *Mol. Phys.* **1984**, *52*, 255−268.
(10) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695−1697.
(11) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(12) Coles, M.; Diercks, T.; Muehlenweg, B.; Bartsch, S.; Zolzer, V.; Tschesche, H.; Kessler, H. *J. Mol. Biol.* **1999**, *289*, 139−157.
(13) Baber, J. L.; Levens, D.; Libutti, D.; Tjandra, N. *Biochemistry* **2000**, *39*, 6022−6032.
(14) Song, J.; Markley, J. L. *Biochemistry* **2003**, *42*, 5186−5194.
(15) Berjanskii, M.; Riley, M.; Van Doren, S. R. *J. Mol. Biol.* **2002**, *321*, 503−516.
(16) Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A. *J. Am. Chem. Soc.* **1995**, *117*, 12562−12566.
(17) Redfield, C.; Boyd, J.; Smith, L. J.; Smith, R. A.; Dobson, C. M. *Biochemistry* **1992**, *31*, 10431−10437.
(18) Zhang, F.; Bruschweiler, R. *J. Am. Chem. Soc.* **2002**, *124*, 12654−12655.
(19) Nabuurs, S. B.; Nederveen, A. J.; Vranken, W.; Doreleijers, J. F.; Bonvin, A. M.; Vuister, G. W.; Vriend, G.; Spronk, C. A. *Proteins* **2004**, *55*, 483−486.
(20) Schwarzinger, S.; Kroon, G. J.; Foss, T. R.; Chung, J.; Wright, P. E.; Dyson, H. J. *J. Am. Chem. Soc.* **2001**, *123*, 2970−2978.
(21) Schwarzinger, S.; Kroon, G. J.; Foss, T. R.; Wright, P. E.; Dyson, H. J. *J. Biomol. NMR* **2000**, *18*, 43−48.
(22) Zhang, H.; Neal, S.; Wishart, D. S. *J. Biomol. NMR* **2003**, *25*, 173−195.

JA054842F