

# NMR data collection and analysis protocol for high-throughput protein structure determination

Gaohua Liu<sup>\*†‡</sup>, Yang Shen<sup>\*†‡</sup>, Hanudatta S. Atreya<sup>\*†‡</sup>, David Parish<sup>\*\*</sup>, Ying Shao<sup>\*\*</sup>, Dinesh K. Sukumaran<sup>\*</sup>, Rong Xiao<sup>§</sup>, Adelinda Yee<sup>¶</sup>, Alexander Lemak<sup>¶</sup>, Aneerban Bhattacharya<sup>‡§</sup>, Thomas A. Acton<sup>§</sup>, Cheryl H. Arrowsmith<sup>¶</sup>, Gaetano T. Montelione<sup>§</sup>, and Thomas Szyperski<sup>\*¶</sup>

<sup>\*</sup>Departments of Chemistry and Structural Biology, University at Buffalo, State University of New York, Buffalo, NY 14260; <sup>§</sup>Center for Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, and Robert Wood Johnson Medical School, Piscataway, NJ 08854; and <sup>¶</sup>Department of Medical Biophysics and Banting and Best Department of Medical Research, University of Toronto, Toronto, ON, Canada M5G 1L5

Communicated by Wayne A. Hendrickson, Columbia University, New York, NY, May 25, 2005 (received for review November 6, 2004)

**A standardized protocol enabling rapid NMR data collection for high-quality protein structure determination is presented that allows one to capitalize on high spectrometer sensitivity: a set of five G-matrix Fourier transform NMR experiments for resonance assignment based on highly resolved 4D and 5D spectral information is acquired in conjunction with a single simultaneous 3D <sup>15</sup>N, <sup>13</sup>C<sub>aliphatic</sub>, <sup>13</sup>C<sub>aromatic</sub>-resolved [<sup>1</sup>H, <sup>1</sup>H]-NOESY spectrum providing <sup>1</sup>H-<sup>1</sup>H upper distance limit constraints. The protocol was integrated with methodology for semiautomated data analysis and used to solve eight NMR protein structures of the Northeast Structural Genomics Consortium pipeline. The molecular masses of the hypothetical target proteins ranged from 9 to 20 kDa with an average of ≈14 kDa. Between 1 and 9 days of instrument time were invested per structure, which is less than ≈10–25% of the measurement time routinely required to date with conventional approaches. The protocol presented here effectively removes data collection as a bottleneck for high-throughput solution structure determination of proteins up to at least ≈20 kDa, while concurrently providing spectra that are highly amenable to fast and robust analysis.**

G-matrix Fourier transform projection NMR | NMR structure determination | structural genomics

**M**ultidimensional NMR spectroscopy is an indispensable tool to determine atomic resolution structures of biological macromolecules in solution (1). Hence, NMR plays an important role for structural genomics (2–4), which aims at making 3D structural information available for each protein domain family in nature. However, typical NMR measurement times on the order of ≈2–6 weeks per structure (e.g., ref. 3) have so far limited throughput. Structure determination nowadays can be accelerated by using highly sensitive spectrometers equipped with cryogenic probes (5). These probes allow reducing measurement times by approximately an order of magnitude, indicating that data collection for structure determination could be accomplished within a few days (e.g., ref. 6).

When using conventional multidimensional NMR, however, fast data collection for structure determination is impeded by the need to record several spectra, each of which requires sampling of two or more indirect dimensions (7). With highly sensitive instrumentation, this protocol can lead to data acquisition in the “sampling limited” regime (4), in which a large fraction (or even most) of the spectrometer time is invested to sample indirect dimensions and not for achieving sufficient signal-to-noise ratios. G-matrix Fourier transform (GFT) NMR spectroscopy (8–10) offers a solution to this “NMR sampling problem” (11) by joint sampling of several indirect dimensions. This approach leads to detection of “chemical shift multiplets” in which each component encodes a defined linear combination of jointly sampled shifts. To avoid spectral crowding, G-matrix transformation enables one to edit

the multiplets; that is, each type of linear combination of shifts is registered in a separate subspectrum.

Here, we present a protocol for rapid NMR data collection based on GFT NMR and simultaneous 3D <sup>15</sup>N, <sup>13</sup>C<sub>aliphatic</sub>, <sup>13</sup>C<sub>aromatic</sub>-resolved [<sup>1</sup>H, <sup>1</sup>H]-NOESY (3D NOESY) (12, 13) for high-quality NMR structure determination. The protocol was used for eight targets of the Northeast Structural Genomics (NESG) consortium (www.nesg.org). Molecular masses of uniformly <sup>13</sup>C, <sup>15</sup>N-double-labeled polypeptides expressed with tags for structural studies ranged from 10 to 22 kDa (average: 16.2 kDa), and NMR experiments were recorded with ≈1 mM protein solutions at ambient temperature. The study demonstrates feasibility and robustness of high-throughput solution NMR structure determination of domain-sized proteins.

## Materials and Methods

**NMR Sample Preparation.** Seven uniformly (*U*) <sup>13</sup>C, <sup>15</sup>N-labeled samples were produced at the NESG production site at Rutgers University as described in ref. 14 for targets encoded by genes *Pyrococcus furiosus* PF0470 (SwissProt accession no. Q8U3J6; NESG ID PFR14), *Bacillus cereus* BC4709 (Q816V6; BcR68), *Bacillus subtilis* yqbG (P45923; SR215), *Escherichia coli* yhgG (P64639; ET95), *Methanosarcina mazei* rps24e (Q8PZ95; MaR11), *Bacillus halodurans* BH1534 (Q9KCN5; BhR29), and *Homo sapiens* UFC1 (Q9Y3C8; HR41). The expressed proteins contained a C-terminal tag with sequence LEH<sub>6</sub> to facilitate purification, and ≈1 mM solutions were prepared (Table 1) in 95% H<sub>2</sub>O/5% <sup>2</sup>H<sub>2</sub>O (20 mM Mes, pH 6.5/100 mM NaCl/10 mM DTT/5 mM CaCl<sub>2</sub>/0.02% NaN<sub>3</sub>). The eighth *U*-<sup>13</sup>C, <sup>15</sup>N-labeled sample was produced for a target encoded by *E. coli* gene yqfB (P67603; ET99). The sample was produced at the Toronto site as described in ref. 3, contained a 22-residue N-terminal tag with sequence MGTSH<sub>6</sub>SSGRENLYFQGH, and was concentrated to ≈1 mM in 90% H<sub>2</sub>O/10% <sup>2</sup>H<sub>2</sub>O (25 mM Na phosphate, pH 6.5/400 mM NaCl/1 mM DTT/20 mM ZnCl<sub>2</sub>/0.01% NaN<sub>3</sub>). The predicted *in vivo* molecular masses of the target proteins range from 9 to 20 kDa (average: 14.0 kDa). However, when

Abbreviations: GFT, G-matrix Fourier transform; 3D NOESY, 3D <sup>15</sup>N, <sup>13</sup>C<sub>aliphatic</sub>, <sup>13</sup>C<sub>aromatic</sub>-resolved [<sup>1</sup>H, <sup>1</sup>H]-NOESY; NESG, Northeast Structural Genomics; NOE, nuclear Overhauser effect; PDB, Protein Data Bank; rmsd, rms deviation.

Data deposition: Chemical shift data were deposited in the BioMagResBank, www.bmrb.wisc.edu, and the Protein Data Bank, www.pdb.org (accession no. and PDB ID code (gene name): 6207 and 1te7 (yqfB); 6364 and 1xn6 (PF0470); 6365 and 1xn6 (BC4709); 6366 and 1xn8 (yqbG); 6367 and 1xn7 (yhG); 6368 and 1xn9 (rps24e); 6369 and 1xn5 (BH1534); 6546 and 1yww (UFC1); and 6363 and 1xpv (XCC2852)).

<sup>†</sup>G.L., Y. Shen, and H.S.A. contributed equally to this work.

<sup>‡</sup>G.L., Y. Shen, H.S.A., D.P., Y. Shao, R.X., A.Y., A.L., A.B., T.A.A., C.H.A., G.T.M., and T.S. are members of the Northeast Structural Genomics Consortium.

<sup>¶</sup>To whom correspondence should be addressed at: Department of Chemistry, University at Buffalo, State University of New York, Buffalo, NY 14260. E-mail: szyperski@chem.buffalo.edu.

© 2005 by The National Academy of Sciences of the USA