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

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

[*]Departments of Chemistry and Structural Biology, University at Buffalo, State University of New York, Buffalo, NY 14260; [§]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 [¶]Department of Medical Biophysics and Banting and Best Department of Medical Research, University of Toronto, Toronto, ON, Canada M5G IL5

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 ^{15}N,^{13}C$^{\text{aliphatic}}$,^{13}C$^{\text{aromatic}}$-resolved [^{1}H,^{1}H]-NOESY spectrum providing ^{1}H-^{1}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

Multidimensional 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 ^{15}N,^{13}C$^{\text{aliphatic}}$,^{13}C$^{\text{aromatic}}$-resolved [^{1}H,^{1}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 ^{13}C,^{15}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) ^{13}C,^{15}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$_6$ to facilitate purification, and ≈1 mM solutions were prepared (Table 1) in 95% H_2O/5% $^{2}H_2O$ (20 mM Mes, pH 6.5/100 mM NaCl/10 mM DTT/5 mM $CaCl_2$/0.02% NaN_3). The eighth U-^{13}C,^{15}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$_6$SSGRENLYFQGH, and was concentrated to ≈1 mM in 90% H_2O/10% $^{2}H_2O$ (25 mM Na phosphate, pH 6.5/400 mM NaCl/1 mM DTT/20 mM $ZnCl_2$/0.01% NaN_3). 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 ^{15}N,^{13}C$^{\text{aliphatic}}$,^{13}C$^{\text{aromatic}}$-resolved [^{1}H,^{1}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 1xne (PF0470); 6365 and 1xn6 (BC4709); 6366 and 1xn8 (yqbG); 6367 and 1xn7 (yhgG); 6368 and 1xn9 (rps24e); 6369 and 1xn5 (BH1534); 6546 and 1ywz (UFC1); and 6363 and 1xpv (XCC2852)].

[†]G.L., Y. Shen, and H.S.A. contributed equally to this work.

[‡]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.

[‖]To whom correspondence should be addressed at: Department of Chemistry, University at Buffalo, State University of New York, Buffalo, NY 14260. E-mail: szypersk@chem.buffalo.edu.

Table 1. Survey of NMR structure determinations

Parameters	yqfB[a] (ET99)	PF0470 (PfR14)	BC4709 (BcR68)	yqbG (SR215)	yhgG (ET95)	rps24e (MaR11)	BH1534[b] (BhR29)	UFC1 (HR41)
Molecular mass,[c] kDa	15.3/11.9	15.7/13.8	18.1/16.1	16.7/14.7	10.3/8.7	13.5/11.7	18.0/15.9	21.7/19.5
Correlation time τ_r at 25°C, ns	≈7.7	≈8.1	≈10	≈8.5	≈5.1	≈6.5	≈8.7	≈11
Protein concentration, mM	≈1.0	≈1.0	≈1.5	≈0.9	≈1.1	≈1.0	≈0.8	≈1.0
BMRB accession no./PDB ID	6207/1te7	6364/1xne	6365/1xn6	6366/1xn8	6367/1xn7	6368/1xn9	6369/1xn5	6546/1ywz
NMR Measurement time								
HNNC$^{\alpha\beta}$C$^{\alpha}$ and C$^{\alpha\beta}$C$^{\alpha}$(CO)NHN, hr.	10.2	44	49	39	13	34	39	67
HACACONHN/H$^{\alpha\beta}$C$^{\alpha\beta}$(CO)NHN, hr.	1/–	–/26	–/26	–/26	2/–	–/17	–/18	–/28
HCCH aliphatic/aromatic, hr.	4.0/1.4	21.5/6.5	26/13	26/13	9/–	15/6.5	22/13.5	29/16
NOESY (750 MHz), hr. [mixing time, ms]	9.1[d] [70]	103 [70]	51 [60]	23 [60]	24 [60]	46 [60]	46 [60]	73 [60]
Total measurement time, days	1.1	8.5	6.9	5.3	2.0	5.0	5.7	8.9
Expert time, days								
Assignment bb/sc [total]	–	2/3 [5]	1/3 [4]	1/2 [3]	2/2 [4]	0.5/1 [1.5]	1/2 [3]	3/5 [8]
Structure refinement	–	10	11	6	5	5.5	6	15
Total expert time	–	15	15	9	9	7	9	23
Structure statistics								
Completeness bb/sc assign.,[e] %	98/95	84/89	99/99	100/99	98/99	100/99	99/99	97/97
Consensus NOE assign.,[f] %	–	56	67	33	35	53	81	57
Total NOE peaks assigned, (N/C^{aliphatic}/C^{aromatic})	1393/3178/241	1561/4169/228	2488/6039/454	2161/5923/244	1156/3123/109	1797/4825/127	2303/5703/423	2572/6381/431
NOE constraints: intraresidue/ sequential/medium-range/ long-range[g]	454/511/ 208/280	505/622/ 418/683	561/861/ 625/1326	583/923/ 962/685	466/403/ 221/251	567/736/ 462/866	666/787/ 590/1010	667/955/ 838/916
No. NOE constraints	1453	2228	3373	3153	1341	2631	3084	3376
No. dihedral angle constraints, ϕ/ψ	68/68	51/51	68/68	53/53	40/40	54/54	79/79	80/80
Total no./long-range NOE constraints per residue	15.4/2.7	20.6/6.0	24.5/9.3	24.9/5.2	18.5/3.3	27.1/8.6	23.2/7.3	21.3/5.6
Completeness of SA β/isopropyl,[h] %	34/58	64/61	68/70	55/80	45/67	67/59	57/90	58/73
DYANA target function, Å^2	1.89 ± 0.16	0.12 ± 0.02	0.18 ± 0.03	0.30 ± 0.08	0.24 ± 0.02	0.21 ± 0.04	0.08 ± 0.02	1.71 ± 0.05[i]
rmsd[j] regular secondary,[k] Å	0.43 ± 0.11	0.38 ± 0.06	0.34 ± 0.08	0.34 ± 0.06	0.27 ± 0.06	0.22 ± 0.05	0.28 ± 0.05	0.61 ± 0.15
rmsd heavy atoms best defined,[l] Å	0.42 ± 0.07	0.38 ± 0.06	0.27 ± 0.05	0.43 ± 0.08	0.30 ± 0.07	0.19 ± 0.04	0.24 ± 0.04	0.58 ± 0.12
rmsd all heavy atoms,[m] Å	1.19 ± 0.25	0.97 ± 0.07	0.76 ± 0.09	0.91 ± 0.13	0.66 ± 0.07	0.80 ± 0.05	0.83 ± 0.07	1.06 ± 0.12
Structure quality validation								
R/P/DP scores, %[n]	94/89/67[a,d]	96/93/77	95/93/81	96/96/82	95/91/74	96/96/83	96/95/79	95/98/78
Ramachandran plot,[m,o] %	72/24/4/0	79/18/2/1	73/22/4/1	81/18/1/0	91/7/2/0	85/14/1/0	71/26/2/1	73/23/4/0
G-factors,[m] ϕ & ψ/all	−0.99/−1.26	−0.68/−0.92	−0.86/−0.93	−0.36/−0.68	−0.23/−0.65	−0.50/−0.86	−0.95/−1.02	−0.82/−0.92
MOLPROBITY clash score[p]	74.6 ± 5.9	24.6 ± 4.3	38.8 ± 3.1	36.8 ± 5.3	28.3 ± 4.8	31.0 ± 4.4	39.3 ± 2.7	10.9 ± 3.9

Gene names are given with NESG ID codes in parentheses.

[a] Measurement time minimized; testing of resonance assignment protocol (see text).

[b] Protein precipitated during data collection with a rate of 6% per day.

[c] Molecular masses are listed for *U*-[^{15}N, ^{13}C] labeled protein with His-tag [since tags affect τ_r (38)]/for expected *in vivo* expressed protein.

[d] Recorded with cryogenic probe at 600 MHz. Measurement time corresponds to ≈24 hours with conventional probe at 750 MHz. Minimization of measurement time is reflected in somewhat decreased scores.

[e] For backbone (bb); the assignment yields was calculated by excluding N-terminal NH_3^+, Pro ^{15}N, and ^{13}C′ shifts of residues preceding Pro residues. For side-chains (sc); excluding side-chain OH, ^{13}C′ and aromatic quaternary ^{13}C shifts, and Lys NH_3^+, Arg NH_2.

[f] Only conformationally restricting NOEs. Intraresidue [$i = j$], sequential [$|i - j| = 1$], medium-range [$1 < |i - j| \leq 4$], long-range [$|i - j| > 4$] with NOE connecting residues *i* and *j*.

[g] Obtained from parallel run using AUTOSTRUCTURE and CYANA (see text).

[h] Stereospecific assignment (SA) of diastereotopic moieties with non-degenerate shifts. β: β-CH_2; isopropyl: Val and Leu methyl groups.

[i] Structure calculation was performed with CYANA 2.0.

[j] Average rmsd values relative to the mean DYANA coordinates.

[k] For N, C$^{\alpha}$, and C′ atoms of regular secondary structure elements; yqfB: residues 10–17, 69–72, 78–88 (α-helices), and 6–7, 22–26, 38–42, 51–61, 95–101 (β-stands); PF0470: 10–18, 61–66, 79–85 (α-helices), and 2–8, 23–25, 40–45, 47–57, 101–107 (β-strands); BC4709: 20–26, 31–35, 118–143 (α-helices), and 10–16, 49–54, 59–67, 71–76, 81–90, 94–102 (β-strands); yqbG: 6–12, 16–20, 23–41, 55–72, 105–111, 125–129 (α-helices), residues 125–129 are flexible and excluded; yhgG: 4–14, 19–25, 30–43 (α-helices), and 16–18, 46–50, 73–77 (β-strands); rps24e: 32–45, 74–84 (α-helices), and 2–11, 16–24, 50–58, 63–70 (β-strands); BH1534: 17–23, 27–32, 114–128, 130–138 (α-helices), and 7–14, 46–51, 54–64, 68–73, 78–87, 90–98 (β-strands); UFC1: 4–12, 25–49, 121–127, 134–149, 142–148, 150–155 (α-helices), and 54–59, 64–73, 76–85 (β-strands).

[l] Residues with best-defined side chains. yqfB: 7, 16, 22–25, 35, 38–39, 52, 55–67, 71, 77, 79, 81, 84–86, 93, 97, 98; PF0470: 3, 6, 8, 13, 15– 18, 23, 28, 35, 40–42, 46–49, 51, 52, 63, 68, 69, 72, 77, 78, 82, 102, 103, 107, 108; BC4709: 10, 13–15, 17–20, 23–29, 33–34, 43–44, 46, 63–64, 66, 69, 71– 73, 75, 82–83, 87, 90, 95, 98, 100, 126, 129–134, 140–141; yqbG: 4, 6, 9, 11, 13, 18, 22, 25–26, 28, 30–31, 33–35, 37–39, 41, 53–54, 56–57, 59–60, 62, 64, 69– 71, 102, 104, 106–108, 112; yhgG: 3, 11–12, 19–20, 22–23, 25–26, 28–31, 33–35, 37, 40, 42, 46, 47, 74–77; rps24e: 3, 5–6, 19, 22–24, 30, 35, 37, 39–41, 43, 45– 47, 49–52, 55, 57, 64, 68, 70, 74, 80; BH1534: 7, 11, 14–17, 20–21, 23–27, 30–31, 33, 37, 39, 41, 58, 60–61, 63, 66, 67, 69–70, 72, 79–81, 84, 91–92, 94– 96, 125, 128–129, 133, 136–137; UFC1: 9–11, 13–16, 18, 20, 22, 27–29, 32, 37–40, 43, 48, 53, 58, 63, 65, 82, 86–89, 91–95, 97, 98, 102, 115, 124, 125, 129, 130, 135–137, 139–140, 142, 144–148, 150, 151, 153, 154, 159.

[m] Ordered residues for yqfB: 4–101; PF0470: 20–86 and 100–110; BC4709: 7–143; yqbG: 3–73 and 100–110; yhgG: 3–51, 73–77; rps24e: 2–84; BH1534: 4–138; UFC1: 2–101, 119–159.

[n] Recall/precision/DP-scores (''NMR R-factors'') as defined in ref. 23.

[o] Most-favored regions/additional allowed regions/generously allowed regions/disallowed regions.

[p] Ref. 39. Except for yqfB, van der Waals violations were minimized yielding reduced clash scores and target function values.

considering tags and $^{13}C/^{15}N$ double-labeling, the masses of polypeptides expressed for the NMR structural studies ranged from 10 to 22 kDa (average: 16.2 kDa; Table 1). Approximate isotropic overall rotational correlation times, τ_r, between 5.1 and 11 ns (Table 1) were inferred from ^{15}N nuclear spin relaxation time $T_1/T_{1\rho}$ ratios (4), which demonstrates that these proteins are largely monomeric in solution.

NMR Data Collection Protocol. For each protein, five GFT NMR experiments were acquired for resonance assignment (see *Supporting Text*, Figs. 3–5, Scheme 1, and Table 2, which are published as supporting information on the PNAS web site) in conjunction with simultaneous 3D NOESY providing ^1H-^1H upper distance limit constraints. This strategy enables one to adapt measurement times to sensitivity requirements while obtaining high-dimensional spectral information and keeping the number of experiments small. (4,3)D $\mathrm{HNN}\underline{\mathrm{C}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha}/\underline{\mathrm{C}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha}\mathrm{(CO)NHN}$ (10) were selected for assignment of polypeptide backbone, and $^{13}C^{\beta}$ resonances, (5,2)D $\underline{\mathrm{HACACON}}\mathrm{HN}$ (8), or (4,3)D $\underline{\mathrm{H}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha\beta}\mathrm{(CO)NHN}$ were chosen for $^1H^{\alpha}$ or $^1H^{\alpha\beta}$ assignment, and aliphatic/aromatic (4,3)D $\underline{\mathrm{HC}}\mathrm{CH}$ served for side-chain assignment. All GFT NMR spectra (Table 1) were recorded on a Varian INOVA 600 spectrometer equipped with a cryogenic $^1H\{^{13}C,^{15}N\}$ triple resonance probe. For proteins dissolved in 90% H_2O/10% ^2H_2O containing 100 (400) mM NaCl at pH 6.5, this cryogenic probe increases sensitivity by ≈3-fold (≈2-fold) when compared with a conventional probe. Except for protein yqfB, 3D NOESY spectra were acquired on a Varian INOVA 750 spectrometer equipped with a conventional probe. Spectra were processed by using the program NMRPIPE (15).

Resonance Assignment Protocol. The program XEASY (16) is capable of processing GFT NMR peak lists encoding linear combinations of shifts (Fig. 3) and was used for spectral analysis. Sequential resonance assignments were achieved in three stages. (*I*) (4,3)D $\mathrm{HNN}\underline{\mathrm{C}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha}$ and $\underline{\mathrm{C}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha}\mathrm{(CO)NHN}$ are represented by two subspectra each. These spectra were analyzed as described in ref. 10 in conjunction with 3D NOESY for backbone and $^{13}C^{\beta}$ assignment. This step was initiated with the program AUTOASSIGN (17) for analysis of scalar connectivities and then completed manually. (*II*) Assignments of $^1H^{\alpha\beta}$ (or $^1H^{\alpha}$) were obtained from (4,3)D $\underline{\mathrm{H}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha\beta}\mathrm{(CO)NHN}$ [or (5,2)D $\underline{\mathrm{HACACON}}\mathrm{HN}$] as described in refs. 4 and 8. (*III*) Starting from $^1H^{\alpha\beta}$ (or $^1H^{\alpha}$) and $^{13}C^{\alpha\beta}$ shifts, the three subspectra of aliphatic and aromatic (4,3)D $\underline{\mathrm{HC}}\mathrm{CH}$ correlation experiments were analyzed in conjunction with 3D NOESY for nearly complete side-chain assignment.

Nuclear Overhauser Effect (NOE) Peak Assignment Protocol. Based on chemical shifts, the locations of regular secondary structure elements were identified (18), and a "starting peak list" was generated for 3D NOESY containing expected intraresidue, sequential, and α-helical medium range NOE peaks. This peak list was manually edited by visual inspection of the NOESY spectra, and subsequent manual peak picking was pursued to identify the remaining, primarily long-range NOEs. After peak integration, the programs CYANA (19, 20) and AUTOSTRUCTURE (21) were used in parallel to automatically assign long-range NOEs. Assignments identically obtained by both programs ("consensus assignments") were retained and established the starting point for manual completion of iterative NOE assignment, peak picking, and structure calculation.

Final Structure Calculations. Stereospecific assignments were obtained by using the FOUND and GLOMSA modules of DYANA (19). For residues located in regular secondary structure segments, ϕ and ψ backbone dihedral angle constraints were derived from chemical shifts by using the program TALOS (22). No hydrogen bond constraints were used. DYANA structure calculations were started with 100 random conformers, and the 20 conformers with the lowest target function values were selected.

Results

By using the protocol described above, eight NMR solution structures were solved. Table 1 provides a survey of measurement times, completeness of resonance assignments, and statistics for structure determination and validation. First, lower limits for NMR measurement times were established for protein yqfB. It was shown that ≈26 h of instrument time enabled high-quality structure determination if 3D NOESY is recorded with the cryogenic probe. Second, the resonance assignment protocol was tested with proteins yqfB (Table 1) and XCC2852 (see *Supporting Text*, Table 3, and Fig. 6, which are published as supporting information on the PNAS web site). For the latter, (5,2)D $\underline{\mathrm{HACACON}}\mathrm{HN}$ was replaced by (4,3)D $\underline{\mathrm{H}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha\beta}\mathrm{(CO)NHN}$ to also measure $^1H^{\beta}$ shifts before analysis of (4,3)D $\underline{\mathrm{HC}}\mathrm{CH}$. Third, the NOE assignment protocol was evaluated. Then, seven protein structures were solved to explore feasibility and robustness of high-throughput structure determination using the thus-standardized protocol. Comparably long measurement times were chosen initially and reduced after unnecessarily high signal-to-noise ratios were registered: 8.5 days of instrument time were invested for protein PF0470 (≈1 mM), but only ≈2–5 days were invested for yqbG (≈0.9 mM), yhgG (≈1.1 mM), and rps24e (≈1 mM). Finally, feasibility in the 20-kDa molecular mass range was documented with protein UFC1 (≈1 mM), for which NMR data were collected in 8.9 days (Table 1).

Resonance Assignment. 2D $[^{15}N,^1H]$-HSQC spectra (Fig. 1) show that the target proteins exhibit varying degrees of chemical shift dispersion, which is representative for a high-throughput pipeline. In several cases, significant $^{15}N/^1H^N$ shift degeneracy is encountered in the central region, but (4,3)D $\underline{\mathrm{C}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha}$-type experiments render spin system identification unambiguous because they encode 4D spectral information. Furthermore, about doubled dispersion is observed in the GFT dimension along $\omega_1(^{13}C^{\alpha};^{13}C^{\alpha\beta})$ (10) when compared with conventional CACB-congeners (7). As a result, (4,3)D $\underline{\mathrm{C}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha}$-based sequential assignment efficiently breaks both $^{15}N/^1H^N$ and $^{13}C^{\alpha/\beta}$ shift degeneracy, and nearly complete backbone and $^{13}C^{\beta}$ assignments were obtained for all proteins within 0.5–3 days of an expert's time (Table 1). While side-chain assignment with conventional 3D H(C)CH relies solely on correlation of $\Omega(^1H)$ detected along $\omega_1(^1H)$, (4,3)D $\underline{\mathrm{HC}}\mathrm{CH}$ affords correlation of $\Omega(^{13}C)$, $\Omega(^{13}C+^1H)$ and $\Omega(^{13}C$-$^1H)$ along $\omega_1(^{13}C;^1H)$. The resulting redundancy and improved resolution (10) ensures high robustness of side-chain assignment, addressing a critical bottleneck of the assignment process. Hence, nearly complete side-chain assignments were obtained in ≈1–5 days of an expert's time when using (5,2)D $\underline{\mathrm{HACACON}}\mathrm{HN}$/(4,3)D $\underline{\mathrm{H}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha\beta}\mathrm{(CO)NHN}$, (4,3)D $\underline{\mathrm{HC}}\mathrm{CH}$ and 3D NOESY. Chemical shift data were deposited in the BioMagResBank (Table 1).

NOE Peak Assignment. 3D NOESY provided in a single data set the information of all three 3D NOESY experiments routinely acquired for structure determination of $^{13}C/^{15}N$ labeled proteins (Table 1; for a quality assessment of the NOESY data, see *Supporting Text*). Typically, ≈20–35% of the peaks represent long-range NOEs. Of those, between 33% and 81% were assigned "by consensus" using the programs CYANA and AUTOSTRUCTURE (Table 1). This strategy yielded protein folds "within" a rms deviation (rmsd) value relative to the refined structure of ≈2 Å for backbone heavy atoms. Subsequent manual structure refinement was accomplished within 5–15 days of an expert's time per structure.

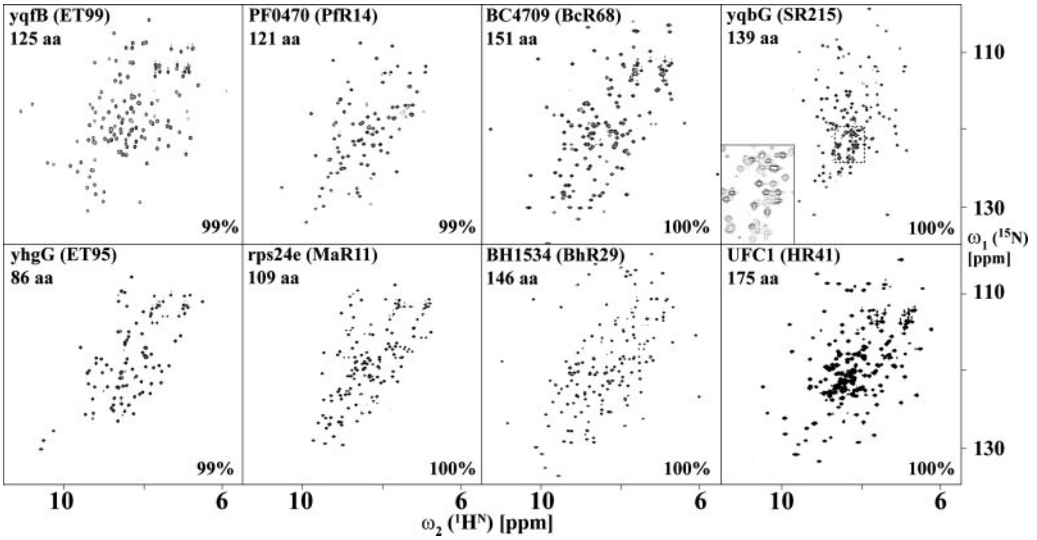

Fig. 1. Composite plot of 2D [^{15}N,^{1}H] HSQC spectra recorded at 750 MHz for target proteins. Gene name, NESG target ID, and number of amino acid residues (including tags) are indicated in the top left of each plot. At the lower right, the fraction of the peaks registered in these spectra is indicated for which sequence specific resonance assignments were obtained. For the highly α-helical protein yqbG (Fig. 2), the central region is expanded in an *Inset*.

Solution Structures and Quality Assessment. High quality is evidenced for all eight structures (Fig. 2 and Table 1) by (*i*) the small size and number of residual constraint violations, (*ii*) the low average rmsd values relative to the mean coordinates of 20 conformers, (*iii*) the large fractions of stereospecific assignments for β-methylene and the Val and Leu isopropyl moieties, (*iv*) high R-, P-, and DP-scores (23) indicating excellent agreement between experimental NOE peak lists and peak lists back-calculated from DYANA conformers, and (*v*) the fact that nearly all ϕ and ψ dihedral angles are located in the allowed regions of the Ramachandran plot (24). Coordinates were deposited in the Protein Data Bank (PDB) (ref. 25; see Table 1).

These scores suggest that structural quality is quite similar to high-quality NMR structures that were solved in recent years by other leading NMR groups using conventional NOE-based structure determination protocols. This view is supported by a comparison of "rapid" XCC2852 and a "conventional" NMR structure with their corresponding x-ray crystal structures (see *Supporting Text*, Tables 3–5, and Figs. 6 and 7, which are published as supporting information on the PNAS web site). In fact, considering that (*i*) virtually complete resonance assignments were obtained and validated by consistent structure calculations, and (*ii*) NOESY data collection and NOE peak assignment for rapid structure determination are accomplished in a "quasi-conventional" manner, one would not expect to encounter a "quality gap" between NMR structures solved either conventionally or with the protocol used for the present study.

Moreover, "bundles" of DYANA conformers sample the conformational space that is in agreement with experimental constraints and Van der Waals radii (19). In contrast, electrostatic interactions are not considered. Hence, DYANA conformers can be further refined. By using the program CNS (26), we performed short constrained molecular dynamics simulations in explicit solvent (27) (see *Supporting Text*, Tables 6–8, and Figs. 8 and 9, which are published as supporting information on the PNAS web site). The thus-refined NMR structures exhibit structural quality scores typically encountered for medium-resolution (1.8–2.5 Å) x-ray structures. Together, the protocol for rapid structure determination used here yields experimental constraint networks that are well suited for high-quality protein structure determination.

Each of the eight proteins analyzed here are the first representatives from protein domain families selected by the NESG consortium (28). Sequence similarity searches using the program PSI-BLAST (29) revealed that the target structures (Fig. 2) represent protein families with a total of 118 homologues from both eukaryotic and prokaryotic organisms. A search for structural homologues in the PDB using the programs DALI (30) and CE (31) revealed (with z-scores $>$ 4.0) that the target proteins belong to the following (super)families (Fig. 2) according to the "structural characterization of proteins" scheme (32): yqfB and PF0470, "PUA domain"; BC4709 and BH1534, "START domain"; yhgG, "winged-helix DNA-binding domain"; rps24e, "ribosomal protein L23 and L15e-like"; and UFC1, "Ubiquitin conjugating enzyme." For protein yqbG, no structurally similar protein was identified, suggesting that this protein possesses a hitherto-uncharacterized fold. Details and discussion of the implications of protein functions will be published elsewhere.

Discussion

Inspection of signal-to-noise ratios in NMR spectra of Table 1 shows that for $\approx$1 mM solutions of proteins with molecular masses $\approx$10–20 kDa [typical for monomeric domains targeted by structural genomics consortia (28)], NMR data collection times of $\approx$1–9 days suffice to ensure high-quality structure determination (Fig. 2). This amount of time is less than $\approx$10–25% of what was previously invested on a routine basis when using conventional probes (3, 33, 34). The rapid data collection enables one to increase structure production throughput and solve structures of slowly precipitating proteins such as BH1534 (Table 1).

GFT NMR affords nearly complete resonance assignments with $\approx$1–6 days of data collection time based on 4D and 5D spectral information encoded at high digital resolution (Fig. 3), which warrants robust data analysis even when encountering significant shift degeneracy. The minimal measurement time for the suite of five (4,3)D GFT NMR experiments used for most proteins of the present study (Table 1) is less than $\approx$20 h., whereas recording of the corresponding set of parent 4D Fourier transform NMR experiments would have taken $\approx$15 times longer. Strategies based on experiments encoding 3D NMR spectral information (e.g., ref. 3) are viable alternatives for assigning proteins with lower molecular masses in high through-

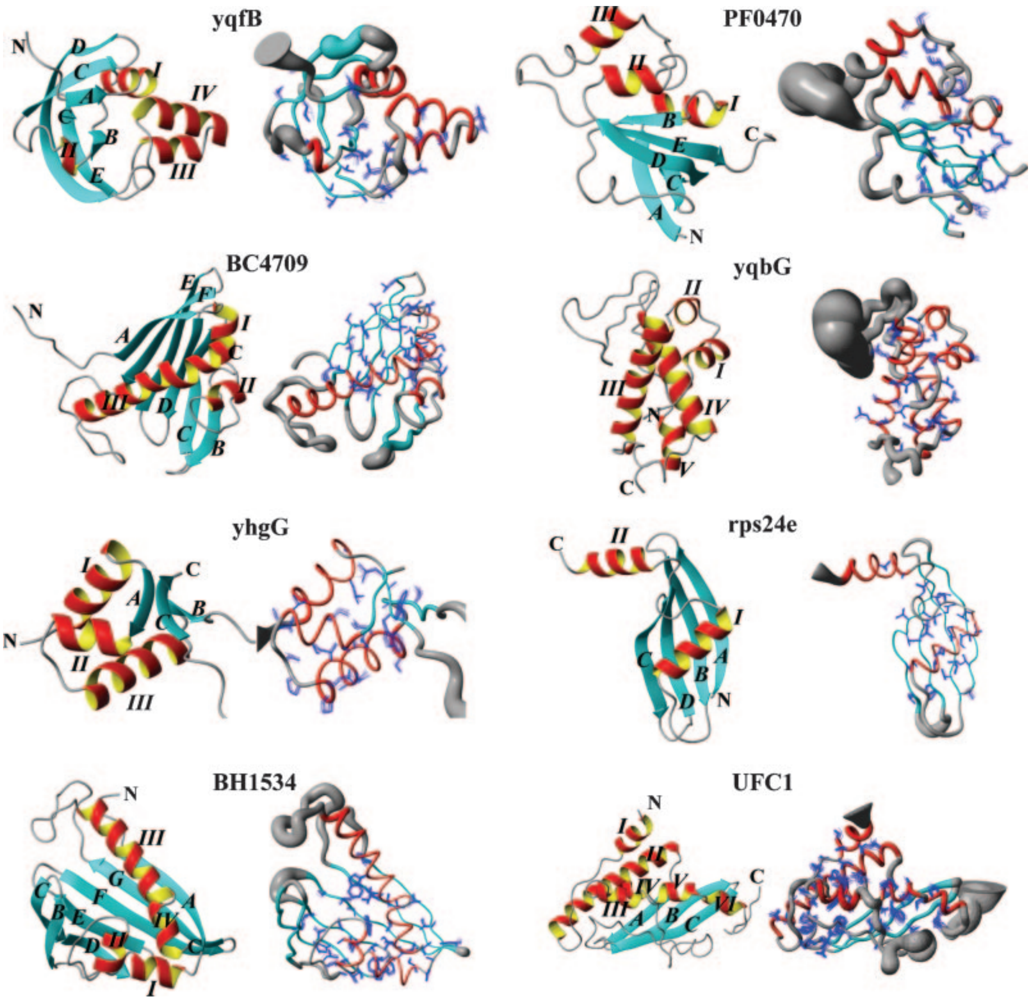

Fig. 2. High-quality NMR solution structures of target proteins are displayed in the order of Table 1. For each structure, a ribbon drawing is shown on the left. α-Helices are enumerated with roman numerals, and β-strands are indicated with letters (for sequence locations of the regular secondary structure elements, see footnote of Table 1). The N and C termini of the polypeptide chains are labeled with N and C. On the right, a ''sausage'' representation of the backbone is shown for which a spline function was drawn through the C$^{\alpha}$ positions and where the thickness of the cylindrical rod is proportional to the mean of the global displacements of the 20 DYANA conformers calculated after superposition of the backbone heavy atoms N, C$^{\alpha}$, and C′ of the regular secondary structure elements for minimal rmsd. Hence, the thickness reflects the precision achieved for the determination of the polypeptide backbone conformation. A superposition of the best-defined side chains having the lowest global displacement for the side-chain heavy atoms also are shown (best third of all residues; for residue numbers, see footnote of Table 1) to indicate precision of the determination of side-chain conformations. Helices are shown in red, the β-stands are depicted in cyan, other polypeptide segments are displayed in gray, and the side chains of the molecular core are shown in blue. The figure was generated by using the program MOLMOL (37).

put. In these cases, (3,2)D GFT NMR can provide within a few hours the information required for assignment (8–11). However, when compared with the current protocol based on 4D and 5D information, such strategies would not offer similar robustness in high throughput. It is of practical interest that manual analysis of (4,3)D GFT NMR experiments is quite generally less challenging than analysis of conventional congeners: G-matrix transformation edits shift doublets into subspectra (so that the total number of peaks per spectrum remains constant despite joint sampling of shifts) while peak dispersion increases because of observation of linear combinations of shifts. Notably, the high resolving power of (4,3)D HCCH renders ^{13}C total correlation spectroscopy (7) unnecessary.

Simultaneous 3D NOESY enabled detection of dense networks of ^{1}H-^{1}H upper distance constraints (Table 1). In such spectra, NOE assignment is greatly facilitated by having in a single data set each X1-H^{1} . . . H^{1}-X2 NOE resolved at the shift of X1 and the corresponding "transposed" peak resolved at the shift of X2. Moreover, the impact of distance constraints involving aromatic rings for structural refinement (1, 35) emphasizes the importance of including ^{13}C$^{\text{aromatic}}$-resolved NOEs in the simultaneous acquisition.

Conclusions

Protein sample preparation, NMR data collection, and data analysis and protein structure calculation have been recognized as major bottlenecks for high-throughput structure determination (2–4). Here we show, first, that collection of data providing 4D/5D NMR spectral information at high digital resolution for resonance assignment and 3D simultaneous NOESY for high-quality structure determination can routinely be accomplished in ≈1–9 days per structure. Secondly, ≈1–2 weeks of an expert's time are required for semiautomated data analysis and structure calculation. The design of the integrated data collection and analysis protocol is robust and effectively removes data acquisition as a bottleneck for rapid structure determination of proteins up to at least ≈20 kDa. Because NOE detection and assignment, as well as (conservative) derivation of experimental constraints are "conventionally" accomplished, the same precision is obtained as with established NOE-based protocols. Considering that (*i*) ≈95% of the NMR structure in the PDB are from proteins with masses <20 kDa, (*ii*) solving the solution structures of slowly precipitating proteins such as BH1534 is feasible only when collecting NMR data rapidly, and (*iii*) sensitivity of NMR spectrometers continues to increase, we expect that the protocol described here, or similar variants, will have high impact for NMR-based structural biology and structural genomics of globular and membrane proteins (36).

We thank Drs. Y. J. Huang, L. C. Ma and H. N. B. Moseley (Rutgers University) for helpful discussions. We acknowledge assistance from the University at Buffalo College of Arts and Sciences. This work was supported by National Science Foundation Grants MCB 0075773 and 0416899 (to T.S.), National Institutes of Health Protein Structure Initiative Grant P50 GM62413, and the University at Buffalo Center for Computational Research.

1. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids* (Wiley, New York).
2. Montelione, G. T., Zheng, D., Huang, Y., Gunsalus, C. & Szyperski, T. (2000) *Nat. Struct. Biol.* **7,** 982–984.
3. Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G. M., Bhattacharyya, S., Gutierrez, P., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 1825–1830.
4. Szyperski, T., Yeh, D. C., Sukumaran, D. K., Moseley, H. N. B & Montelione, G. T. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8009–8014.
5. Styles, P., Soffe, N. F., Scott, C. A., Cragg, D. A., White, D. J. & White, P. C. (1995) *J. Magn. Reson.* **60,** 397–404.
6. Monleon, D., Colson, K., Moseley, H. N. B., Anklin, C., Oswald, R., Szyperski, T. & Montelione, G. T. (2002) *J. Struct. Funct. Genomics* **2,** 93–101.
7. Cavanagh, J., Fairbrother, W. J., Palmer, A. G. & Skelton, N. J. (1996) *Protein NMR Spectroscopy* (Academic, San Diego).
8. Kim, S. & Szyperski, T. (2003) *J. Am. Chem. Soc.* **125,** 1385–1393.
9. Kim, S. & Szyperski, T. (2004) *J. Biomol. NMR* **28,** 117–130.
10. Atreya, H. S. & Szyperski, T. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 9642–9647.
11. Atreya, H. S. & Szyperski, T. (2005) *Methods Enzymol.* **394,** 78–108.
12. Pascal, S. M., Muhandiram, D. R., Yamazaki, T., Forman-Kay, J. D. & Kay, L. E. (1994) *J. Magn. Reson.* **103,** 197–201.
13. Xia, Y., Yee, A., Arrowsmith, C. H. & Gao, X. (2003) *J. Biomol. NMR* **27,** 193–203.
14. Acton, T. B., Gunsalus, K. C., Xiao, R., Ma, L. C., Aramini, J., Baran, M. C., Chiang, Y. W., Climent, T., Cooper, B., Denissova, N. G., *et al.* (2005) *Methods Enzymol.* **394,** 210–243.
15. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995) *J. Biomol. NMR* **6,** 277–293.
16. Bartels, C., Xia, T. H., Billeter, M., Güntert, P. & Wüthrich, K. (1995) *J. Biomol. NMR* **6,** 1–10.
17. Moseley, H. N. B., Monleon, D. & Montelione, G. T. (2001) *Methods Enzymol.* **339,** 91–108.
18. Wishart, D. S. & Sykes, B. D. & Richards, F. M. (1994) *Biochemistry* **31,** 1647–1650.
19. Güntert, P., Mumenthaler, C. & Wüthrich, K. (1997) *J. Mol. Biol.* **273,** 283–298.
20. Herrmann, T., Güntert, P. & Wüthrich, K. (2002) *J. Mol. Biol.* **319,** 209–227.
21. Huang, Y. J., Moseley, H. N. B., Baran, M. C., Arrowsmith, C., Powers, R.,Tejero, R., Szyperski, T. & Montelione, G. T. (2005) *Methods Enzymol.* **394,** 111–141.
22. Cornilescu, G., Delaglio, F. & Bax, A. (1999) *J. Biomol. NMR* **13,** 289–302.
23. Huang, Y. J., Powers, R. & Montelione, G. T. (2005) *J. Am. Chem. Soc.* **127,** 1665–1674.
24. Laskowski, R. A., Rullmann, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. (1996) *J. Biomol. NMR* **8,** 477–486.
25. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2002) *Nucleic Acids Res.* **28,** 235–242.
26. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., *et al.* (1998) *Acta Crystallogr. D* **54,** 905–921.
27. Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M. & Nilges, M. (2003) *Proteins* **50,** 496–506.
28. Liu, J., Hegyi, H., Acton, T. B., Montelione, G.T. & Rost, B. (2004) *Proteins* **56,** 188–200.
29. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
30. Holm, L. & Sander C. (1995) *Trends Biochem. Sci.* **20,** 478–480.
31. Shindyalov, I. N. & Bourne, P. E. (1998) *Protein Eng.* **11,** 739–747.
32. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
33. Markley, J. L., Ulrich, E. L., Westler, W. M. & Volkman, B. F. (2003) *Methods Biochem. Anal.* **44,** 89–113.
34. Adams, M. W. W., Dailey, H. A., Delucas, L. J., Luo, M., Prestegard, J. H., Rose, J. P. & Wang, B. C. (2003) *Acc. Chem. Res.* **36,** 191–198.
35. Skalicky, J. J., Mills, J. L., Sharma, S. & Szyperski, T. (2001) *J. Am. Chem. Soc.* **123,** 388–397.
36. Atreya, H. S., Eletski, A., Szyperski, T. (2005) *J. Am. Chem. Soc.* **127,** 4554–4555.
37. Koradi, R., Billeter, M. & Wüthrich, K. (1996) *J. Mol. Graphics* **14,** 51–55.
38. Nicastro, G., Margiocco, P., Cardinali, B., Stagnaro, P., Cauglia, F., Cuniberti, C., Collini, M., Thomas, D., Pastore, A. & Rocco, M. (2004) *Biophys. J.* **87,** 1227–1240.
39. Word, J. M., Bateman, R. C., Presley, B. K., Lovell, S. C. & Richardson, D. C. (2000) *Protein Sci.* **9,** 2251–2259.