IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Protein NMR Structure Refinement based on Bayesian Inference

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2016 J. Phys.: Conf. Ser. 699 012005

(http://iopscience.iop.org/1742-6596/699/1/012005)

View the table of contents for this issue, or go to the journal homepage for more

Download details:

IP Address: 212.51.137.249
This content was downloaded on 08/04/2016 at 06:46

Please note that terms and conditions apply.

Protein NMR Structure Refinement based on Bayesian Inference

Teppei Ikeya[1,2*], Shiro Ikeda[3], Takanori Kigawa[2,4], Yutaka Ito[1,2] and Peter Güntert[1,5,6*]

[1]Tokyo Metropolitan University, 1-1 Minami-ohsawa, Hachioji, Tokyo 192-0397, Japan
[2]CREST/Japan Science and Technology Agency (JST), Japan
[3]The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[4]Laboratory for Biomolecular Structure and Dynamics, RIKEN Quantitative Biology Center (QBiC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
[5]Institute of Biophysical Chemistry and Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany
[6]Laboratory of Physical Chemistry, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland

E-mail: tikeya@tmu.ac.jp, guentert@em.uni-frankfurt.de

Abstract. Nuclear Magnetic Resonance (NMR) spectroscopy is a tool to investigate three-dimensional (3D) structures and dynamics of biomacromolecules at atomic resolution in solution or more natural environments such as living cells. Since NMR data are principally only spectra with peak signals, it is required to properly deduce structural information from the sparse experimental data with their imperfections and uncertainty, and to visualize 3D conformations by NMR structure calculation. In order to efficiently analyse the data, Rieping et al. proposed a new structure calculation method based on Bayes' theorem. We implemented a similar approach into the program CYANA with some modifications. It allows us to handle automatic NOE cross peak assignments in unambiguous and ambiguous usages, and to create a prior distribution based on a physical force field with the generalized Born implicit water model. The sampling scheme for obtaining the posterior is performed by a hybrid Monte Carlo algorithm combined with Markov chain Monte Carlo (MCMC) by the Gibbs sampler, and molecular dynamics simulation (MD) for obtaining a canonical ensemble of conformations. Since it is not trivial to search the entire function space particularly for exploring the conformational prior due to the extraordinarily large conformation space of proteins, the replica exchange method is performed, in which several MCMC calculations with different temperatures run in parallel as replicas. It is shown with simulated data or randomly deleted experimental peaks that the new structure calculation method can provide accurate structures even with less peaks, especially compared with the conventional method. In particular, it dramatically improves in-cell structures of the proteins GB1 and TTHA1718 using exclusively information obtained in living *Escherichia coli* (*E. coli*) cells.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

1. Introduction

Biomacromolecules such as proteins, DNA and RNA have three-dimensional (3D) structures and dynamical properties, and provide essential functions contributing to almost all biological events and phenomena. It is notably interesting that the structures and dynamics of proteins have been designed exquisitely during their long evolution process in order to accurately perform individual vital tasks. To elucidate the molecular structures and properties, therefore, allows to understand mechanisms of biological activities and human diseases at atomic level, and to promote drug discovery and protein engineering.

There are currently three main methods to determine 3D structures of biomacromolecules at atomic resolution, X-ray crystallography, cryo-electron microscopy (cryo-EM) and nuclear magnetic resonance (NMR) spectroscopy. X-ray crystallography is the most popular and powerful as a 3D structure determination tool for biomacromolecules. It can investigate large molecules in detail routinely once a crystal of the target molecule has been obtained. However, it requires to produce a crystal of the molecule, which is usually not trivial, and can observe only the crystal state in which the molecule is packed, immobilized and occasionally skewed by surrounding molecules in the crystal lattice despite of the importance of the molecular dynamics for the functions of proteins. Cryo-EM currently becomes popular to observe large macromolecular assemblies because it has virtually no limitation on upper molecular size. Its drawback of inherently low resolution has recently ameliorated considerably and cryo-EM is approaching atomic resolution [1]. NMR spectroscopy is the second most commonly used tool to investigate the biomacromolecules. It observes the precession of nuclear spins as nuclear resonance of emitted electromagnetic radiation under strong magnetic fields, which is obtained by a superconducting magnet and modulated by the surrounding electronic environments. By NMR, one can determine, with atomic resolution, the chemical configurations, structures, and dynamics of a molecule.

A principal advantage of NMR is that biological macromolecules can be studied non-invasively in their native form in nearly natural environment, and that besides static structural information also dynamic aspects of protein function can be investigated on timescales ranging from nanoseconds to months. It does not require preparing crystals of the molecules like X-ray, and can obtain structural information at atomic resolution. A large number (thousands) of signals can be recorded and resolved in a single NMR spectrum by spreading them in multiple dimensions. Furthermore, a method to investigate protein structures in living cells by NMR [2, 3], so-called in-cell NMR, has recently been developed, which uncovers new pictures of protein behaviors in the cells. A principal limitation of solution NMR spectroscopy relates to the size of the macromolecule. Due to their slow molecular tumbling larger molecules exhibit rapid transverse relaxation that leads to rapidly decaying signals and broad resonance lines. These are manifested in the spectra by weak, broad and overlapping peaks, which eventually impedes their detection and assignment. Therefore, for structure determination, the NMR method is in general limited to soluble proteins below 30 kDa of molecular weight, or to even smaller masses for proteins that have a large effective size because of their environment, e.g. proteins within living cells. It is of high interest to overcome the molecular size limitation and to establish more accurate structure determination of such difficult molecule samples.

Since the original NMR data are principally only peak intensities and positions in a spectrum, it is necessary to properly deduce the structural information from the data with its imperfections and uncertainties, and to visualize 3D conformations by NMR structure calculation. The structure calculation can be a model fitting approach starting from a model structure that embodies the information of chemical composition and bonds of a molecule, which is generated by some other method. In the case of proteins, the initial model is a conformation of an extended linear amino acid chain, which is optimized to find atomic coordinates that satisfy expected distances and dihedral angles derived from the experimental NMR data. Conventional structure determination is typically achieved by a molecular dynamics simulation (MD) approach with distance and dihedral angle restraints derived from the data, in which a target function is defined as the sum of the (suitably simplified) physical potential energy of a molecule and the sum of squared errors between experimental data and molecular conformations, with a weight factor to keep a balance between them.

Various NMR structure calculation methods have been proposed to determine more accurate 3D structures for proteins of lager molecular size or in cells with sparse and noisy information. Nilges and coworkers presented in 2005 a NMR structure calculation method based on Bayes theorem, so-called Inferential Structure Determination (ISD) [4, 5]. Since, as noted above, the raw NMR experimental data are not direct structural information, it is necessary to convert the NMR data into restraints on distances and dihedral angles among atoms by relations that are based on physical principals but contain also some nuisance parameters. Whereas these parameters are fixed to user-defined values in the conventional methods, the Bayesian method can deduce them with a statistical model by parameter sampling schemes. In a Bayesian NMR structure calculation, the target function is replaced by the posterior, and the contributions of prior and likelihood are computed from the data without a predefined explicit weight factor.

The program CYANA [6, 7] is widely used for NMR structure calculations as well as occasionally for molecular modeling in *de novo* design of proteins. As an optimization method for the target function, it adopts torsion angle molecular dynamics (TAMD) that enables to obtain converged structures quickly with longer step-sizes than MD in Cartesian coordinate space. The potential energy of CYANA is a physical force field optimized for torsion angle space derived from the Cartesian space force field of the MD program Amber [8] and with a Generalized Born (GB) implicit water model [9]. The physical force field and water model can achieve a more accurate estimation of the prior distribution of the structure ensemble, in which TAMD can reduce the computational cost for obtaining the marginal likelihood as well as molecular conformations. Here, we present a new NMR structure refinement method based on Bayesian inference implemented into CYANA. We employed the Bayesian modeling to Nuclear Overhauser Effect (NOE) data, which has the form of restraints on distances between two hydrogen atoms that are shorter than approximately 5 Å. The method is composed of a MD calculation to quickly obtain rough global structures with automatic NOE assignment, and subsequent structure refinement by Bayesian inference addressing both unambiguously and ambiguously assigned NOE peak lists. The final conformational ensembles are evaluated by principal component analysis (PCA).

Three different types of peak lists were prepared as input data for the validation of the new method. (1) Fully simulated data reconstructed from known structures. (2) Sparse experimental data obtained by randomly deleting a given amount of peaks from the original experimental data set. (3) Complete in-cell NMR data sets that are intrinsically sparse and noisy and one of the main targets of this method. Using fully simulated and experimental data allow us in-depth analyses and verification for practical non-ideal data, respectively.

2. Theory

The target function (T) of a conventional NMR structure calculation usually consists of the χ^2-term between predicted and observed data, the physical potential energy (E), and a weight factor (w):

$$T(\theta) = \chi^2(\theta) + wE(\theta), \tag{1}$$

where θ denotes the set of torsion angles used to describe the protein structure. In Bayes' theorem, the target function can be replaced by the posterior probability for the evaluation of an ensemble of conformations:

$$P(\theta|D) \propto P(\theta)P(D|\theta) \tag{2}$$

where θ is coordinates in torsion angle space, and D is the experimental data. In modern NMR structure determination, the most essential experimental data yielding 3D structure information is the NOE peak volume V_{kl} derived from a dipolar interaction between spin k and l, and hence in this study the NOE data is described in the framework of Bayesian inference, while the other experimental data are treated in conventional forms, e.g. dihedral angle restraints. The likelihood function is described by a lognormal distribution

$$V_{kl} = \frac{\gamma}{r_{kl}{}^{6}} \tag{3}$$

$$P(V|\theta) = \prod_{i=1}^{n} L(V_{k_i l_i}|\theta) \tag{4}$$

$$L(V_{kl}|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}V_{kl}} \exp\left(-\frac{1}{2\sigma^2}\ln^2\left(\frac{V_{kl}}{\mu}\right)\right) \tag{5}$$

where *n* is the number of peaks *σ* is the standard deviation and *μ* is an expected peak volume. It is necessary to convert the peak volume into distance information between two spins r_{kl}. In the conventional approach this is achieved using the simple equation (3) with a calibration constant γ. Thus, the prior of the Bayesian modeling is

$$P(\sigma,\theta,\gamma) = P(\sigma)P(\theta)P(\gamma) \tag{6}$$

$$P(\theta) = \frac{1}{Z(\beta)}\exp(-\beta E(\theta)) \tag{7}$$

$$\sigma \sim G[a,b]$$
$$\gamma \sim LN[\mu_\gamma,\sigma_\gamma]$$

where *a* and *b* are shape and scale parameters of the Gamma function *G*, respectively, *LN* is the lognormal distribution, and *P*(*θ*) is described as the canonical ensemble of molecular structures with partition function *Z* and inverse temperature *β*.

The sampling algorithm for obtaining the posterior is based on Markov chain Monte Carlo (MCMC) with the Gibbs sampler for the *σ* and *γ* parameters, and MD for sampling the canonical ensemble of conformations. In a general sense, this MC method along with MD is known as hybrid MC or Hamiltonian MC (HMC) in Bayesian statistics, by which high acceptance rates on Metropolis criteria can be achieved even for large parameter dimensions. While the general HMC method introduces an independent auxiliary variable describing a pseudo kinetic energy term for effective sampling in MC, the method here is a more direct combination of MC and MD. The MD method is used for sampling molecular conformations because it is known that its sampling of protein conformations is far superior to MC due to covalent structure restrictions and the tightly packed globular shape [10, 11]. The target function *T*(*θ*) in MD is composed of the physical potential *E*(*θ*) and a pseudo energy *L*(*θ* , *γ*, *σ*) with *γ* and *σ*, which are sampled by the Gibbs sampler,

$$T(\theta) = \beta E(\theta) + L(\theta,\gamma,\sigma) \tag{8}$$

$$E(\theta) = \sum E_{dihedral} + \sum E_{vdw} + \sum E_{vdw14} + \sum E_{electro} + \sum E_{electro14} + \sum E_{GB}$$

$$L(\theta,\gamma,\sigma) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\ln\left(\frac{V_i r_i^{\,6}}{\gamma}\right)$$

where $E_{dihderal}$, E_{vdw}, E_{vdw14}, $E_{electro}$, $E_{electro14}$, and E_{GB} are energy terms for dihedral angles, van der Waals, 1-4 van der Waals, electrostatic, 1-4 electrostatic, and Generalized Born (GB) implicit water model interactions.

In the case of an ambiguously assigned distance restraint with *m* assignment possibilities, the distance r_i is replaced by

$$\bar{r} = \left(\sum_{j=1}^{m} r_{k_j l_j}{}^{-6} \right)^{-1/6} \tag{9}$$

3. Material and Methods

Figure 1 shows a flow chart of this method. The initial step of the structure calculation is principally identical with the conventional structure calculation with automatic Nuclear Overhauser Effect SpectroscopY (NOESY) cross peak assignments by the program CYANA [12]. The calibration constant, *γ*, is determined such that the median value of all peak volumes or intensities in a given NOESY spectrum corresponds to a predefined distance, usually 4.0 Å. All upper distance bound restraints are calculated from the corresponding peak intensity in the spectrum using the calibration constant. The conformation with the lowest target function value is selected from the 20 conformers calculated with the distance restraints obtained in the 7th cycle, and regularized for the molecular geometry of the Amber ff03 force field which is slightly different from the original CYANA geometry used in the initial step. The regularization was achieved by recalculating the structures in CYANA using distance and torsion angle restraints. Distance restraints with an upper bound of 0.1 Å were used for the distances to the corresponding N, C$^\alpha$, and C' atoms in the structured regions of the conformers after the initial 7th cycle of CYANA. Torsion angle restraints were set with a width of 20° centered around the value of each torsion angle in that structure [13]. The regularized coordinates were used for the next refinement stage based on Bayesian inference and as reference structures for the calculation of root-mean-square deviations (RMSDs).

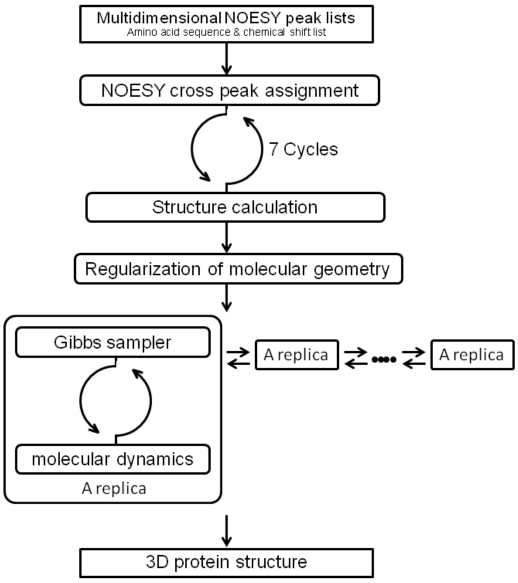
Multidimensional NOESY peak lists
Amino acid sequence & chemical shift list
NOESY cross peak assignment
7 Cycles
Structure calculation
Regularization of molecular geometry
Gibbs sampler
molecular dynamics
A replica
A replica
A replica
3D protein structure

Figure 1. Structure refinement by Bayesian inference.

A replica in Fig. 1 denotes a hybrid Monte Carlo (HMC) calculation for sampling *θ*, *γ*, *σ*, and evaluating the posterior. The torsion angle coordinates, *θ*, are sampled by molecular dynamics simulation with the current *γ* and *σ*, in which the Amber physical force field with the GB model is modified for torsion angle space and used for sampling the canonical ensemble. The calibration constant and standard deviation of NOE volumes, γ and σ, are sampled with the current θ by the Gibbs sampler. These three parameters are optimized by alternatively performing MD and the Gibbs sampler in iterative steps. In the practical sampling of these parameters, it is not trivial to search the entire functional space particularly on exploring the conformational prior due to the extraordinarily large conformational space of proteins. Hence, the replica exchange method is employed, in which several MCMC calculations with different temperatures are performed in parallel as replicas, and each is exchanged at certain intervals.

The accuracy of structures was quantified by the RMSD value to the reference structure for the backbone atoms N, C$^\alpha$, and C' in the structured regions of the proteins. Bayesian modelling allows to obtain a posterior probability for all parameters including atomic coordinates and NOE calibration constants, and final structure ensembles can be validated based on the posterior distribution. Since the parameter space is high-dimensional, principal component analysis (PCA) were performed to visualize the distributions of structure ensembles. Cartesian coordinates of the atoms in the final structure ensembles were used as input variables for PCA. The posterior distribution on the principal components elucidates structure ensembles from the estimated distribution due to wider search of the conformational space satisfying current experimental data by MCMC. In here, we show all conformers within 1 standard deviation of the maximum of the posterior. Structures were visualized using the program MOLMOL [14].

3.1. Conventional structure calculation as reference

Conventional structure calculations were done using the standard protocol implemented in CYANA [7, 12]. 100 initial conformers with random torsion angle values were subjected to simulated annealing using 10000 torsion angle dynamics steps, and the 20 lowest target function conformers were selected to represent the NMR structure of the protein.

For comparison, we alternatively refined the CYANA structures with the program OPALp [15], which we routinely calculate to obtain final structures deposited to Protein Data Bank (PDB). The program can perform the conjugate gradient minimization in Cartesian space with Amber ff94 and TIP3P implicit water model. For comparison of the conventional and Bayesian algorithms, the same peak lists were given to both approaches.

3.2. Data set generation for structure calculation

Two types of simulation data were prepared to validate this method: (1) Fully simulated data that are reconstructed from known structures. (2) Experimental data that were modified by randomly deleting peaks. The first data sets permit to evaluate the predictive accuracy of the method, and the second ones can assess its performance in more realistic cases.

The fully simulated peak lists were derived from 3D structures of the proteins TTHA1718 (66 aa, PDB 2ROE) [2] and GB1 (57 aa, PDB 2J52) [16]. The peak intensities were calculated from the structures according to the relationship $V_{kl} = \gamma\ r_{kl}^{-6}$ for all atom pairs within the range 2.4–5.0 Å for ^{13}C-separated, ^{13}C-separated aromatic and ^{15}N-separated NOESYs. The individual peaks were given slightly different calibration constant values taken from a Gaussian distribution centered at the given γ. Table 1 shows the numbers of simulated peaks for TTHA1718 and GB1 that were randomly selected from atom pairs fulfilling the above criteria. Other restraints like hydrogen bond and torsion angle were not used. The data sets were designed so as to reflect challenging conditions for the conventional structure determination by CYANA, e.g. cases with a severely limited amount of structural information such as in-cell NMR. The peak lists were assigned with the original atom pairs in order to validate the performance of the conventional and new methods exclusively in terms of the quality and amount of distance information, but excluding contributions from the accuracy of the CYANA automatic NOE assignment algorithm.

Reduced experimental peak lists were generated by randomly deleting given amounts of the NMR distance restraint data for three different proteins, i.e. TTHA1718, the 140-residue ENTH-VHS domain At3g16270(9–135) from *Arabidopsis thaliana* (ENTH) [17, 18], and the 114-residue Src homology domain 2 from the human feline sarcoma oncogene Fes (SH2) [19]. The amounts of deleted peaks were chosen such that structures calculated by the conventional CYANA method were disturbed by 1.0–4.0 Å RMSD from the original ones, since such data sets are the prime practical targets for the present structure refinement method. Considering the dynamics of molecules and imperfections of experimental data, it does not make sense to attempt obtaining structures with less than 1.0 Å RMSD. On the other hand, it is not trivial to dramatically improve structures with more than 4.0 Å RMSD because the conformational space that should be search for this would become vast. The manually determined chemical shift assignments were used from the BMRB with accession numbers 11035 for TTHA1718, 5928 for ENTH, and 6331 for SH2.

The chemical shifts for GB1 were determined in our group. Peak lists were obtained with the automated peak picking algorithm of the program NMRView [20] and Azara [21] with manual corrections or modifications for ^{15}N-, and ^{13}C-resolved NOESY spectra. Torsion angle and hydrogen bond restraints were not used in any of the calculations.

In-cell NMR experiments and data collection were performed as reported previously [2, 3].

4. Results and Discussion

4.1. Fully simulated peak lists of TTHA1718 and GB1

First, we validated the performance of the method with the fully simulated NOE peak lists from 3D structures of TTHA1718 and GB1. Structure calculations using either the conventional CYANA-OPALp approach or the Bayesian refinement were independently performed with these data sets. Table 1 shows the γ and σ values used for simulating peaks, and the predicted values by CYANA and the Bayesian method. Assuming that the peak intensities include not only distance information but also various other physical properties and noise, they were produced on fluctuations by normal distribution with the standard deviation σ. Figure 2 shows the structures derived from the two methods and distributions of the calibration constants in TTHA1718 generated by the Bayesian method. The fluctuations of the calibration constants and standard deviations are insensitive to the number of iterations in the Gibbs sampler, showing that the number of iterations were sufficient (Fig. 2C). In addition, it was confirmed that the replicas with the highest and lowest temperatures exchanged one or more times during the replica exchange MC process, and all variables as well as the conformational space of the protein were sampled sufficiently (data not shown). The expected γ and σ values were computed from the distributions (Table 1). Although it is not trivial to infer those parameters because the process requires simultaneously distance information among atoms, the results demonstrate that the Bayesian method determined those parameters rather precisely in all spectra of both TTHA1718 and GB1. There is no σ in the CYANA calculations since γ was determined such that the median value of all peak intensities in a given NOESY spectrum corresponds to a predefined distance of 4.0 Å. These γ values by CYANA are relatively accurate despite the fairly naïve presumption for obtaining the distances. Nevertheless, the Bayesian method shows better predictions than CYANA in all cases. The peak list of the ^{13}C-separated aromatic NOESY of TTHA1718 contained only 9 peaks, so that the predicted values differ slightly more from the original than those for the other spectra. The estimated σ are slightly smaller than the original values used to simulate the data, which means that the predicted distributions of γ were narrower and less ambiguous or underestimated.

Table 1. Parameters set for creating simulated NOESY peak lists

	peaks[a]	γ			σ		
		set[b]	CYANA	Bayesian	set[b]	CYANA	Bayesian
TTHA1718:							
^{13}C-NOESY	86	20.74	19.51	20.69	2.39	-	2.20
^{13}C-aro NOESY[c]	9	17.05	19.28	17.72	5.71	-	4.70
^{15}N NOESY	61	12.27	11.25	12.23	3.30	-	3.05
GB1:							
^{13}C-NOESY	97	20.53	19.56	20.65	2.33	-	1.23
^{13}C-aro NOESY[c]	15	13.17	11.70	12.96	4.76	-	2.82
^{15}N NOESY	41	12.32	10.95	12.34	3.74	-	2.68

[a] Number of used peaks

[b] γ and σ values used for creating the fully simulated NOESY peak lists.

[c] ^{13}C-separated aromatic NOESY

While the conventional method selects the 20 conformers with the lowest target function values from 100 computed with different random seeds, the Bayesian approach predicts the posterior distribution and allows to choose structure ensembles from the estimated distribution due to wider search on conformational space satisfying current experimental data by MCMC. In here, we exhibit all conformers within 1 standard deviation of the maximum of the posterior. Although Figure 2a shows 564 conformers of TTHA1718, the bundle structures are noticeably more converged than the 20 of the conventional method (Figure 2a), and closer to the reference structure deposited in PDB. The tube models depicting the coordinates of C$^{\alpha}$ atoms clearly demonstrate that the ensemble of the new method is less broadening than those of the conventional. Whereas the backbone RMSDs to the mean and the reference by CYANA were 1.62 and 2.22 Å, respectively, the Bayesian approach improved the accuracy of the structures to 1.45 Å of backbone RMSD to the reference.

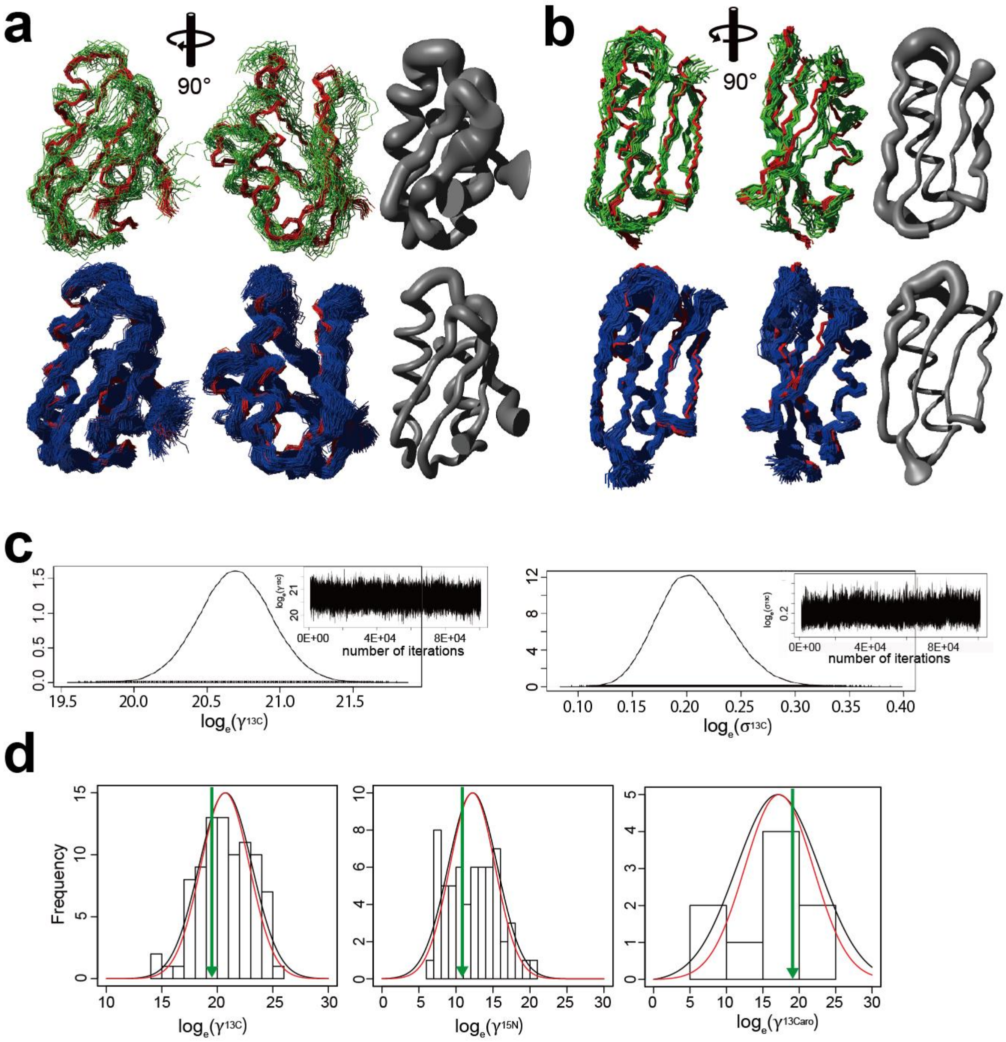

Figure 2. Bayesian structure calculations with fully simulated data of TTHA1718 and GB1. Superpositions of the 20 reference structures from the full data sets (red) conformations yielded by the Bayesian method (blue), showing the backbone (N, C$^\alpha$, C') atoms. Green superpositions are the 20 structures by the conventional method. Deviations of C$^\alpha$ atoms in the Bayesian-refined conformers are shown as tube models. (**a**) TTHA1718. (**b**) GB1. (**c**) Distributions of γ and σ for the ^{13}C-NOESY of TTHA1718. Insets show trajectories of γ and σ in the HMC sampling. (**d**) Histograms and distribution curves of γ for the ^{13}C-, ^{15}N-, and ^{13}C aromatic-NOESY spectra. The distributions of γ in the input peaks (black) and predicted by the Bayesian method (red), as well as the values determined by CYANA (green arrows) are shown.

These results demonstrate that the new Bayesian method for refining protein structures enables a more accurate interpretation of peak intensities than the conventional CYANA approach even in the case of having less structural information in the case that the data are composed of relatively ideal distributions of the peak intensities without additional noise signals. Next, we validated it for more realistic data containing noise signals and non-ideal NOE distributions.

4.2. Imperfect experimental peak lists of four different proteins

The robustness of the algorithm for more realistic data was investigated with respect to missing peaks. Starting from experimental data sets for the proteins TTHA1718, SH2, and ENTH, approximately 55–90% of the peaks were randomly deleted. (see Methods).

Table 2. RMSD values of three proteins

run	1		2		3	
	conv.[a]	Bayes	conv.	Bayes	conv.	Bayes
ENTH:						
Peaks[b]	1229/6823		1010/6823		903/6823	
RMSD to mean (Å)[c]	1.48/2.01	0.69/1.06	1.92/2.46	0.74/1.18	2.13/2.80	1.24/1.70
RMSD to reference (Å)[d]	1.92/2.21	1.25/1.86	2.41/2.74	1.04/1.65	3.51/3.77	1.86/2.47
SH2:						
Peaks	2832/5422		2175/5422		2033/5422	
RMSD to mean (Å)	1.71/2.28	0.65/0.96	2.19/2.76	1.13/1.57	2.92/3.45	1.38/1.80
RMSD to reference (Å)	1.39/1.90	1.27/1.82	2.24/2.76	2.28/2.83	3.10/3.55	2.25/2.65
TTHA1718:						
Peaks	472/3205		348/3205		299/3205	
RMSD to mean (Å)	1.47/2.03	0.84/1.32	2.13/2.83	0.82/1.33	2.43/3.18	1.52/2.06
RMSD to reference (Å)	1.57/1.89	1.55/1.97	2.40/2.77	1.32/2.00	3.21/3.59	2.36/2.85

[a] conventional structure determination by CYANA
[b] Number of used/original peaks
[c] RMSD to the mean structure of the structure ensemble (backbone/sidechain)
[d] RMSD to the reference structure (backbone/all heavy atoms)

Results of the calculations used with three different data sets for ENTH, SH2 and TTHA1718 are summarized in Table 2. The reference structures of the three proteins were obtained by conventional CYANA calculations and OPALp refinement with the original experiment data set. The RMSD to the reference structure depended on the proteins and simulated data set. The Bayesian refinements exquisitely resulted in lower RMSDs to the mean and reference than the conventional structure calculations except for the RMSD of SH2 to the reference. Representative structure ensembles in the three independent calculations, run 3 of ENTH, run 3 of SH2 and run 2 of TTHA1718, clearly demonstrate that the Bayesian-refined structures approach the reference structure more closely than those of the conventional method (Fig. 3). The structures with 2–4 Å RMSD by the conventional method were most improved by the Bayesian method, whereas the improvement was less significant for the ones with approximately 1 Å RMSD to the reference, probably owing to the dynamics of molecules and imperfections of experimental data.

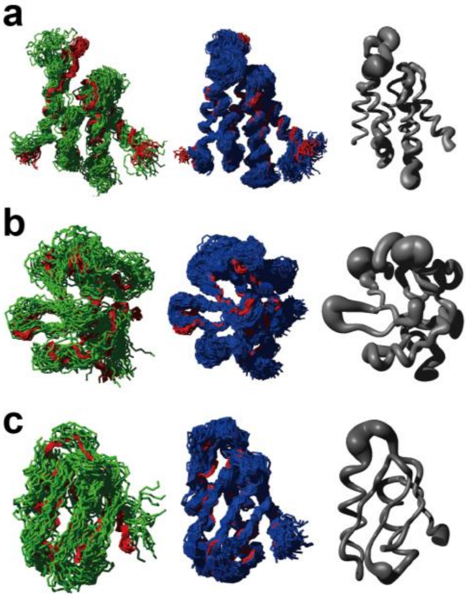

Figure 3. Structures bundles of 3 proteins calculated from reduced data sets by the conventional CYANA method (green) and Bayesian refinement (blue) superimposed on conventionally determined reference structures from the full data set (red), (a) ENTH. (b) SH2. (c) TTHA1718. Deviations of C^{α} atoms in the Bayesian-refined conformers are shown as tube models.

4.3. In-cell NMR structure determination

One of the targets of this method is to determine protein structures in living cells, so called in-cell NMR structure determination. So far, we obtained 3D NOESY spectra of two proteins in living cells, TTHA1718 and GB1, which were of sufficient quality for structure calculation. In the case of the first in-cell structure determination of the protein TTHA1718 [2, 3], we employed backbone hydrogen bond restraints for the β-sheet region where their existence was indicated by inter-strand NOEs. While this approach has been used also for *in vitro* NMR structure determinations, it may obscure deviations from canonical secondary structure manifested in the experimental data because it explicitly fixes standard secondary structure hydrogen bonds for ranges of residues identified by spectroscopists. It is instructive to improve in-cell NMR by our present approach of data-driven structure determination with minimum prior information. Thus, the in-cell TTHA1718 structures were calculated by the Bayesian approach without the hydrogen bond restraints. Since the chemical shift differences between *in vitro* and in-cell were relatively small for these proteins except for the metal binding region of TTHA1718, the accuracy of the in-cell structures could be validated by comparison with the *in vitro* conformers. The Bayesian approach was compared to the conventional refinement method with CYANA and OPALp.

Figure 4a shows the final, Bayesian-refined structures of GB1 in living cells. The Bayesian calculation computes the posterior distribution by replica exchange Monte Carlo simulation, which can efficiently sample the marginal likelihood (Fig. 4b,c). The structure ensemble was validated by PCA, in which the distribution of the first principal component (PC1) was composed of one major and other minor populations in the vicinity of the major region (Fig. 4d). The minor populations include structures that differ from the independently determined *in vitro* structure by approximately 1.6 Å RMSD. This suggests that PCA could efficiently exclude inaccurate ensembles. Thus, we chose from the posterior distribution 1416 conformations within 1 standard deviation from the mean of PC1 (Fig. 4a,d), whereas the 20 structures with the lowest energies are shown as the NMR structure ensemble in the conventional structure determination. The conventional structure ensemble is well converged with an average backbone RMSD of 0.40 Å to the mean structure. The backbone RMSD between the mean structure and the *in vitro* structure is 1.17 Å. Selecting the 20 highest posterior probability structures for comparison with the conventional method, the RMSD of these 20 structures is 0.49 Å to its mean and 1.02 Å to the *in vitro* structure. In addition, Bayesian inference provided the distributions of the calibration constants and their standard deviations for the three NOESY spectra. These distributions are broader for the ^{13}C/^{13}C-separated NOESY than for the ^{15}N-separated and ^{13}C-separated NOESYs, presumably due to the smaller number of peaks (Fig. 4e,f).

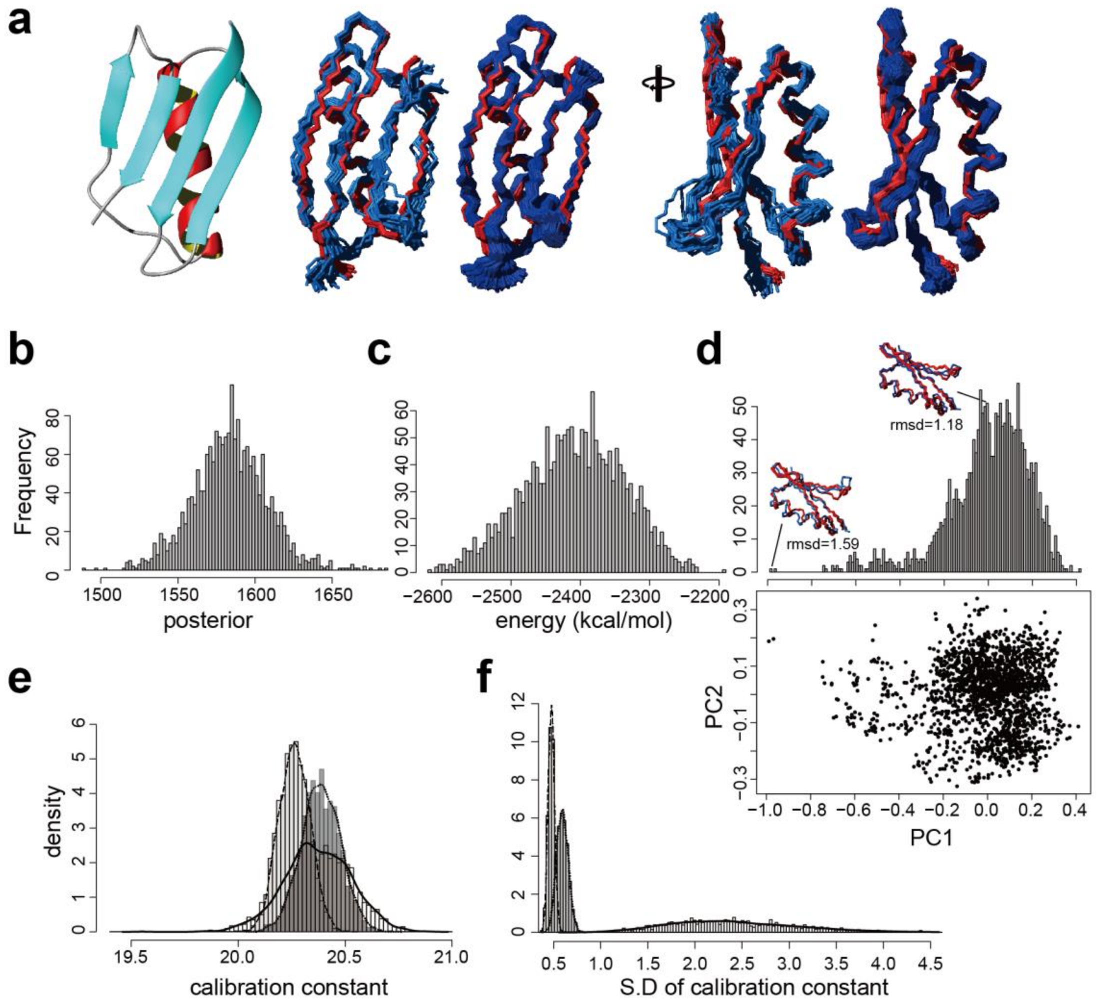

Figure 4. NMR structure of the protein GB1 in living *E. coli* cells. (**a**) Ribbon diagram of the structure of GB1 in living *E. coli* cells with the highest posterior (left). Superpositions of the 20 final structures of purified GB1 in vitro (red) and the 1416 conformations representing the in-cell structure of GB1 within one standard deviation of the first principal component of the posterior distribution (blue). Green superpositions are the conventional CYANA structure calculation. (**b**) Posterior distribution of the in-cell GB1 conformations. (**c**) Potential energy distribution of the in-cell GB1 conformations. (**d**) Distributions of the first principal component (top) and the first and second ones (bottom). (**e**) Distributions of the calibration constants computed by the Bayesian method of ^{13}C/^{13}C-separated (bold), ^{13}C-separated (dashed), and ^{15}N-separated (dotted) NOESY spectra. (**f**) Corresponding distributions of the standard deviation of the calibration constants.

5. Conclusion

In summary, the results of this paper show that the Bayesian-based structure calculation implemented into CYANA enables to significantly improve structures calculated by the original CYANA/OPALp method. A currently routinely used approach for NMR structure determination using CYANA is to obtain global structures by TAMD and automated NOESY cross peak assignment, and to refine the structures by other software like OPALp, CNS or Amber. However, those structure calculation and subsequent refinement methods have considered the weight factor of the potential energy and the calibration constant describing the relationship between NOE intensities and atom-atom distances exclusively in the initial stage by user-predefined or naive approaches. The weight and calibration constants are closely associated with possible conformational ensembles, and conversely those conformations are restricted by distance restraints derived from those parameters. Thus, it is rational to compute the structures and parameters simultaneously from the experimental data. Since it usually requires substantial calculation time to obtain global folds of proteins and accurate NOE assignments due to the recursive steps of the structure calculation and assignment, our method first performs the conventional structure calculation with a simplified target function to obtain the global fold, and then refines these structures by the Bayesian method.

In addition, it is cumbersome for many spectroscopists to use several software packages in which parameters and data formats are different, due to strong parameter dependence, non-standard format requirements, lack of documentation for conversion from one software to others, or high computation time demands. In contrast, using the CYANA refinement algorithm is straightforward. Moreover, seamless processes without manual intervention are a prerequisite for fully automatic structure determination like the FLYA method. The full automatic calculation is usually composed of automatic peak picking, chemical shift and NOE assignment, structure calculation, and refinement. All the stages should be principally validated by structures and experimental data, and a recursive approach of the stages can be the most reliable.

A principal advantage of the Bayesian refinement algorithm is that it searches a wider conformational space with the other explanatory parameters by the replica exchange and Gibbs sampler methods. It can also be used with solid state NMR data, for structure-based assignment, or resonance assignment based exclusively on NOESY spectra, which will be treated elsewhere.

References

[1] Wang R Y R, Kudryashev M, Li X M, Egelman E H, Basler M, Cheng Y F, Baker D and DiMaio F 2015 *Nat. Methods* **12** 335–338

[2] Sakakibara D, Sasaki A, Ikeya T, Hamatsu J, Hanashima T, Mishima M, Yoshimasu M, Hayashi N, Mikawa T, Wälchli M, Smith B O, Shirakawa M, Güntert P and Ito Y 2009 *Nature* **458** 102–105

[3] Ikeya T, Sasaki A, Sakakibara D, Shigemitsu Y, Hamatsu J, Hanashima T, Mishima M, Yoshimasu M, Hayashi N, Mikawa T, Nietlispach D, Wälchli M, Smith B O, Shirakawa M, Güntert P and Ito Y 2010 *Nat. Protoc.* **5** 1051–1060

[4] Rieping W, Habeck M and Nilges M 2005 *Science* **309** 303–306

[5] Habeck M, Nilges M and Rieping W 2005 *Phys. Rev.* E **72** 031912

[6] Güntert P 2003 *Prog. Nucl. Magn. Reson. Spectrosc.* **43** 105–125

[7] Güntert P, Mumenthaler C and Wüthrich K 1997 *J. Mol. Biol.* **273** 283–298

[8] Duan Y, Wu C, Chowdhury S, Lee M C, Xiong G M, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J M and Kollman P 2003 *J. Comput. Chem.* **24** 1999–2012

[9] Baker N A 2005 *Curr. Opin. Struct. Biol.* **15** 137–143

[10] Yamashita H, Endo S, Wako H and Kidera A 2001 *Chem. Phys. Lett.* **342** 382–386

[11] Northrup S H and McCammon J A 1980 *Biopolymers* **19** 1001–1016

[12] Güntert P and Buchner L 2015 *J. Biomol. NMR* **62** 453–471

[13] Kirchner D K and Güntert P 2011 *BMC Bioinformatics* **12** 170

[14] Koradi R, Billeter M and Wüthrich K 1996 *J. Mol. Graphics* **14** 51–55

[15] Koradi R, Billeter M and Güntert P 2000 *Comput. Phys. Commun.* **124** 139–147

[16] Frick I M, Wikström M, Forsén S, Drakenberg T, Gomi H, Sjöbring U and Björck L 1992 *Proc. Natl. Acad. Sci. USA* **89** 8532–8536

[17] López-Méndez B, Pantoja-Uceda D, Tomizawa T, Koshiba S, Kigawa T, Shirouzu M, Terada T, Inoue M, Yabuki T, Aoki M, Seki E, Matsuda T, Hirota H, Yoshida M, Tanaka A, Osanai T, Seki M, Shinozaki K, Yokoyama S and Güntert P 2004 *J. Biomol. NMR* **29** 205–206

[18] Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Sugano S, Yokoyama S and Güntert P 2005 *J. Biomol. NMR* **31** 357–361

[19] Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Sugano S, Yokoyama S and Güntert P 2004 *J. Biomol. NMR* **30** 463–464

[20] Johnson B A and Blevins R A 1994 *J. Biomol. NMR* **4** 603–614

[21] Boucher W 2010 Azara v. 2.8, in, Cambridge, UK