J Biomol NMR (2015) 63:255–264
DOI 10.1007/s10858-015-9982-z

ARTICLE

A robust algorithm for optimizing protein structures with NMR chemical shifts

Mark Berjanskii[1] · David Arndt[1] · Yongjie Liang[1] · David S. Wishart[1,2,3]

Received: 1 June 2015 / Accepted: 27 August 2015 / Published online: 7 September 2015

Abstract Over the past decade, a number of methods have been developed to determine the approximate structure of proteins using minimal NMR experimental information such as chemical shifts alone, sparse NOEs alone or a combination of comparative modeling data and chemical shifts. However, there have been relatively few methods that allow these approximate models to be *substantively* refined or improved using the available NMR chemical shift data. Here, we present a novel method, called Chemical Shift driven Genetic Algorithm for biased Molecular Dynamics (CS-GAMDy), for the robust optimization of protein structures using experimental NMR chemical shifts. The method incorporates knowledge-based scoring functions and structural information derived from NMR chemical shifts via a unique combination of multi-objective MD biasing, a genetic algorithm, and the widely used XPLOR molecular modelling language. Using this approach, we demonstrate that CS-GAMDy is able to refine and/or fold models that are as much as 10 Å (RMSD) away from the correct structure using only NMR chemical shift data. CS-GAMDy is also able to refine of a wide range of approximate or mildly erroneous protein structures to more closely match the known/correct structure and the known/correct chemical shifts. We believe CS-GAMDy will allow protein models generated by sparse restraint or chemical-shift-only methods to achieve sufficiently high quality to be considered fully refined and "PDB worthy". The CS-GAMDy algorithm is explained in detail and its performance is compared over a range of refinement scenarios with several commonly used protein structure refinement protocols. The program has been designed to be easily installed and easily used and is available at http://www.gamdy.ca.

Keywords Protein · Structure · Accuracy · NMR · Chemical shifts

Introduction

Using NMR to solve the structures of larger proteins (>15kD), weakly soluble proteins or disordered proteins is often complicated by the poor quality of their NMR spectra and, consequently, the small number of experimental restraints. As a result, protein structure determination from sparse NMR data has been a very active area of research for more than two decades. While most sparse-data methods have focused on finding intelligent ways to use limited numbers of distance restraints from Nuclear Overhauser Effects (NOE), there has been increased interest in using chemical shifts to help solve the sparse restraint problem. Initially, chemical shifts were only used for secondary structure restraints (Wishart and Sykes 1994; Wishart et al. 1992) or torsion angle constraints (Berjanskii et al. 2006; Cheung et al. 2010; Shen et al. 2009) to help supplement NOE data. More recently, chemical shifts, either alone or in combination with other non-NOE data (such as RDCs),

Electronic supplementary material The online version of this article (doi:10.1007/s10858-015-9982-z) contains supplementary material, which is available to authorized users.

✉ David S. Wishart
david.wishart@ualberta.ca

[1] Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada

[2] Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada

[3] National Research Council, National Institute for Nanotechnology (NINT), Edmonton, AB T6G 2M9, Canada

have been used to determine or refine 3D protein structures. Indeed, impressive results have been reported with programs such as Cheshire (Cavalli et al. 2007), CS-Rosetta (Shen et al. 2008), CS23D (Wishart et al. 2008), CS-MD (Robustelli et al. 2010), and CS-Torus (Boomsma et al. 2014). Typically, these methods rely on supplementing chemical shift data with pre-existing information, such as knowledge-based scoring functions, structures of homologous proteins or protein fragments.

While encouraging progress continues to be made in the field (Rosato et al. 2015), many protein structures generated by these sparse-restraint modeling techniques still are only approximately correct or have obvious structural errors. Consequently, most structures generated via sparse-restraint methods or chemical-shift-only methods are looked upon by the NMR community as working structural "hypotheses". Indeed, less than 20 of the 11,000 NMR structures deposited in the PDB have been solved using sparse restraint methods. Efforts to use traditional structure determination and refinement programs, such as XPLOR-NIH (Schwieters et al. 2003), CNS (Brunger et al. 1998), or CYANA (Guntert 2004) to improve these structure have rarely succeeded. Improvements are only seen if large numbers of additional NMR restraints (usually NOE) are added. Also, it is still very computationally challenging to refine, optimize or otherwise improve the initial models using sparse NMR data (Robustelli et al. 2010). As a result, the potential time savings or experimental simplification offered by chemical-shift only or other sparse restraint NMR methods for protein structure determination have largely remained unrealized.

Ideally, what is needed is a robust, easy-to-use program that can take approximate protein structures, such as those generated via comparative modeling, CS23D, CS-Rosetta or even sparse NOE data, and use the existing experimental NMR data (primarily chemical shifts) to further optimize and improve the structure. Here, we present just such a program, called CS-Chemical Shift driven Genetic Algorithm for biased Molecular Dynamics (GAMDy). CS-GAMDy is a hybrid molecular dynamics (MD) program that combines knowledge-based potentials with conventional MD-based NMR modelling to perform robust chemical shift refinement and structural optimization. Because derivatives cannot be calculated from many experimental parameters and knowledge-based potentials, CS-GAMDy employs a novel combination of multi-objective MD biasing and a genetic algorithm (GA) to perform its model optimization (Fig. 1).

The molecular dynamics in CS-GAMDy is performed using the XPLOR-NIH molecular modelling package (Schwieters et al. 2003), which is one of the most commonly used structure determination programs in the protein NMR community. Thus, CS-GAMDy can be easily

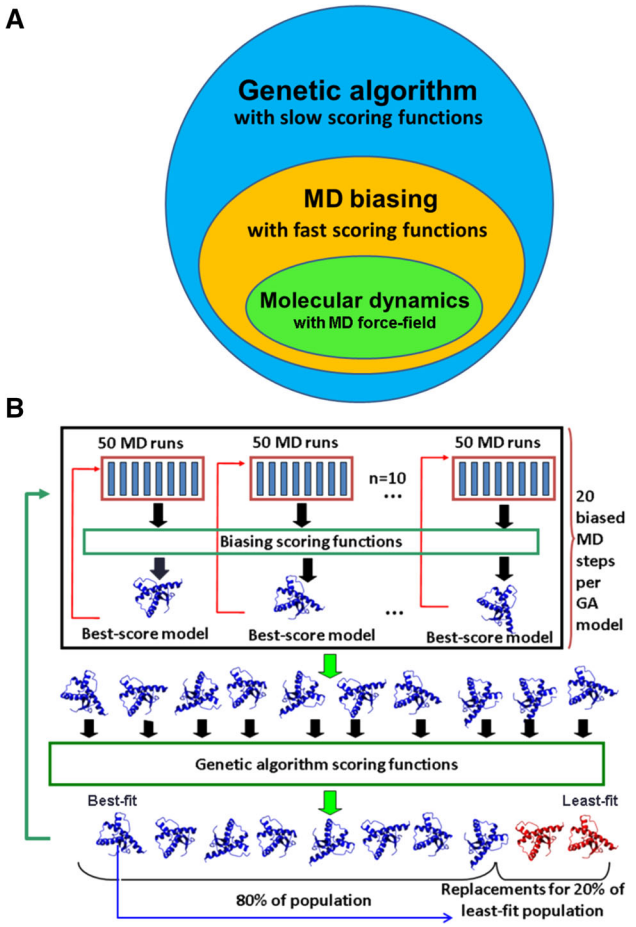

Fig. 1 CS-GAMDy protocol. **a** The main components of CS-GAMDy, **b** MD biasing and genetic algorithms in CS-GAMDy. See text for details

adopted to use a wide range of restraints commonly employed in XPLOR refinement methodologies. CS-GAMDy also allows users to take advantage of the latest knowledge-based scoring functions, such as GOAP (Zhou and Skolnick 2011), RW (Zhang and Zhang 2010), GeNMR (Berjanskii et al. 2009), and various MD force-fields [e.g. CHARMM (MacKerell et al. 1998), Amber (Cornell et al. 1995), and OPLS (Jorgensen and Tirado-Rives 1988)]. CS-GAMDy is also the first program to incorporate the Random Coil Index (RCI) and a novel RCI-ASA score to improve the agreement between model's accessible surface area (ASA) and the ASA derived from chemical shifts by RCI (Berjanskii and Wishart 2013).

Here, we describe the CS-GAMDy algorithm in detail and discuss its performance for optimizing approximate or moderately incorrect protein structures (such as those generated via comparative modelling, 3D-threading, NOE-only methods or chemical shift-only methods like CS23D or CS-Rosetta) using NMR chemical shifts as the only source of experimental information. We demonstrate that

CS-GAMDy is able to refine and/or fold protein models that are, in some cases, as much as 10 Å (RMSD) away from the reference structure using only NMR chemical shift data. Based on its performance over a wide range of refinement scenarios, we believe CS-GAMDy will allow protein models initially generated by sparse restraint or chemical-shift-only methods to achieve sufficiently high quality to be considered fully refined and worthy of PDB deposition.

Materials and methods

A brief summary of the CS-GAMDy protocol

CS-GAMDy consists of three major components: (1) quenched, restrained molecular dynamics, (2) multi-objective MD biasing by experimental data and knowledge-based scores, and (3) a multi-objective genetic algorithm, as shown on Fig. 1a. Briefly, multiple MD runs are performed at each MD biasing step. The final models of the MD runs are evaluated and ranked by computationally fast scoring functions. The model with the best score is used as the starting model for the next set of MD runs (Fig. S1; Fig. 1b). A population of several independent MD biasing trajectories is generated and the final model of each trajectory is assessed with the use of computationally more demanding fitness functions. The least-fit models in the population are replaced by the best-scoring (i.e. most fit) models (Fig. 1b). This process is repeated until certain CS-GAMDy stop criteria are met (vide infra). We describe the details of these three components and testing the CS-GAMDy protocol below.

CS-GAMDy has four operational modes: (A) the default, full mode, with both the MD biasing and genetic algorithm active; (B) a mode with only the genetic algorithm for unbiased MD; (C) a mode with only MD biasing; (D) a mode with only molecular dynamics (Fig. S2). Instructions for how to switch between different operational modes is given below. All tests in this paper were done with the default CS-GAMDy mode. The other modes can be used to optimize or test individual parts of the CS-GAMDy framework by developers or advanced users.

Molecular dynamics

The molecular dynamics protocols in CS-GAMDy were programmed with the XPLOR-NIH molecular modelling language (Schwieters et al. 2003) and Python. We chose XPLOR-NIH because it is one of the most popular, well-tested programs for NMR-based protein structure modelling and refinement. Also, because of its ability to accept the majority of modern types of NMR experimental restraints, the XPLOR-NIH molecular modelling package can be used for almost any kind of model optimization with NMR data. The most important MD parameters are listed in Table S2 and described below.

CS-GAMDy allows users to select a variety of MD force-fields that come with XPLOR-NIH including the CHARMM force-fields (versions 11, 19, and 22), Amber94, OPLS, as well as the PARALLHDG force-field of XPLOR (Schwieters et al. 2003) that is commonly used in NMR-based structure determination. In our preliminary tests, we found that selecting the PARALLHDG force-field in CS-GAMDy leads to the best model accuracy (data not shown). This is likely because we used MD conditions (a high virtual temperature) that are similar to those of a typical NMR structure determination protocol. Therefore, we used the PARALLHDG force-field as the default in CS-GAMDy. We have made the other force-fields available to developers and advanced users. CS-GAMDy also provides support for employing XPLOR's knowledge-based database potentials (Kuszewski et al. 1996, 1997), a self-guiding hydrogen bond potential (Grishaev and Bax 2004), and a radius of gyration energy term (Kuszewski et al. 1999). All of these potentials are enabled by default and were used to generate the results described in this paper.

Cartesian coordinate molecular dynamics is the default MD method of choice in CS-GAMDy. XPLOR's torsion angle dynamics (Stein et al. 1997) is also supported but it was found to produce less accurate results (data not shown) for most types of models that we tested in this work.

CS-GAMDy permits XPLOR MD simulations to be conducted in a vacuum or with a Generalized Born implicit solvent (Wagner and Simonson 1999). Interestingly, using implicit solvent in combination with experimental restraints, did not result in improved accuracy with CS-GAMDy (data not shown). This may be due to the fact that the solvent slows down the model refinement process and reduces the positive effects of MD biasing and genetic algorithm optimization on model quality. Therefore, MD simulations in CS-GAMDy are conducted *in vacuo* by default.

In order to generate conformational changes of various amplitudes and directions, many starting parameters for XPLOR's molecular dynamics can be randomized. In fact, this is the preferred way to run CS-GAMDy because the magnitude of optimal conformational changes to refine a protein model is not known a priori. By default, randomization is automatically applied to several MD parameters, such as velocities, the length of the MD run, the MD time-step, the temperature, the contributions of the torsion angle restraints, the radius of gyration, and the electrostatic contribution to the XPLOR force-field.

Molecular dynamics runs are quenched using Powell's minimization (Powell 1977) to remove temperature-

induced distortions of the model's local structure. The minimization is critical to properly evaluate and rank the models by CS-GAMDy's scoring functions since the scores were optimized on protein models with proper local geometry. Molecular dynamics is the part of the CS-GAMDy protocol where NMR- and template-based restraints in an XPLOR-compatible format can be applied. In this work, we utilized torsion angles derived from chemical shifts (Shen and Bax 2015) to restrain XPLOR's molecular dynamics. Standard deviations of torsion angles were used to define the torsion angle restraint errors. CS-GAMDy can run only XPLOR molecular dynamics without MD biasing and without the genetic algorithm if the number of biased MD runs and the genetic algorithm population are both equal to 1 (Fig. S2). This mode can be used to conduct traditional NMR structure refinement.

MD biasing

MD biasing in CS-GAMDy is conducted using the CONTRA MD biasing method (CONformational TRAnsitions by Molecular Dynamics with minimum biasing) (Harvey and Gabb 1993). We chose this technique because it allows a user to bias a MD program "as is", without changing its code or force-field. CS-GAMDy is the first example of a successful application of the CONTRA MD method with collective variables derived from chemical shifts or knowledge-based normality scores. In short, this approach involves generating multiple MD trajectories starting from the same initial model but with different initial velocities. Once the trajectories are generated, their final models are assessed and ranked by a scoring function and the model with the best score is used as the initial model for the next iteration of the CONTRA MD protocol (Fig. S1; Fig. 1b). The simulations are terminated after a certain number of biasing iterations (default is 10), which is typically a compromise between biasing efficiency and available computational time. The most significant settings for MD biasing in CS-GAMDy are documented in Table S3 and discussed below.

In CS-GAMDy, we use a multi-objective version of the CONTRA MD protocol. This simply means that we use more than one scoring function to assess the MD models. For each biasing iteration, we randomly select one of the two scoring functions: a GeNMR scoring function (Berjanskii et al. 2009) and a RCI-derived accessible surface area score (RCI-ASA) as seen in Table S1. We selected these two scoring schemes because they are quick to calculate and permit the use of both experimental information (raw NMR chemical shifts, secondary structure derived from NMR chemical shifts, and RCI-based ASA) and pre-existing knowledge (e.g. threading scores, normality of the Ramachandran plot, omega angle normality, etc.). Technically, protein models in the MD biasing step can be ranked by the scoring functions that are used in the genetic algorithm of the CS-GAMDy protocol (vide infra). However, these scoring functions are computationally demanding and not recommended for MD biasing.

The number of individual MD runs per biasing iteration can be specified by the user. One MD run per biasing iteration means no biasing is done by CS-GAMDy and only the genetic algorithm is enabled (if the genetic algorithm population > 1, Fig. S2) or only pure MD is performed (if the genetic algorithm population $= 1$, Fig. S2). In practice, the number of MD runs should depend on the speed of the biasing scoring programs, available computational time, and the model difficulty. If simulations do not improve the value of the scoring function significantly, a user may want to consider increasing the number of MD runs per biased MD iteration to capture less abundant conformations that can help to lead the model to better refinement paths. The current default number of MD runs per iteration is set to 50, which appears to be a reasonable compromise between performance and computing time for the examples described in this paper.

Genetic algorithm

A multi-objective genetic algorithm was implemented to manage biased MD runs that get stuck in local energy minima. We observed that some biased MD simulations fail to optimize the same starting protein models that other biased MD runs with identical MD conditions (except MD velocities) can refine. The "unlucky" runs fail to achieve a satisfactory level of optimization no matter how long the simulations are continued. We found that we could achieve a better overall performance if we run multiple MD biasing runs and periodically replace "unlucky" runs with successful ones. The most important parameters for CS-GAMDy's genetic algorithm with their default values are listed in Table S4 and explained below.

In CS-GAMDy, each iteration of the genetic algorithm evolves a population of several biased MD models (Fig. 1b). All trajectories are periodically scored and ranked with a scoring function. A portion of the population (20 %) with the worst scores gets is replaced by the model with the best score. Due to the randomization of MD parameters (e.g. temperature, velocities, time steps, etc.), the best-scoring model and its "clones" follow different optimization paths during the next iteration of the genetic algorithm. This helps to maintain diversity in the population of protein models. Mutations in the CS-GAMDy genetic algorithm correspond to changes in atom coordinates during XPLOR molecular dynamics. The magnitude of these mutations depends on local defects in protein models (as sensed by the chosen MD force-field) and on

global MD parameters, such as temperature, time step, length of MD runs, etc. We do not perform coordinate cross-overs in CS-GAMDy because they can result in severe atom overlaps and high energies that, in turn, can cause significant model distortions and simulation crashes.

Since model evaluation in the genetic algorithm happens less frequently than in MD biasing, we can afford to use scoring functions that are more computationally expensive than those in MD biasing. At each step, all models in the population are assessed and ranked by a scoring function that is randomly selected from the computationally demanding scoring functions, such as GOAP (Zhou and Skolnick 2011) and RW (Zhang and Zhang 2010), and the less computationally demanding GeNMR and RCI-ASA functions. The use of multiple knowledge-based scoring methods from different research groups can help to minimize structural distortions due to imperfections or inaccuracies in a particular scoring function. The size of the genetic algorithm population can also be changed, depending on available computational resources and model difficulty. The current default size of a population is 10. It was sufficient for optimizing most models in this work. A larger population size may help if simulations struggle to meet success criteria. To disable the genetic algorithm and use only MD biasing in CS-GAMDy, the population size should be set to 1 (Fig. S2).

Termination conditions and simulation time

In order to decide when CS-GAMDy runs can be stopped, we normally monitor the GeNMR scoring function as an indicator of structural changes. We terminate simulations when the GeNMR target function levels off and does not significantly change for a long period of time (i.e. five times longer than the initial function decay, see Fig. S3).

The time that is required for a model optimization in CS-GAMDy can vary (e.g. from several hours to >100 h), depending on the quality of the starting structure and experimental data, protein size, computational resources, and selected parameters of MD, biased MD and genetic algorithm. In this work, a single CS-GAMDy run took on average 72 CPU hours on a single 2.6 GHz CPU computer with 3 GB RAM. This level of time and CPU requirements is typical for sparse-data modelling methods where the lack of complete experimental data often leads to a shallow energy landscape, which requires a more time-consuming sampling of conformational space (i.e. the "no free lunch" principle). Because CS-GAMDy's genetic algorithm is easily parallelized and readily adapted to larger multi-core installations, the time needed to perform these refinements will be substantially shortened in future program distributions.

Success criteria

To assess whether a refinement has been successful or not, we use well-established criteria for experimentally restrained protein modelling that are commonly used for CS-Rosetta simulations: the RMSD criterion and the score-drop criterion (Raman et al. 2010; Shen et al. 2008; Thompson et al. 2012). First, we rank all output models by the GeNMR score and take a cluster of ten models with the best (i.e. lowest) GeNMR score. During the next step, we identify the best-score model in this cluster and measure backbone RMSD of rigid secondary structure elements [α-helices and β-sheets as identified by CSI (Wishart and Sykes 1994; Wishart et al. 1992)] of the remaining models with respect to this best-scoring model. If the average backbone RMSD of the nine models is within 1.5 Å from the best-scoring model, we consider the RMSD criterion satisfied (Fig. S4, black lines). In order to perform the score-drop test, we conduct simulations with the same parameters and inputs but exclude the experimental data. If we observe that the average GeNMR score of the simulations with the experimental data is better than the average GeNMR score of the simulations without experimental data, we consider the score-drop criterion satisfied. Both the RMSD and score-drop criteria need to be met for a simulation to be considered successful. (Fig. S4, green lines).

For some starting models with poor GeNMR score values (above 0), indications of success or failure can be obtained from the Pearson correlation coefficient between the GeNMR score and the backbone RMSD to the best-scoring model (Fig. S4, red lines). Successful simulations often have correlation coefficients above 0.5, whereas failed simulations have correlation coefficients near 0. While this criterion can be useful to evaluate CS-GAMDy success for models with significant 3D distortions (non-coil backbone RMSD to the reference model >3Å), it frequently fails for refinement of near-native models (non-coil backbone RMSD to the reference model <2 Å). To assess the uncertainty of the CS-GAMDy results, we run ten or more independent CS-GAMDy simulations. If an ensemble of the best-scoring models from five successful runs (see the success criteria above) has a backbone RMSD to the ensemble mean within 2 Å, we consider the uncertainty of the CS-GAMDy results to be acceptable.

Testing CS-GAMDy

A total of four tests of the CS-GAMDy protocol were performed (vide infra). The experimental data provided to CS-GAMDy consisted of chemical shift derived torsion angles and secondary structures (Shen and Bax 2015), ASA

(Berjanskii and Wishart 2013), and chemical shift scores from the GeNMR scoring function (Berjanskii et al. 2009). Hence, success or failure of chemical shift refinement was estimated by monitoring violations of dihedral angle restraints, secondary structure score, RCI-ASA score, and Pearson correlation between predicted and experimental chemical shifts (Tables S7–13, S15–18). Model coordinate errors were estimated by the backbone RMSD of non-coil regions with respect to the native protein structure (Tables 1, 2, 3; Fig. 2). To compare the performance of CS-GAMDy with a common model refinement in XPLOR (Schwieters et al. 2003), XPLOR simulations was done using torsion angle restraints predicted from chemical shifts by TALOS-N (Shen and Bax 2015) and an XPLOR script for gentle refinement (refine_gentle.inp) that comes with XPLOR-NIH distributions.

Limitations

As with any data-driven method, the accuracy of CS-GAMDy's results will be limited by the quality or accuracy of the input data (i.e. "garbage in equals garbage out"). Poorly estimated torsion angles will have a greater impact on CS-GAMDy's performance than errors in ASA or secondary structure. This is because torsion angle restraints are used during every MD step while ASA and secondary structure restraints are used for model assessment less frequently. More specifically, ASA and secondary structure are used only in the MD biasing and the genetic algorithm, so any inaccuracies will have a somewhat smaller effect on the quality of CS-GAMDy's results.

CS-GAMDy is currently limited to refining monomeric proteins without any ligands. Efforts are underway in our lab to extend CS-GAMDy to multimeric proteins and complexes of proteins with small ligands or/and with other proteins. While CS-GAMDy can technically take distance restraints in XPLOR format (i.e. with a flag "–noe"), its conformational sampling and scoring functions have not yet been optimized for handling distance restraints. Therefore, this option should be used with some caution. Currently CS-GAMDy does not accept any XPLOR restraints other than torsion angle restraints, radius of gyration, and distance restraints.

CS-GAMDy installation

CS-GAMDy can be installed on any modern Linux computer with 3 GB RAM, at least 1 GB of hard-drive space, Python 2, and a GCC compiler. Users will also need to obtain and install several third-party programs, most importantly XPLOR-NIH, GOAP, and RW. An installation script is included with the program. Installation instructions can be found in the README file that comes with a CS-GAMDy distribution (located at www.gamdy.ca).

Results and discussion

For the first test, CS-GAMDy and the XPLOR refinement were evaluated on their ability to refine protein models that were deliberately distorted by unrestrained dynamics. We started from models of ubiquitin that were misfolded with

Table 1 Model accuracy of the distorted protein models under different refinement scenarios

Protein name	PDB ID	Model accuracy (backbone RMSD to the reference model, Å)				
		Initial model	Refined by CS-MD	Refined by XPLOR with NMR data	Refined by CS-GAMDy without NMR data	Refined by CS-GAMDy with NMR data
PyJ	1FAF	6.35	2.02	8.97	3.74	1.81
Ubiquitin	1UBQ	3.57	1.92	2.33	0.64	0.49
GB3	1P7E	3.47	0.84	2.82	0.88	0.89
Q5E7H1	2JVW	6.40	1.11	14.19	2.3	1.11
RPA3401	2JTV	3.20	1.30	2.87	1.13	0.74
RHOS4 26430	2JVM	4.60	1.51	4.38	2.15	1.0
Protein LX	2JXT	3.33	1.59	6.24	3.01	0.75
PefI	2JT1	6.75	1.67	14.89	1.06	1.71
tRNA hydrolase domain	2JVA	5.27	1.88	5.23	3.35	1.85
CSPA	1MJC	7.07	2.08	8.04	2.3	1.56
Calbindin D9 K	3ICB	4.81	2.15	4.68	3.26	1.17
NE1242	2JV8	3.63	2.48	6.37	2.38	1.18
Average		4.9	1.7	6.8	2.2	1.2

CS-MD performance data was taken from the work of Robustelli et al. (2010)

Table 2 Model accuracy of comparative models of ubiquitin under different refinement scenarios

Template PDB ID	ID %	Model accuracy (backbone RMSD to the reference model, Å)			
		Initial model	Refined by XPLOR with NMR data	Refined by CS-GAMDy without NMR data	Refined by CS-GAMDy with NMR data
1OTR	96	0.91	1.14	1.37	0.66
2GBK	92	3.99	7.03	3.23	0.87
1UD7	91	0.83	0.92	0.86	0.65
2GBJ	90	4.18	3.24	3.33	0.83
2GBM	90	1.0	1.15	0.86	0.83
1WY8	39	3.85	19.27	3.21	0.65
2DZI	39	0.88	4.88	0.94	0.53
2FAZ	37	4.51	25.60	4.44	0.63
1OQY	36	2.42	3.72	1.81	0.63
1WH3	36	0.84	15.84	1.11	0.64
1WX9	34	1.06	1.08	0.82	0.83
1Z2M	33	0.92	0.95	0.87	0.64
1MG8	32	1.41	1.48	0.94	0.64
1UEL	32	0.76	11.83	0.91	0.66
1IYF	30	1.91	10.59	1.38	0.68
1WE7	28	3.98	6.84	3.46	0.5
1TTN	26	1.17	15.23	1.73	0.88
Average		2.03	7.69	1.8	0.69

Table 3 Accuracy of comparative models with different sizes and types of protein architecture under different refinement scenarios

Protein name	PDB ID	Model accuracy (backbone RMSD to the reference model, Å)			
		Initial model	Refined by XPLOR with NMR data	Refined by GAMDy without NMR data	Refined by GAMDy with NMR data
PyJ	1FAF	1.89	2.11	2.6	1.65
Elongation Factor 1	1B64	1.13	1.47	1.89	1.08
GB3	1P7E	2.17	2.45	1.99	1.65
Foxo4	1E17	3.12	25.46	3.01	2.07
Hamster PrP	1B10	1.63	3.99	2.26	1.52
Vts1	2D3D	1.15	1.21	1.66	0.96
NifU-like protein	2LTL	4.62	7.46	4.16	1.49
cg2496	2KPT	3.58	6.39	3.25	2.47
Cadherin	1SUH	1.93	1.51	1.85	1.68
Adenylate kinase	2CDN	1.37	40.41	1.71	1.24
NFU1 homolog	2M5O	3.82	9.67	4.18	1.43
Average		2.4	9.2	2.6	1.5

RMSD coordinate errors ranging from 1 to 16 Å. As shown in Fig. 2, CS-GAMDy could consistently refine ubiquitin models with starting RMSD's ranging from 1 to 10 Å to a near-correct structure (i.e. RMSD < 1 Å). In contrast, XPLOR's refinement could only refine near-native misfolded models (i.e. RMSD < 4 Å).

For the second test, we ran CS-GAMDy and the XPLOR refinement on a protein evaluation set used by the aforementioned CS-MD (Robustelli et al. 2010), a program for protein 3D refinement with chemical shifts. This was done to compare the performance of the three programs on similar data sets of moderately damaged structures

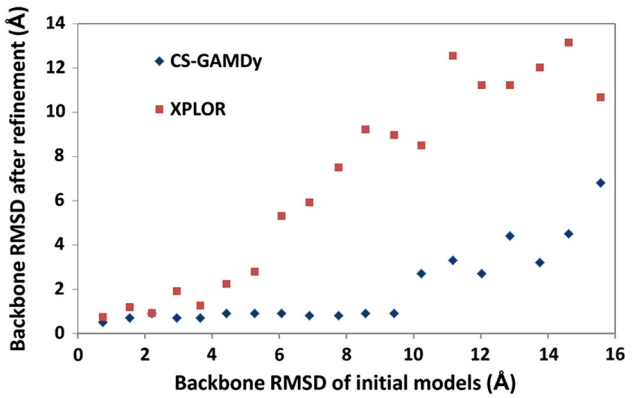

Fig. 2 Comparison of the performance of CS-GAMDy (*blue diamonds*) and XPLOR (*red squares*) for distorted models of ubiquitin. Model accuracy (backbone RMSD of non-coil regions with respect to the PDB entry 1UBQ) is plotted on the X axis (before refinement) and Y axis (after refinement), respectively

(RMSD < 7 Å). We distorted the 3D structure of each protein to the RMSD value and by the misfolding method described in the CS-MD publication (Robustelli et al. 2010). In order to assess the influence of experimental data on the refinement process, we tested CS-GAMDy with and without chemical shift scores and restraints. As seen in Table 1 and Tables S6-9, CS-GAMDy was able to consistently refine the starting models towards the correct native structure and improve the chemical shift based scores for all proteins. Table S7 illustrates the striking improvement in backbone chemical shift correlations achieved by CS-GAMDy, with the starting structures having average correlation coefficients of just 0.35, as calculated by ShiftX (Neal et al. 2003), and the final structures having correlation coefficients of 0.72 (matching that of the native proteins). Examples of the improvements in structural quality for several distorted protein models are shown on Figs. 3a–c. The average level of RMSD improvement, with and without experimental data, was 3.6 and 2.7 Å, respectively. This result indicates that CS-GAMDy has a capacity to efficiently refine protein structures even without chemical shift data. However, using chemical shifts improves the refinement outcome by an additional ~1 Å. The average improvement in backbone RMSD by the CS-MD method (using chemical shift data) was 3.2 Å (Robustelli et al. 2010). The standard XPLOR refinement actually made model accuracy worse by 1.9 Å. This test demonstrates that CS-GAMDy's performance for refining distorted protein models is better than the performance of CS-MD (Table 1) and XPLOR (Table 1; Tables S6–9). We also tested CS-GAMDy on misfolded models of these proteins with RMSDs ranging from 1 to 11 Å (Fig. S5). In all but one case (Protein LX), CS-GAMDy was able to refine the proteins to a near-native

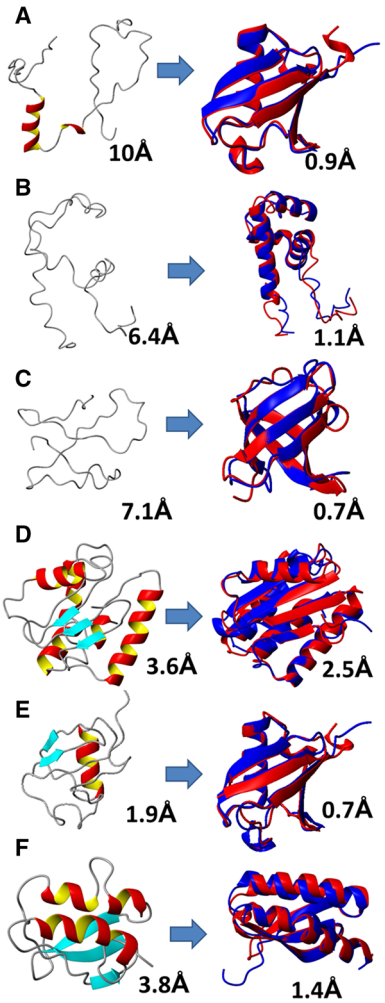

Fig. 3 Improvement in model accuracy after CS-GAMDy refinement. Starting models are shown on the left. β-strands are colored blue, α-helices are colored red and yellow, coil regions are *colored gray*. Alignments of refined models (*red*) with the reference models (*blue*) are shown on the right. *Numbers* represent the backbone RMSD (in Å) between the models and the reference structure. **a** Distorted model of ubiquitin, reference PDB ID: 1UBQ, **b** Distorted model of Q5E7H1, reference PDB ID: 2JVW, **c** Distorted model of CSPA, reference PDB ID: 1MJC, **d** Comparative model of cg2496, reference PDB ID: 2KPT, template PDB ID: 2KW7, sequence ID: 24 %, **e** Comparative model of ubiquitin, reference PDB ID: 1UBQ, template PDB ID: 1IYF, sequence ID: 30 %, **f** Comparative model of NFU1 homolog, reference PDB ID: 2M5O, template PDB ID: 1TH5, sequence ID: 20 %

structure with an RMSD below 2 Å, even when the accuracy of the initial model was as poor as 6 Å.

Near-native protein models (such as those generated by comparative modelling, CS23D, CS-Rosetta or NOE-based methods) are often more challenging to refine than uniformly distorted models. Their structural defects can be very localized and, therefore, not easily detected by global scoring functions. To simulate this scenario, we tested CS-GAMDy and the XPLOR refinement on 17 comparative

models of ubiquitin that were generated from templates ranging from 26 to 96 % sequence identity. These models had RMSDs ranging from 1.37 to 4.86 Å. All comparative models were prepared using the homology modelling functionality of the CS23D webserver (Wishart et al. 2008). PDB IDs of templates and reference proteins are listed in Tables 2 and S14. In all cases, CS-GAMDy was able to improve model accuracy and chemical shift based scores (Table 2 and Tables S10–13). The RMSD to the reference structure was below 1 Å for all refined models. An example of the level of structural improvement achieved is shown on Fig. 3e. When chemical shifts were used, the average improvement in model accuracy was 1.34 Å with the maximum improvement being 3.9 Å. Removal of experimental data led to much more modest enhancements of model accuracy (i.e. average improvement of 0.2 Å). A standard XPLOR refinement protocol made average model accuracy worse by 5.7 Å.

In the final test, we evaluated how well CS-GAMDy and the standard XPLOR refinement could optimize homology models of 11 different proteins with different architectures (all α, α/β, and all β), created from templates with sequence identity levels ranging from 19 to 95 % (Table S14). The RMSD of these models relative to the corresponding reference structures ranged from 1.1 to 4.6 Å (Table 3). In all cases, CS-GAMDy succeeded in improving chemical shift based scores and decreasing coordinate errors, with RMSD reductions ranging from 0.05 to 3.13 Å (Table 3 and Tables S15–18). The average changes in model accuracy (with and without experimental data) were 0.8 and −0.2 Å, respectively, indicating that the experimental data was essential for the refinement. In contrast, the XPLOR refinement actually decreased model accuracy by 7 Å, on average. Examples of improvements of comparative models by CS-GAMDy are shown on Fig. 2d–f.

The aforementioned results demonstrate that the CS-GAMDy protocol can tolerate modest errors in the input data. Indeed, the chemical shift input data did not have complete agreement with the reference structures (Tables S6–9 and S15–18). Yet, the CS-GAMDy refinement achieved good accuracy for many of proteins tested (Figs. 2, S5; Tables 1, 2, 3). Not surprisingly, CS-GAMDy showed the best performance for experimental input with the smallest errors, especially with errors in torsion angle restraints. This can be seen with the data sets for ubiquitin and GB3 (Figs. 2, S5).

Conclusion

Protein structure determination from sparse NMR data is critical for expanding NMR's reach to higher molecular weight proteins, disordered proteins, and poorly soluble proteins. Current approaches generally combine comparative modelling or fragment-based assembly with sparse NMR data such as chemical shifts or small number of NOEs. However, these limited-data methods often generate unrefined, approximate models with clear structural errors. The inability to easily refine and optimize these structures using sparse NMR data (i.e. chemical shifts) has limited their deposition frequency to the PDB and prevented their widespread uptake and use within the NMR community. To address these problems, we have developed a new algorithm, called CS-GAMDy, for performing chemical shift optimization with the widely used XPLOR-NIH molecular modelling package. Extensive assessments using four different test sets showed that CS-GAMDy was able to consistently drive all starting (approximate) structures towards the correct structure while at the same time improving the level of agreement with the observed chemical shifts. CS-GAMDy employs a unique combination of multi-objective MD biasing and a genetic algorithm to incorporate pre-existing and experimental NMR information, including the novel RCI-ASA score, into its protein model optimization. CS-GAMDy represents the first successful implementation of the CONTRA MD biasing method with collective variables derived from chemical shifts and knowledge-based scores.

Based on its performance over a wide range of refinement scenarios we believe CS-GAMDy will now allow protein models initially generated by sparse restraint or chemical-shift-only methods to achieve sufficiently high quality to be considered fully refined and worthy of PDB submission. Furthermore, CS-GAMDy should also allow the time and labour savings originally projected for sparse-restraint NMR structure determination to be fully realized. Efforts to improve the program's speed (through parallelization) and accuracy, through the use of ShiftX+ (Han et al. 2011) and improved ASA calculations, are actively underway. Extending CS-GAMDy to work with other types of sparse NMR data (NOEs, RDCs, PREs, cross-linking data) and to perform ab initio folding is also under development and will be described in future publications. CS-GAMDy is available from www.gamdy.ca.

Acknowledgments This work was supported by the Natural Sciences and Engineering Research Council (NSERC) and Compute Canada.

References

Berjanskii MV, Wishart DS (2013) A simple method to measure protein side-chain mobility using NMR chemical shifts. J Am Chem Soc 135:14536–14539. doi:10.1021/ja407509z

Berjanskii MV, Neal S, Wishart DS (2006) PREDITOR: a web server for predicting protein torsion angle restraints. Nucleic Acids Res 34:W63–W69. doi:10.1093/nar/gkl341

Berjanskii M et al (2009) GeNMR: a web server for rapid NMR-based protein structure determination. Nucleic Acids Res 37:W670–W677. doi:10.1093/nar/gkp280

Boomsma W, Tian P, Frellsen J, Ferkinghoff-Borg J, Hamelryck T, Lindorff-Larsen K, Vendruscolo M (2014) Equilibrium simulations of proteins using molecular fragment replacement and NMR chemical shifts. Proc Natl Acad Sci USA 111:13852–13857. doi:10.1073/pnas.1404948111

Brunger AT et al (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905–921

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 104:9615–9620. doi:10.1073/pnas.0610313104

Cheung MS, Maguire ML, Stevens TJ, Broadhurst RW (2010) DANGLE: a Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. J Magn Reson 202:223–233. doi:10.1016/j.jmr.2009.11.008

Cornell WD et al (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197. doi:10.1021/ja00124a002

Grishaev A, Bax A (2004) An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. J Am Chem Soc 126:7281–7292. doi:10.1021/ja0319994

Guntert P (2004) Automated NMR structure calculation with CYANA. Methods Mol Biol 278:353–378. doi:10.1385/1-59259-809-9:353

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57. doi:10.1007/s10858-011-9478-4

Harvey SC, Gabb HA (1993) Conformational transitions using molecular dynamics with minimum biasing. Biopolymers 33:1167–1172. doi:10.1002/bip.360330803

Jorgensen WL, Tirado-Rives J (1988) The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 110:1657–1666. doi:10.1021/ja00214a001

Kuszewski J, Gronenborn AM, Clore GM (1996) Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci 5:1067–1080. doi:10.1002/pro.5560050609

Kuszewski J, Gronenborn AM, Clore GM (1997) Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. J Magn Reson 125:171–177. doi:10.1006/jmre.1997.1116

Kuszewski J, Gronenborn AM, Clore GM (1999) Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. J Am Chem Soc 121:2337–2338. doi:10.1021/ja9843730

MacKerell AD et al (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616. doi:10.1021/jp973084f

Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. J Biomol NMR 26:215–240. doi:10.1023/A:1023812930288

Powell MJD (1977) Restart procedures for the conjugate gradient method. Math Program 12:241–254. doi:10.1007/bf01593790

Raman S et al (2010) NMR structure determination for larger proteins using backbone-only data. Science 327:1014–1018. doi:10.1126/science.1183649

Robustelli P, Kohlhoff K, Cavalli A, Vendruscolo M (2010) Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. Structure 18:923–933. doi:10.1016/j.str.2010.04.016

Rosato A et al (2015) The second round of critical assessment of automated structure determination of proteins by NMR: CASD-NMR-2013. J Biomol NMR. doi:10.1007/s10858-015-9953-4

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65–73. doi:10.1016/S1090-7807(02)00014-9

Shen Y, Bax A (2015) Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. Methods Mol Biol 1260:17–32. doi:10.1007/978-1-4939-2239-0_2

Shen Y et al (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690. doi:10.1073/pnas.0800256105

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223. doi:10.1007/s10858-009-9333-z

Stein EG, Rice LM, Brunger AT (1997) Torsion angle molecular dynamics as a new efficient tool for NMR structure calculation. J Magn Reson 124:154–164. doi:10.1006/jmre.1996.1027

Thompson JM et al (2012) Accurate protein structure modeling using sparse NMR data and homologous structure information. Proc Natl Acad Sci USA 109:9875–9880. doi:10.1073/pnas.1202485109

Wagner F, Simonson T (1999) Implicit solvent models: combining an analytical formulation of continuum electrostatics with simple models of the hydrophobic effect. J Comput Chem 20:322–335. doi:10.1002/(sici)1096-987x(199902)20:3<322:aid-jcc4>3.0.co;2-q

Wishart DS, Sykes BD (1994) The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. J Biomol NMR 4:171–180

Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31:1647–1651

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36:W496–W502. doi:10.1093/nar/gkn305

Zhang J, Zhang Y (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS One 5:e15386. doi:10.1371/journal.pone.0015386

Zhou H, Skolnick J (2011) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 101:2043–2052. doi:10.1016/j.bpj.2011.09.012