FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

PhD Thesis

Anders S. Christensen

Inferential Protein Structure Determination Using Chemical Shifts

Academic supervisor: Jan H. Jensen

January 29, 2014

Acknowledgements

I would like to thank the following people

- Jan Jensen
- Casper Steinmann
- More people

Publication list

List of publications:

1. <u>Anders S. Christensen</u>, Stephan P. A. Sauer, Jan H. Jensen (2011) Definitive benchmark study of ring current effects on amide proton chemical shifts. *Journal of Chemical Theory and Computation*, 7:2078-2084.

2. Wouter Boomsma, Jes Frellsen, Tim Harder, Sandro Bottaro, Kristoffer E. Johansson, Pengfei Tian, Kasper Stovgaard, Christian Andreetta, Simon Olsson, Jan B. Valentin, Lubomir D. Antonov, <u>Anders S. Christensen</u>, Mikael Borg, Jan H. Jensen, Kresten Lindorff-Larsen, Jesper Ferkinghoff-Borg, Thomas Hamelryck (2013) PHAISTOS: A framework for Markov chain Monte Carlo simulation and inference of protein structure. *Journal of Computational Chemistry*, 34:1697-1705.

3. <u>Anders S. Christensen</u>, Troels E. Linnet, Mikael Borg, Wouter Boomsma, Kresten Lindorff-Larsen, Thomas Hamelryck, Jan H. Jensen (2013) Protein Structure Validation and Refinement Using Amide Proton Chemical Shifts Derived from Quantum Mechanics. *PLoS ONE* 8:e84123.

4. <u>Anders S. Christensen</u>, Thomas Hamelryck, Jan H. Jensen (2014) FragBuilder: An efficient Python library to setup quantum chemistry calculations on peptides models. *PeerJ* (accepted).

5. <u>Anders S. Christensen</u>, Lars Bratholm, Simon Olsson, Thomas Hamelryck, Jan H. Jensen (2014) Weighting of chemical shift evidence in Monte Carlo simulation of proteins. *ShareLatex* (unpublished).

6. Torus-DBN-CS-LARS

7. J-coupling Casper/Kongen

List of public code:

1. FragBuilder (BSD license) `https://github.com/jensengroup/fragbuilder`

2. CamShift module (BSD license) `https://github.com/jensengroup/camshift-phaistos`

3. ProCS module (BSD license) `https://github.com/jensengroup/procs-phaistos`

4. Phaistos (GPL license) `https://svn.code.sf.net/p/phaistos/code/trunk`

5. GAMESS patch FMO-RHF:MP2 (GAMESS license/free) `https://github.com/andersx/fmo-rhf-mp2`

List of other publications:

1. Casper Steinmann, Kristoffer L. Blædel, <u>Anders S. Christensen</u>, Jan H. Jensen (2013) Interface of the polarizable continuum model of solvation with semi-empirical methods in the GAMESS program. *PLoS ONE* 8:e67725.

2. <u>Anders S. Christensen</u>, Casper Steinmann, Dmitri G. Fedorov, Jan H. Jensen (2013) Hybrid RHF/MP2 geometry optimizations with the Effective Fragment Molecular Orbital Method. *PLoS ONE* (accepted).

3. HF-3c Jimmy

4. PM6 Jimmy

5. h-bond Jimmy

Preface

This is the introduction

Contents

Chapter 1

Chemical shifts in a probabilistic framework

This section introduces the formalism for Monte Carlo simulations which includes both physical energy terms as well as a probabilistic energy terms based on experimentally observed chemical shifts. These equations presented are not new, but have not been published in the form in which they are presented here. The intention is to present the equations in the form in which they are implemented in PHAISTOS, so that they can easily be re-implemented in other programs by others.

1.1 Defining an energy function from Bayes' theorem

A simplistic approach to this problem is to is to define a hybrid energy by defining a penalty function that describes the agreement between experimental data and data calculated from a proposed protein structure with a physical energy (such as from a molecular mechanics force field). A structure is then determined by minimizing

$$E_{\text{hybrid}} = w_{\text{data}} \ E_{\text{data}} + E_{\text{physical}}. \tag{1.1}$$

This approach, however, does not uniquely define neither shape nor weight of E_{data}. Chemical shifts have been combined with physical energies in a multitude of ways, e.g., weighted RMSD values or harmonic constraints. Vendruscolo and co-workers implemented a "square-well soft harmonic potential", and corresponding molecular gradients and were able to run a chemical shift-biased MD simulation. In all cases the parameters and weights of E_{data} had to be carefully tweaked by hand, and it is not clear how to choose optimal parameters.

The inferential structure determination (ISD) principles introduced by Rie-ping, Habeck and Nigles [Rieping et al., 2005]defines a Bayesian formulation of Eq XX. In the following section the equations of an ISD approach are derived for combining the knowledge of experimental chemical shifts with a physical energy. First remember Bayes' theorem which relates a conditional probability (A given B) with its inverse:

$$p(A|B) = \frac{p(B|A) \, p(A)}{p(B)} \tag{1.2}$$

Now consider a set of chemical shifts $\{\delta_i\}$, and the uncertainty to which these can be predicted $\{\sigma_i\}$ from a structure, $\boldsymbol{X}$ (the experimental uncertainty is negligibly small compared to this). We have to make the basic assumption, that the error, given as $\Delta\delta_i = \left|\delta_i^{\text{predicted}} - \delta_i^{\text{experimental}}\right|$, approximately follows a Gaussian distribution with some standard deviation, but we need not

hand-pick and assign any numeric value to the standard deviation.Furthermore, the Gaussian distribution is the least biasing distribution according to the principle of maximum entropy.

In this case, the most likely structure, $\boldsymbol{X}$, and optimal choice of $\{\sigma_i\}$ is found by maximizing (via Bayes' theorem)

$$p\left(\boldsymbol{X}, \{\sigma_i\}\Big|\{\delta_i\}\right) = \frac{p\left(\{\delta_i\}\Big|\boldsymbol{X}, \{\sigma_i\}\right)p\left(\boldsymbol{X}, \{\sigma_i\}\right)}{p\left(\{\delta_i\}\right)}. \tag{1.3}$$

Here, *marginal distribution* of $p\left(\{\delta_i\}\right)$ merely serves as a normalizing factor and the *likelihood* of $p\left(\{\delta_i\}|\boldsymbol{X}, \{\sigma_i\}\right)$, is obtained as the product of the individual, Gaussian probabilities over all n single chemical shift measurements. Nuclei of the same atom-type, here denoted by index j, (e.g. C^α, H^α, etc.) are assumed to carry the same uncertainly denoted by σ_j:

$$\begin{aligned}
p\left(\{\delta_i\}\Big|\boldsymbol{X}, \{\sigma_i\}\right) &\simeq \prod_{i=0}^{n} p\left(\Delta\delta_i|\boldsymbol{X}, \sigma_i\right) \tag{1.4}\\
&= \prod_{j=0}^{m}\prod_{i_j=0}^{n_j} p\left(\Delta\delta_{i_j}|\boldsymbol{X}, \sigma_j\right) \tag{1.5}\\
&= \prod_{j=0}^{m}\prod_{i_j=0}^{n_j} \frac{1}{\sigma_j\sqrt{2\pi}}\exp\left(-\frac{\Delta\delta_{i_j}^2}{2\sigma_j^2}\right) \tag{1.6}\\
&= \prod_{j=0}^{m}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right)^{n_j}\exp\left(\sum_{i_j=0}^{n_j}-\frac{\Delta\delta_{i_j}^2}{2\sigma_j^2}\right) \tag{1.7}\\
&= \prod_{j=0}^{m}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right)^{n_j}\exp\left(\frac{-\chi_j^2(\boldsymbol{X})}{2\sigma_j^2}\right) \tag{1.8}
\end{aligned}$$

Furthermore, $p\left(\boldsymbol{X}, \{\sigma_j\}\right)$ can be simplified as

$$\begin{aligned}
p\left(\boldsymbol{X}, \{\sigma_j\}\right) &\propto p\left(\{\sigma_j\}\Big|\mathbf{X}\right)p\left(\mathbf{X}\right) \tag{1.9}\\
&= p\left(\{\sigma_j\}\right)p\left(\boldsymbol{X}\right), \tag{1.10}
\end{aligned}$$

where it is assumed that the errors in the chemical shift prediction model are independent of the particular protein structure and *vice versa*. The *prior* distribution of $p\left(\{\sigma_j\}\right)$ is accounted for by proposing updates from a log-normal distribution (see next subsection). $p(\boldsymbol{X})$ of the molecular protein structure is here simply the Boltzmann distribution, i.e.

$$p(\boldsymbol{X}) = \frac{1}{Z}\exp\left(-\frac{E(\boldsymbol{X})}{k_\text{B}T}\right) \tag{1.11}$$

where $E(\boldsymbol{X})$ is the (physical) potential energy of the protein structure, most often described by a molecular mechanics force field. k_B is the Boltzmann constant and T is the temperature of interest. Luckily we need not calculate the partition function, Z, because the relative energy landscape is invariant under choice of normalization constant. Note that $p(\boldsymbol{X})$ also can be introduced via conformational sampling from a biased distribution, such as for example TorusDBN or BASILISK (mimicking the Ramachandran plot and side chain rotamer distributions, respectively).

Neglecting normalization constants, the total probability to be maximized is thus proportional to:

$$p\Big(\boldsymbol{X}, \{\sigma_i\}\Big|\{\delta_i\}\Big) \propto p\Big(\{\delta_i\}\Big|\boldsymbol{X}, \{\sigma_i\}\Big)p\Big(\boldsymbol{X}\Big)p\Big(\{\sigma_i\}\Big) \tag{1.12}$$

$$\propto \prod_{j=0}^{m}\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right)^{n_j}\exp\left(-\frac{1}{2\sigma_j^2}\chi_j^2\right)\exp\left(-\frac{E(\boldsymbol{X})}{k_{\mathrm{B}}T}\right)p\Big(\{\sigma_j\}\Big) \tag{1.13}$$

When $p(\{\sigma_j\})$ is introduced via biased sampling, the associated hybrid-energy to be evaluated is (again neglecting constant terms)

$$E_{\mathrm{hybrid}}\left(\boldsymbol{X}\right) = -k_{\mathrm{B}}T\ln\Big(p\Big(\boldsymbol{X}, \{\sigma_i\}\Big|\{\delta_i\}\Big)\Big) \tag{1.14}$$

$$= E(\boldsymbol{X}) - k_{\mathrm{B}}T\sum_{j=0}^{n_j}n_j\ln\left(\frac{1}{\sigma_j\sqrt{2\pi}}\right) + \frac{\chi_j^2}{2\sigma_j^2} \tag{1.15}$$

1.1.1 Jeffreys' prior (general, one-parameter case)

The prior distribution of the nuisance parameter is inherently unknown. In such cases, it is necessary to use a prior distribution that will have only very little influence on the sampled value. One such *uninformative prior* could for instance be a flat distribution over the positive real line. The concept of Jeffreys' priors are a generalization of flat priors. In the one parameter case the Jeffrey's prior is given as

$$p(\theta) \propto \sqrt{\mathbf{I}(\theta)}, \tag{1.16}$$

where $\mathbf{I}(\theta)$ is the *Fisher information* defined (in the one parameter case) as

$$\mathbf{I}(\theta) = \left\langle\left(\frac{\partial}{\partial\theta}\ln p(x|\theta)\right)^2\right\rangle. \tag{1.17}$$

1.1.2 Jeffreys' prior (Gaussian and Cauchy distributions)

Here we derive Jefferys' prior for the uncertainty of a Gaussian distribution, i.e. a distribution on the form

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)}{2\sigma^2}\right). \tag{1.18}$$

This immediately gives us the Jeffreys' prior:

$$
\begin{aligned}
p(\sigma) &\propto \sqrt{\left\langle\left(\frac{\partial}{\partial\sigma}\ln p(x|\mu, \sigma)\right)^2\right\rangle} \\
&= \sqrt{\left\langle\left(\frac{\partial}{\partial\sigma}\ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)}{2\sigma^2}\right)\right]\right)^2\right\rangle} \\
&= \sqrt{\left\langle\left(\frac{(x-\mu)-\sigma^2}{\sigma^3}\right)^2\right\rangle} \\
&= \sqrt{\int_{-\infty}^{\infty}p(x|\mu, \sigma)\left(\frac{(x-\mu)-\sigma^2}{\sigma^3}\right)^2 dx} \\
&= \sqrt{\frac{2}{\sigma^2}} \propto \frac{1}{\sigma}
\end{aligned}
\tag{1.19}
$$

Similarly for the γ parameter of the Cauchy distribution of the form

$$p(x|x_0, \gamma) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}, \tag{1.20}$$

we obtain the following Jeffreys' prior:

$$
\begin{aligned}
p(\gamma) \quad &\propto \quad \sqrt{\left\langle \left(\frac{\partial}{\partial\gamma} \ln p(x|x_0,\gamma)\right)^2 \right\rangle} \\
&= \quad \sqrt{\left\langle \left(\frac{\partial}{\partial\gamma} \ln \left[\frac{1}{\pi\gamma \left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}\right]\right)^2 \right\rangle} \\
&= \quad \sqrt{\left\langle \left(-\frac{\gamma^2 - (x-x_0)^2}{\gamma^3 + \gamma(x-x_0)^2}\right)^2 \right\rangle} \\
&= \quad \sqrt{\int_{-\infty}^{\infty} p(x|x_0,\gamma) \left(-\frac{\gamma^2 - (x-x_0)^2}{\gamma^3 + \gamma(x-x_0)^2}\right)^2 dx} \\
&= \quad \sqrt{\frac{1}{2\gamma^2}} \quad \propto \quad \frac{1}{\gamma}
\end{aligned}
\tag{1.21}
$$

1.1.3 Sampling of nuisance parameters

Since the nuisance parameters of the energy functions are unknown, they too must be sampled. The move used to update the value of the nuisance parameters must obey detailed balance:

$$p(\theta \rightarrow \theta') = p(\theta' \rightarrow \theta) \tag{1.22}$$

The simplest Monte Carlo move is simply adding a number from a normal distribution with $\mu = 0$, this clearly obeys detailed balance, since the distribution is symmetric. For the scale parameter, γ and σ, of the Cauchy and Gaussian distributions, respectively, we found a variance of 0.05 in the normal distributed move to converge quickly and stably. Figure 1.1 show a histogram of sampled values of γ and σ for the NMR structure of Protein G (PDB-id: 2OED). 55 C-alpha

1.2 Prior distributions for protein structure

1.2.1 Molecular mechanics force field

One reasonable prior distribution for protein structure, $p(\mathbf{X})$, is the Boltzmann distributino, e.g.:

$$p(\mathbf{X}) \propto \exp\left(\frac{-E}{k_{\mathrm{B}}T}\right) \tag{1.23}$$

where E is the energy of the structure, $\mathbf{X}$ and k_{B} and T are Boltzmann's constant and the temperature, respectively. The energy of the structure is in this context usually approximated by a molecular mechanics force field that is taylor-made for protein simulations. PHAISTOS currently supports two different protein force field: The OPLS-AA/L force field with a GB/SA solvent term, and the PROFASI force field.
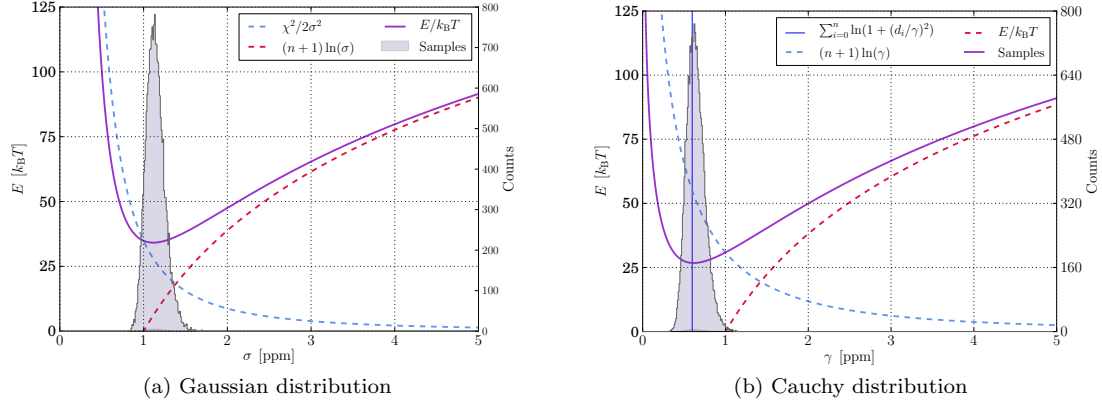
(a) Gaussian distribution

(b) Cauchy distribution

Figure 1.1: Sampling of σ and γ for 2OED for Ca-chemical shifts. $n = 55$ and $\chi^2 = 69.7$.

1.2.2 Generative probabilistic models

Another way to introduce the prior distribution for a protein structure is to bias the conformational sampling. Conventional conformational sampling will proposed (ϕ, ψ) backbone angles uniformly (i.e. in the range $[-180°, 180°]$ and let the energy function filter and construct the target distribution (e.g. the canonical ensemble, etc.) via energy evaluation. Since only a fraction of the possible (ϕ, ψ) backbone angles are allowed (i.e. the Ramachandran plot), it is computationally very convenient to only sample from the allowed regions.

TorusDBN

BASILISK

1.2.3 Or both

Definitive Benchmark Study of Ring Current Effects on Amide Proton Chemical Shifts

Anders S. Christensen,* Stephan P. A. Sauer, and Jan H. Jensen*

Department of Chemistry, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark

S *Supporting Information*

ABSTRACT: The ring current effect on chemical shifts of amide protons ($\Delta\delta_{RC}$) is computed at the B3LYP/6-311++G(d,p)// B3LYP/aug-cc-pVTZ level of theory for 932 geometries of dimers of *N*-methylacetamide and aromatic amino acid side chains extracted from 21 different proteins. These $\Delta\delta_{RC}$ values are scaled by 1.074, based on MP2/cc-pVQZ//B3LYP/aug-cc-pVTZ chemical shift calculations on four representative formamide/benzene dimers, and are judged to be accurate to within 0.1 ppm based on CCSD(T)/CBS//B3LYP/aug-cc-pVTZ calculations on formamide. The 932 scaled $\Delta\delta_{RC}$ values are used to benchmark three empirical ring current models, including the Haigh—Mallion model used in the SPARTA, SHIFTX, and SHIFTS chemical shift prediction codes. Though the RMSDs for these three models are below 0.1 ppm, deviations up to 0.29 ppm are found, but these can be decreased to below 0.1 ppm by changing a single parameter. The simple point-dipole model is found to perform just as well as the more complicated Haigh—Mallion and Johnson—Bovey models.

1. INTRODUCTION

Prediction of chemical shifts in proteins based on protein structure serves many uses in structure verification or fast generation of structures in accordance with relatively inexpensive experimental nuclear magnetic resonance (NMR) data.[1-5] The most popular protein chemical prediction software packages include the SHIFTX,[6] SHIFTS,[7,8] SPARTA,[9] and PROSHIFT[10] servers. These programs use empirical models that relate various features of protein structure, such as secondary structure, hydrogen bond geometry, and ring current effects, to changes in chemical shifts. The contributions from several different sources of chemical shift perturbations are in all cases assumed to work additively. These small additive terms are in many cases given as classical approximations to well-known physical interactions. For example, SPARTA, SHIFTX, and SHIFTS use the approximation due to Haigh and Mallion[11] to model the changes in chemical shifts due to ring current effects in the aromatic side chains of phenylalanine, tyrosine, tryptophan, and histidine residues. The Haigh—Mallion model contains a single adjustable parameter for each side chain, which must be parametrized from a data set. In the case of SHIFTX, these parameters are obtained by a data mining approach, which correlates experimental chemical shifts with corresponding, known X-ray protein structures, based on a series of predefined physical and empirical terms.[6] In the case of SHIFTS, the parameters are obtained by fitting parameters for a set of known physical terms, which relates protein structure and chemical shifts to a data set which combines empirical chemical shift data as well as data obtained through quantum chemical calculations.[12] Ultimately, these fitting methods include uncertainties from many terms in the underlying approximations of the fit, as well as the uncertainties connected to the experimental data. It is thus unclear how accurate the obtained parameters are in reproducing the underlying physics of the system.[13]

Other methods exist to approximate the ring current effect, most notably[11,14] the Johnson—Bovey model[15] and the simpler point-dipole model due to Pople.[16,17] In general, the three methods describe the change in chemical shift due to a nearby aromatic ring, formally as

$$\Delta\delta_{RC} = i\,B\,G \tag{1}$$
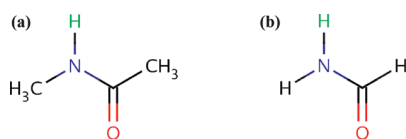
where G is a geometric factor, which depends on the spatial orientation and distance of the ring relative to the proton, B is a "natural constant" denoting the ring current intensity for a benzene ring, and i is the ring current intensity relative to that of a benzene ring, such that $i_{benzene} \equiv 1$. A thorough description of the three models mentioned above can be found in the Supporting Information. It has previously been attempted to derive analytical expressions for the values of i and B for different functional forms of G, but these have not been successful in reproducing experimental results.[11] Consequently, various numerical methods have been widely used to obtain the intensity values.

The study by Case[18] is one of the few that has addressed these issues using *ab initio* calculations. A methane molecule is used to probe the chemical shift perturbation due to a nearby aromatic ring, with the chemical shift calculated at the CSGT/PW91/ IGLO-III level of theory. However, it is not clear whether the level of theory used at that time (1995) is sufficiently accurate, and second, it is unknown whether the ring current parameters obtained for a methane proton can directly be transferred to amide protons. The parameters obtained by Case are used in the SPARTA program.[9]

In this study, we use CCSD(T) and MP2 methods to benchmark the accuracy of $\Delta\delta_{RC}$ calculations at the B3LYP/ 6-311++G(d,p)//B3LYP/aug-cc-pVTZ level of theory. This level of theory is then used to compute nearly 1000 representative

ACS Publications © 2011 American Chemical Society

2078

dx.doi.org/10.1021/ct2002607 | *J. Chem. Theory Comput.* 2011, 7, 2078–2084

Figure 1. The molecules used as probes for the ring current effects on amide protons: *trans-N*-methylacetamid (NMA) (a) and formamide (FMA) (b). The probe nucleus for which the shielding constants are calculated are the amide proton of NMA and the amide proton *trans* to the C=O bond in FMA, here marked in green color.

$\Delta\delta_{RC}$ values, which, in turn, are used to obtain parameters for three empirical $\Delta\delta_{RC}$ models.

The paper is organized as follows. First, we describe the computational methodology used to isolate the ring current effect and to obtain data sets, against which the ring current intensity parameters are fitted. Then, we benchmark various levels of theory, in order to estimate error bounds on our data. This is followed by a presentation of the obtained intensities and a comparison to the intensities obtained by other authors.
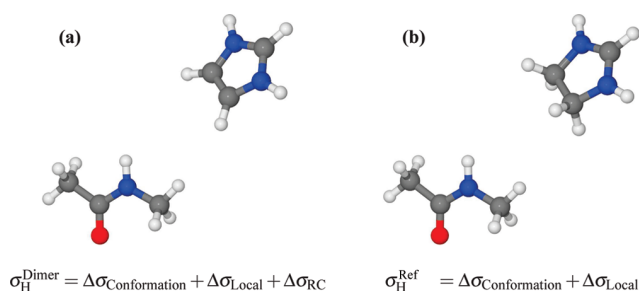
2. COMPUTATIONAL METHODOLOGY

2.1. Isolating the Ring Current Effect. As a probe nucleus, for which the isotropic shielding is calculated using quantum chemical methods, the amide proton of a *trans-N*-methylacetylamide (NMA; Figure 1a) molecule is used in order to provide a small and inflexible molecule with a high degree of resemblance to the amide group of the protein backbone. For more computationally demanding calculations, such as MP2, CCSD, and CCSD(T) calculations, the two methyl groups are removed in order to save computational time, and the probe nucleus is then the amide proton *trans* to the C=O bond in formamide (FMA; Figure 1b).

A large number of dimer systems (see next subsection), consisting of an amide probe molecule and a nearby aromatic ring in different conformations is constructed. Following the general approach of Boyd and Skrynnikov,[19] the absolute shielding of the probe hydrogen in the dimer system can be written as

$$\sigma_H^{Dimer} = \Delta\sigma_{Conformation} + \Delta\sigma_{Local} + \Delta\sigma_{RC} \qquad (2)$$

The chemical shift of the probe hydrogen atom in the dimer system will, apart from the ring current effect, also be influenced by the exact geometry and type of the probe molecule (described in the $\Delta\sigma_{Conformation}$ term) as well as any possible interactions with the aromatic moiety, such as electrostatic forces, possible hydrogen bonding, spin–spin repulsion, and other effects which can be difficult to quantify and separate (described in the $\Delta\sigma_{Local}$ term). Finally, the chemical shift perturbation due to the aromatic ring is approximated as $\Delta\sigma_{RC}$.

Reference systems for each dimer system, which have approximately identical local interactions between the molecules, apart from the ring current effect, are constructed in order to filter out these hard-to-quantify effects (see Figure 2). These are modeled as corresponding dimer systems, where the aromatic ring has been replaced by an olefinic analogue. The definition of an olefinic analogue here is an aromatic ring which, by the addition of two hydrogen atoms, has lost its aromaticity. The protonation is done such that the planar geometry of the ring is still enforced, causing deviations in the spatial positions relative to the corresponding aromatic ring dimer to be negligible. The olefinic

$$\sigma_H^{Dimer} = \Delta\sigma_{Conformation} + \Delta\sigma_{Local} + \Delta\sigma_{RC} \qquad \sigma_H^{Ref} = \Delta\sigma_{Conformation} + \Delta\sigma_{Local}$$

Figure 2. Two example geometries demonstrating the two different dimers used in the calculation scheme to isolate the ring current effect for one amide-ring conformation. The shown geometries correspond to the amide proton of ILE64 and the side chain of HIS69 in the HIV-1 protease, PDB-code 2I4V. In part a, the chemical shift of the probe nucleus is determined by the conformation of the NMA molecule, electrostatic and spin–spin repulsion interactions to the positively charged imidazolium molecule, and, by comparison, a small ring current interaction. In part b, the aromaticity of the imidazolium is broken, but the spatial distribution of charge as well as the internal conformation of NMA is approximately identical to those found in part a.

analogue is placed such that the ring center corresponds to the center of the aromatic ring, relative to the probe hydrogen, and the coordinates of the carbon/nitrogen atoms are matched as closely as possible. This approach ensures that $\Delta\sigma_{Conformation}$ and $\Delta\sigma_{Local}$ are largely retained, while $\Delta\sigma_{RC}$ is removed. Using this substitution scheme, the absolute shielding of the hydrogen atom in the reference system can be written as

$$\sigma_H^{Ref} = \Delta\sigma_{Conformation} + \Delta\sigma_{Local} \qquad (3)$$

which enables us to estimate the ring current contribution to the chemical shift due to the aromatic ring as

$$\Delta\delta_{RC} = -\Delta\sigma_{RC} \approx \sigma_H^{Ref} - \sigma_H^{Dimer} \qquad (4)$$

The aromatic rings studied here are equivalent to the rings found in the aromatic protein side chains. See Table 1 for an overview of the used molecules. Sketches are shown in the Supporting Information as well.

2.2. Construction of Test Systems. Dimers consisting of an amide probe and an aromatic ring were generated from a data set of 21 protein structures obtained from the RCSB Protein Data Bank (PDB),[20] in order to to ensure that only realistic conformations were used in the QM calculations.

The structures used were (listed by PDB code): 1F94, 1GK1, 1IGD, 1JYQ, 1JYR, 1JYU, 1Q3E, 1QJP, 1VJC, 1XA5, 1ZJK, 2ACO, 2B6C, 2D57, 2DRJ, 2ETL, 2F47, 2FZG, 2GOL, 2I4D, and 2I4V. Since the used structures were experimental X-ray structures, no hydrogen atoms were present in the structures, and PDB2PQR 1.5[21,22] was used to protonate the structures in order to obtain hydrogen atom positions. From all of these protein structures, we selected systems where the center of an aromatic ring was within a cutoff distance of 7 Å from an amide proton. This resulted in a total of 932 different dimer conformations (see Table 1). For each of these conformations, a dimer was created with a simpler aromatic ring in place of the aromatic side chain with the ring centers at identical coordinates and in the same plane. A directional vector was used to align the rotation of the ring in the plane, in order to have closely matching coordinates for the heavy atoms. For tyrosine, the center to oxygen vector was used. For benzene, the center to C^{γ} to C^{ζ} vector was used.

Table 1. List of the Side Chain Approximations Used in This Work and Their Olefinic Analogues[a]

side chain	analogue	olefinc analogue	# dimers
phenylalanine	benzene	1,4-cyclohexadiene	276
tyrosine	phenol	cyclohexa-1,4-diene-1-ol	172
tryptophan	indole	2,3,5,6-tetrahydroindole	113
histidine	imidazole	4,5-dihydroimidazole	185
histidine+	imidazolium	4,5-dihydroimidazolium	174

[a] The residue type is listed along with the aromatic and olefinic analogues, as well as the total number of different NMA/ring dimer conformations in each data set.

For histidine (both in the charged and neutral state), the center to $C^{\varepsilon 1}$ vector was used. The same histidine–amide group pairs were used to generate the dimers for both charged and neutral histidine conformations. The $N^{\varepsilon 2}$ nitrogen atom was in all cases of neutral histidine assumed to be the deprotonated nitrogen.

A given backbone amide group within the 7 Å range from the aromatic ring was substituted by an NMA or an FMA molecule, with the nitrogen atoms at identical coordinates. Furthermore, the N–H vector and the C(=O)–N–H plane were also aligned. See Table 1 for the number of dimers for each ring type.

If the aromatic ring corresponded to the ring of the side chain of the previous residue, with respect to the investigated amide group, the dimer construction scheme occasionally caused a clash between the extra hydrogens, where the C^{β} atom was previously located. Other conformations also gave rise to unphysical conformations, due to clashes between the inflexible subunits of the constructed dimers. To avoid computational artifacts from these, all dimers with a shortest intermolecular distance of 3.4 Å or less were discarded, since 3.4 Å is twice the van der Waal radius of the largest atom (carbon) in the system.[23] NMR shielding constants were then calculated for the dimer systems.

2.3. Basis Set Extrapolations and Correlation Effects. In this work, density functional theory (DFT; and the very popular B3LYP functional[24,25]) is used to obtain NMR shielding constants. Due to the partly semiempirical nature of the approximated exchange-correlation functionals used, B3LYP data cannot in general be expected to show convergence toward experimental shielding values or values obtained at very accurate levels of theory when increasing the basis set size.[26] It is, however, often the case that a small error can be obtained in calculated DFT chemical shielding constants, compared to high-level correlated wave function methods, if a simple linear correction or scaling factor is applied to the DFT data.[27] In this work, a comparison of B3LYP to high-level correlated methods is used to obtain such a linear scaling factor.

Unfortunately, CCSD(T) calculations with an appropriate basis set are still not possible for the FMA/benzene dimer, which has 70 electrons. Instead, we benchmark the shielding constants for FMA alone at the B3LYP, MP2, CCSD, and CCSD(T) levels of theory, in order to allow us to estimate error bounds to shielding constants obtained at levels of theory less accurate than CCSD(T).

For complete basis set limit (CBS) estimates, we use the approach of Moon and Case[26] and Kupka et al.[27,28] By using Dunning's correlation consistent basis sets[29] (cc-pVxZ; where $x \in \{D, T, Q, 5, 6, ...\}$ is the valence orbital splitting in the basis set), it is possible to carry out calculations using a sequence of basis sets of well-defined, systematic increasing quality. Kupka et al.

extrapolate calculated NMR shielding values toward infinite basis set size with a three parameter exponential decreasing function:

$$\sigma(x) = \sigma(\infty) + A \exp(-x/B) \tag{5}$$

where $\sigma(x)$ is the shielding obtained using a basis set with the valence orbital splitting number of x and $\sigma(\infty)$, A and B are the fitting parameters, with $\sigma(\infty)$ being the estimated shielding in the complete basis limit. A nonlinear least-squares Marquardt–Levenberg algorithm[30,31] is used to fit the parameters.

Jensen has constructed a set of basis sets for the purpose of Hartree–Fock (HF) and DFT NMR shielding calculations, called the polarization consistent pcS-n basis sets.[32] For basis sets of similar valence orbital splitting, the pcS-n basis sets contain more basis functions of low angular momentum, compared to the Dunning-type basis sets. pcS-1 is a double-ζ quality basis set, pcS-2 is triple-ζ, and so forth. When estimating the complete basis limit based on the pcS-n basis sets, a value of $x = n + 1$ is thus used in eq 5.

Last, we compare the proton chemical shift of the amide proton trans to the C=O bond in FMA at the CCSD(T)/CBS level of theory to the experimental value, in order to verify that CCSD(T)/CBS is, in fact, a reliable method. Inferring the experimental gas-phase ^{1}H shielding values from CH$_4$ (σ_H = 30.61 ppm[33]), an experimental value of σ_H = 26.24 ppm in the gas phase[34] at 483 K is obtained. At this temperature, thermal motion cause rapid switching of the two N-amide protons and the peaks are not separable, so this value has to be considered as an average over the two proton chemical shifts.[34]

It is well-known[35–38] that a zero-point vibrational correction (ZPVC) has to be added to equilibrium geometry ab initio shielding constants in order to obtain close agreement to experimental data. This vibrational averaging correction can easily be calculated using the method of Kern and Matcha.[39] While we had preferred to carry out a ZPVC calculation at the same level of theory as the geometry optimization of the molecules used throughout this work, no program is currently capable of automatically computing a ZPVC at the DFT level of theory with Gaussian-type basis sets. In the following, the ZPVC is calculated at the MP2/cc-pVQZ level of theory instead.

2.4. Software. All geometries were minimized at the B3LYP/aug-cc-pVTZ level of theory using Gaussian 03,[40] except when otherwise noted. All DFT calculations of NMR shielding constants were carried out using Gaussian 03. MP2/6-311++G(d,p) NMR calculations were also carried out in Gaussian 03, while the calculation of MP2 shielding constants using Dunning's correlation consistent basis sets[29] and the polarization consistent pcS-n and aug-pcS-n basis sets of Jensen[32] were carried out with Turbomole 6.2.[41] All calculations at the CCSD and CCSD(T) levels of theory were carried out using CFOUR 1.0.[42] All NMR shielding constants are calculated using the Gauge-Including Atomic Orbital formulation.[43–45] For the calculations of the ZVPC to the FMA chemical shifts, the equilibrium geometry of a planar FMA molecule was obtained at the MP2/cc-pVQZ level of theory with CFOUR 1.0, exploiting the C_S symmetry of the molecule. From this equilibrium geometry, the ZVPC to the NMR isotropic shielding was subsequently calculated at the MP2/cc-pVQZ level of theory using the method of Kern and Matcha[39] as implemented in CFOUR 1.0.

2080

dx.doi.org/10.1021/ct2002607 |*J. Chem. Theory Comput.* 2011, 7, 2078–2084

Table 2. Absolute Isotropic Chemical Shielding of the *cis-N*-amide Proton in Gas-Phase FMA at the B3LYP, MP2, CCSD, and CCSD(T) Levels of Theory Using Dunning's Correlation Consistent Basis Sets and the Polarization Consistent Shielding Basis Sets of Jensen[a]

|		method			
basis set	size	CCSD(T)	CCSD	MP2	B3LYP
cc-pVDZ	57	28.06	28.09	27.90	27.67
cc-pVTZ	132	27.29	27.35	27.16	27.17
cc-pVQZ	255	26.92	27.00	26.80	26.94
cc-pV5Z	438	26.78	26.86	26.65	26.83
$\sigma_{cc-pVxZ}(\infty)$		26.64	26.73	26.50	26.73
pcS-0	44	29.32	29.36	29.31	28.88
pcS-1	66	27.55	27.58	27.40	27.29
pcS-2	141	27.02	27.09	26.89	26.91
pcS-3	321	26.75	26.83	26.62	26.77
$\sigma_{pcS-n}(\infty)$		26.67	26.78	26.57	26.75
$\sigma_{exptl(gas)}$		26.24			

[a] All values are given as ppm. The experimental value is obtained at 483 K.[34] $\sigma(\infty)$ is obtained using eq 5 and fitted over all values in the corresponding series of basis sets. The size indicates the number of basis functions in the system at the given basis set size. All shielding constants are given in ppm.

3. RESULTS

3.1. Correlation and Basis Set Effects on the Chemical Shift of the (N—)H Proton in Formamide.
It is currently not feasible to perform a complete basis set study at the CCSD(T) level for $\Delta\delta_{RC}$ of a FMA/benzene dimer. Instead, we perform such a study of the chemical shielding of the *cis-N*-proton in FMA and use the results to benchmark more approximate methods that can be applied to FMA/benzene dimers (as described in the next subsection).

Table 2 lists CCSD(T) chemical shielding values computed using a B3LYP/aug-cc-pVTZ optimized geometry of FMA and two different, systematic series of basis sets (cc-pVxZ and pcS-n). Each set of calculations is used to extrapolate shielding constants to the complete basis set limit (as described in the previous section) and lead to very similar results: 26.64 and 26.67 ppm for cc-pVxZ and pcS-n, respectively. In the following, we will refer to 26.64 ppm as CCSD(T)/CBS, since this value is extrapolated using the largest basis set (cc-pV5Z) and since the Dunning-type basis sets are constructed for the purpose of correlated wave function calculations, whereas the pcS-n basis sets are constructed specifically for shielding constant calculations at the HF and DFT levels of theory.

The CCSD(T)/CBS value is 0.40 ppm higher than the experimental gas phase value obtained at 483 K of 26.24 ppm. However, this experimental value includes vibrational effects and is an average of the chemical shieldings of both amide protons. The effect of vibrations at 0 K (i.e., the zero-point vibrational correction) can be estimated relatively easily, as described in the previous section. At the MP2/cc-pVQZ level of theory, the ZPVC is −0.26 ppm, which, when used to correct the CCSD-(T)/CBS value, results in a chemical shielding of 26.38 ppm—within 0.14 ppm of experiment. The ZPVC correction is unlikely to contribute significantly to $\Delta\delta_{RC}$, because it is a shielding difference between two molecular systems with very
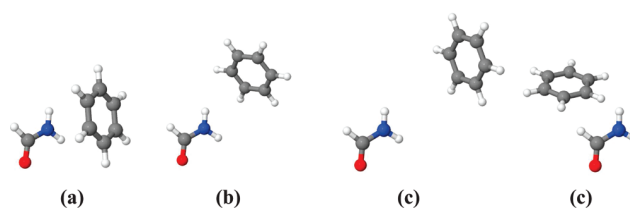
Figure 3. Pictures of the four FMA/benzene dimers used in this study. The resulting $\Delta\delta_{RC}$ calculated at various levels of theory for each conformation can be found in Table 3, where $\Delta\delta_{RC}^1$ corresponds to conformation a, $\Delta\delta_{RC}^2$ to conformation b, and so forth.

similar vibrational normal modes of FMA. Thus, we in the following focus on the electronic contribution to the chemical shielding alone.

The CBS values computed using CCSD, MP2, and B3LYP are all within 0.14 ppm of the CCSD(T)/CBS value, suggesting that the amide proton chemical shielding is relatively insensitive to electronic correlation effects in the limit of large basis sets. However, it is quite basis-set-dependent as at least the cc-pV5Z or the pcS-3 basis set is needed to get within 0.2 ppm of the CCSD(T)/CBS, with the exception of MP2/cc-pVQZ, which deviates by 0.16 ppm. We therefore choose MP2/cc-pVQZ for the $\Delta\delta_{RC}$ calculations using the FMA/benzene dimers described in the next subsection. Since $\Delta\delta_{RC}$ is a relative shielding value between two very similar systems, we expect that the MP2/cc-pVQZ results are well within 0.1 ppm of what would be computed with CCSD(T)/CBS and measured experimentally. A factor not investigated here was the dependence on the used geometry, which is known to cause deviations in calculated 1H shielding constants on the order of ±0.1 ppm—see for instance Rablen et al.[46]

3.2. Scaling B3LYP Results to Those Obtained with Correlated Wave Function Methods.
In this section, high-level correlated wave function methods are used to obtain a linear scaling correction to the chemical shift contribution due to ring current effects, obtained at the B3LYP/6-311++G(d,p)// B3LYP/aug-cc-pVTZ level of theory.

Four dimer systems were selected from the large data set of NMA/benzene dimers (see Figure 3), in such a way that the ring current contributions ($\Delta\delta_{RC}$) in the four dimer conformations cover a range from −0.72 ppm to +0.15 ppm, at the B3LYP/6-311++G(d,p) level of theory, in even sized steps. In these dimers, the NMA molecule was replaced with the much smaller FMA molecule, and the isotropic shielding was calculated using various methods and basis sets. Here, the chemical shift is modeled by

$$\Delta\delta_{RC}^{(uncorrected)} \approx \sigma_H^{Probe} - \sigma_H^{Dimer} \qquad (6)$$

where σ_H^{Probe} is the shielding of the probe nucleus in the probe molecule alone and σ_H^{Dimer} is the shielding of the probe nucleus in the probe molecule in the dimer. Note that an NMR calculation for a reference dimer is not carried out, and the linear scaling factor is unaffected, whether a reference calculation is carried out, since this calculation would also have to be scaled by the same factor. The results are collected in Table 3. We observe the following:

1. Regardless of the basis set or method, the obtained $\Delta\delta_{RC}$'s have a linear correlation to B3LYP/6-311++G(d,p) data of 0.992 or better (see the Supporting Information). It is thus demonstrated that applying a linear correction based

Table 3. Shielding Constant of the FMA Probe Proton in a Vacuum for Each Method and Basis Set Used in This Section, As Well As the Chemical Shift Ring Current Interaction ($\Delta\delta_{RC}^n$) for Each of the Four Different Conformations Used[a]

| method	σ_{FMA}	$\Delta\delta_{RC}^1$	$\Delta\delta_{RC}^2$	$\Delta\delta_{RC}^3$	$\Delta\delta_{RC}^4$	scaling
B3LYP/6-311++G(d,p)	27.64	−0.72	−0.43	−0.21	0.15	
MP2/6-311++G(d,p)	27.70	−0.76	−0.45	−0.22	0.15	1.052
CCSD/6-311++G(d,p)	27.90	−0.75	−0.44	−0.22	0.15	1.033
CCSD(T)/6-311++G(d,p)	27.85	−0.73	−0.43	−0.21	0.15	1.012
MP2/cc-pVDZ	27.90	−0.74	−0.43	−0.22	0.17	1.033
MP2/cc-pVTZ	27.16	−0.77	−0.46	−0.23	0.17	1.076
MP2/cc-pVQZ	26.80	−0.81	−0.45	−0.22	0.13	1.074
B3LYP/pcS-0	28.88	−0.75	−0.42	−0.21	0.16	1.042
B3LYP/pcS-1	27.29	−0.76	−0.44	−0.22	0.16	1.056
B3LYP/pcS-2	26.91	−0.80	−0.47	−0.24	0.14	1.087
B3LYP/pcS-3	26.77	−0.80	−0.47	−0.24	0.16	1.097
B3LYP/pcS-4	26.76	−0.80	−0.47	−0.25	0.16	1.095
CCSD(T)/CBS	26.64					
B3LYP/6-311++G(d,p) (NMA)		−0.74	−0.43	−0.22	0.14	1.004

[a] Furthermore, the resulting scaling factor relative to data obtained at the B3LYP/6-311++G(d,p) level of theory is noted. The B3LYP/6-311++G(d,p) shieldings for the identical conformation with NMA as a probe are given in the bottom row. All shielding constants and $\Delta\delta_{RC}$ values are given in ppm.

on correlated methods is a very good approximation. No constant offset (intercept) greater that 0.01 ppm was found, so the relationship between data obtained at the B3LYP level of theory and data obtained using a correlated wave function method is effectively a simple scaling factor. Thus, the fit was carried out as

$$\Delta\delta_{RC}^{Other} = k_{scaling}\Delta\delta_{RC}^{B3LYP} \qquad (7)$$

where $\Delta\delta_{RC}^{B3LYP}$ is the $\Delta\delta_{RC}$ obtained at the B3LYP/6-311++G(d,p) level of theory, $\Delta\delta_{RC}^{Other}$ is the $\Delta\delta_{RC}$ obtained using another method and/or basis set, and $k_{scaling}$ is the fitted scaling constant. Here, the $\Delta\delta_{RC}$ values are obtained via eq 6.

2. All $\Delta\delta_{RC}$ values are within 0.1 ppm of one another, including B3LYP/6-311++G(d,p), which is used for the 932 dimer calculations. This supports our previous assertion that it is easier to compute $\Delta\delta_{RC}$ accurately compared to the computation of absolute shielding constants. Therefore, the $\Delta\delta_{RC}$ values listed in Table 3 are very likely within 0.1 ppm of what would be computed with CCST(T)/CBS and measured experimentally.

3. As a result, all scaling factors are within 10% and will all yield very similar results. However, on the basis of the results in Table 2, we pick the scaling factor computed at the MP2/cc-pVQZ level, where $k_{scaling} = 1.074$.

4. The difference between ring current effects acting on either an FMA or an NMA probe was found at the B3LYP/6-311++G(d,p) level of theory to be a factor of 1.004 (see Table 3). This suggests that results obtained using FMA as a probe are, to a very good approximation, transferable to systems where NMA is used as a probe.
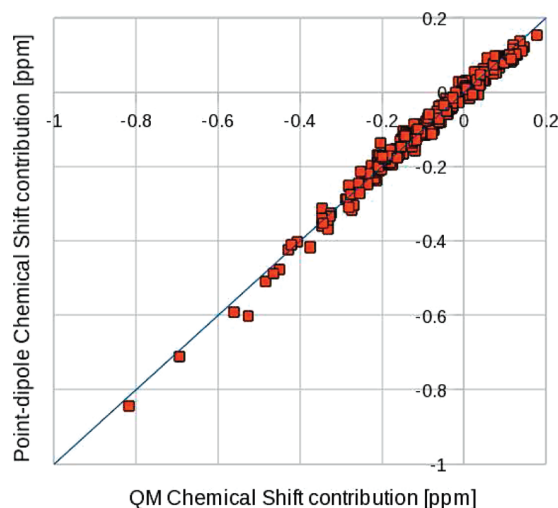
Figure 4. Correlation between the chemical shift predictions of the point-dipole model and the chemical shifts obtained by eq 4 for a set of NMA/benzene dimers using a best fit value of $B_{PD} = 30.42 \pm 0.16$ ppm Å[3]. The blue line represents the best fit between the two methods. The linear correlation of the data set is 0.993.

3.3. Expressions for the B Factors. In the point-dipole model, the definition of $i_{Benzene} \equiv 1$ is used, and the B factor in the point-dipole model (B_{PD}) is obtained via a fit using the chemical shifts obtained for all NMA/benzene dimers to their corresponding G values in the point-dipole model ($G_{PD}(\tilde{r},\theta)$, see Supporting Information), using the following formula:

$$\Delta\delta_{RC} = B_{PD}G_{PD}(\vec{r},\theta) \qquad (8)$$

where $\Delta\delta^{QM}$ is the scaled QM calculated chemical shifts of the amide protons using eq 4 and $G_{PD}(\tilde{r},\theta)$ is the geometric term of the NMA/benzene dimers in the point-dipole model. This gives a value of $B_{PD} = 30.42 \pm 0.16$ ppm Å[3]. The linear correlation of this fit is $r = 0.993$. See Figure 4 for a scatter plot of the fitted data set.

In the literature, the trend has been to use formally derived B factors in the Johnson−Bovey model (B_{JB}) and scale the relative intensities accordingly.[18] Following this, the analytical values of B_{JB} are used in the Johnson−Bovey model in this work. These evaluate to −3.79 ppm and −3.25 ppm for five- and six-membered rings, respectively. To facilitate an easy comparison of the ring current intensities to those found by Case,[18] a B factor in the Haigh−Mallion model of $B_{HM} = 5.455$ ppm Å is adopted.

3.4. Fitting the Relative i Factors. Using the B factor obtained in the previous subsection, the relative ring current intensities of all ring types in the three ring current models are obtained as the best fit of i when fitting the right-hand side of eq 4 to the right-hand side of eq 1.

The relative intensities of the two rings in tryptophan were trivially fitted using a two parameter fitting routine, although the contributions from the five- and six-membered rings were somewhat correlated. The fitted relative ring current intensities can be found in Table 4, which also features a comparison to the i factors found in other studies. A comparison is made to the values used in the SHIFTX and SHIFTS programs and to the values obtained for methane hydrogen by Case[18] as used in SPARTA. The linear correlation between the B3LYP/6-311++G(d,p) ring current contributions and the predictions of the three approximations

Table 4. Relative Ring Current Intensity Factors of the Different Side Chains, As Found in This Study, Compared to the Value of Other Studies[a]

model	point-dipole	Haigh−Mallion				Johnson−Bovey	
reference	This Work	SHIFTX[6]	SHIFTS[47]	Case[18] SPARTA[9]	This Work	Case[18]	This Work
PHE	1.00 (0.02, 0.07)	1.05 (0.05, 0.18)	1.00 (0.05, 0.17)	1.46 (0.07, 0.17)	1.18 (0.03, 0.06)	1.27 (0.03, 0.14)	1.13 (0.02, 0.06)
TYR	0.81 (0.02, 0.10)	0.92 (0.02, 0.09)	0.84 (0.02, 0.08)	1.24 (0.06, 0.22)	0.93 (0.02, 0.09)	1.10 (0.04, 0.10)	0.91 (0.02, 0.07)
HIS+	0.69 (0.02, 0.05)	0.43 (0.08, 0.29)	0.90 (0.06, 0.12)	1.35 (0.05, 0.07)	1.26 (0.03, 0.05)	1.40 (0.03, 0.08)	1.27 (0.03, 0.05)
HIS	0.68 (0.03, 0.06)	0.43 (0.08, 0.28)	0.90 (0.07, 0.11)	1.35 (0.06, 0.08)	1.22 (0.03, 0.07)	1.40 (0.04, 0.09)	1.25 (0.03, 0.06)
TRP5	0.57 (0.03, 0.08)	0.90 (0.03, 0.11)	1.04 (0.03, 0.08)	1.32 (0.04, 0.15)	0.97 (0.02, 0.10)	1.02 (0.02, 0.10)	1.06 (0.02, 0.09)
TRP6	1.02	1.04	1.02	1.24	1.18	1.27	1.18
B-factor	30.42 ppm Å^3	5.13 ppm Å	5.455 ppm Å	5.455 ppm Å	5.455 ppm Å	−3.25 ppm[b]	−3.25 ppm[b]
						−3.79 ppm[b]	−3.79 ppm[b]

[a] The RMSD associated with using the given intensity factor and B value is given as the first entry in the parentheses, and MAD as the second entry, for each intensity factor. The RMSD and MAD are calculated over all dimer systems used in the fits to obtain intensity factors. The RMSD given for tryptophan is the RMSD for a sum of both rings with the given intensities. [b] In the Johnson−Bovey model, values of −3.25 ppm and −3.79 ppm are used for six- and five-membered rings, respectively.

were found to be $r = 0.980$ or better (see Supporting Information), so using linear fits to determine the i factors is evidently a very good approximation. For the data sets for each ring type, chemical shift predictions of different sets of i factors are compared to our QM data. We present the *root-mean-square deviation* (RMSD) and the *maximum absolute deviation* (MAD) of the data set. The RMSDs to our QM values are seemingly very small for all sets of intensity parameters. However, this is mostly due to the magnitude of the ring current effect in the dimers used in the data set being on average very small, and RMSD is thus not a very good measure in this case. Our intensities, however, do have the lowest maximum RMSD of up to just 0.03 ppm, while the competing methods have RMSDs of up to 0.08 ppm (SHIFTX), 0.07 ppm (SHIFTS), 0.07 ppm (Case, Haigh−Mallion), and 0.04 ppm (Case, Johnson−Bovey) for a residue type. A much better metric than RMSD is in this case the maximum average deviation (MAD) to QM values, which loosely corresponds to the largest error one can expect from using a certain set of intensities. In this metric, our method has a MAD of 0.05−0.10 ppm or better, while the corresponding numbers for the competing methods are 0.09−0.29 ppm (SHIFTX), 0.08−0.17 ppm (SHIFTS), 0.09−0.29 ppm (Case, Haigh−Mallion), and 0.09−0.14 ppm (Case, Johnson−Bovey), for all residue types.

We note that the SHIFTX and SHIFTS predictions are, on average, slightly lower than our prediction and those of Case. One possible explanation is that the SHIFTX and SHIFTS predictions are based on empirical parameters fitted to chemical shifts measured for solution phase structures. These structures may exhibit larger conformational fluctuations, leading to a net larger average distance between the ring and the amide proton compared to the X-ray structure as well as fluctuations in the direction of the magnetic dipole arising from the aromatic side chains and, therefore, a smaller ring current effect.

4. SUMMARY

We have presented sets of ring current intensity parameters for chemical shift predictions with the point-dipole, Haigh−Mallion, and Johnson−Bovey models. The maximum errors arising from use of the presented parameters are judged to be within ±0.1 ppm from what would have been computed at the CCSD(T)/CBS for a set of 932 test cases. Further improvements in computational

methodology are thus not expected to yield any significant qualitative or quantitative improvement in chemical shift prediction in proteins. Preliminary calculations at the B3LYP/6-311++G(d,p) level using methane as a probe and intermolecular geometries corresponding to those in Table 3 suggest that the current parameters can be used to predict ring current effects on CH protons to within 0.2 ppm.

The presented parameters are rigorously based on the underlying physical properties of aromatic molecules. Parameters based on empirical models were found to perform worse on our amide proton test set. Our report of the superiority of a physics-based method over empirical methods is backed up by the fact that the parameters obtained by Case[18] through QM methods have the same disagreements with the empirical methods as our parameters, despite the fact that the computational methodology used by Case was somewhat different than ours.

Finally, we have made a detailed numerical comparison between the point-dipole, Haigh−Mallion, and Johnson−Bovey models. The chemical shift predictions of the three models were nearly identical, and no outliers compared to our quantum mechanical calculations were found in any of the three models. Apart from reported problems with predictions of ring current effects in macrocyclic rings, such as those found in porphyrins,[48] which should be a nonissue for most uses in protein chemical shift predictions, the three methods should yield results of identical accuracy. Hence, we suggest that the point-dipole model should be used in future chemical shift prediction software, since (1) it both is computationally faster than competing models, since it does not require any integral evaluation as opposed to the Johnson−Bovey model, and contains significantly fewer geometric terms than the Haigh−Mallion model and (2) it is much easier to implement than the competing models.

■ ASSOCIATED CONTENT

ⓢ Supporting Information. Thorough descriptions of the point-dipole, Haigh−Mallion, and Johnson−Bovey models and our implementations are described. The material also includes sketches of the used molecules, linear correlation values and scatter plots of the fits used to obtain the intensity values of Table 4, and an additional investigation of the equilibrium geometry dependence of FMA NMR shielding calculations.

This information is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Authors
*E-mail: andersx@nano.ku.dk; jhjensen@chem.ku.dk.

■ REFERENCES

(1) Raman, S.; Lange, O.; Rossi, P.; Tyka, M.; Wang, X.; Aramini, J.; Liu, G.; Ramelot, T.; Eletsky, A.; Szyperski, T.; Kennedy, M. A.; Prestegard, J.; Montelione, G. T.; Baker, D. *Science* **2010**, *327*, 1014–1018.

(2) Meiler, J.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 15404–15409.

(3) Jensen, M. R.; Salmon, L.; Nodet, G.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 1270–1272.

(4) Vila, J. A.; Scheraga, H. A. *Acc. Chem. Res.* **2009**, *42*, 1545–1553.

(5) Robustelli, P.; Kohlhoff, K. J.; Cavalli, A.; Vendruscolo, M. *Structure* **2010**, *18*, 923–933.

(6) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *J. Am. Chem. Soc.* **1991**, *111*, 9436–9444.

(7) Xu, X. P.; Case, D. A. *J. Biomol. NMR* **2001**, *21*, 321–333.

(8) Xu, X. P.; Case, D. A. *Biopolymers* **2002**, *65*, 408–423.

(9) Shen, Y.; Bax, A. *J. Biomol. NMR* **2007**, *38*, 289–302.

(10) Meiler, J. *J. Biomol. NMR* **2003**, *26*, 25–37.

(11) Haigh, C. W.; Mallion, R. B. *Prog. NMR Spectrosc.* **1980**, *13*, 303–344.

(12) Moon, S.; Case, D. A. *J. Biomol. NMR* **2007**, *38*, 139–150.

(13) Mulder, F. A.; Filatov, M. *Chem. Soc. Rev.* **2010**, *395*, 578–590.

(14) Monya, G.; Zauhar, R. J.; Williams, H. J.; Nachman, R. J.; Scott, A. I. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 702–709.

(15) Johnson, C. E.; Bovey, F. A. *J. Chem. Phys.* **1958**, *29*, 1012–1014.

(16) Pople, J. A. *J. Chem. Phys.* **1956**, *24*, 1111.

(17) Pople, J. A. *Mol. Phys.* **1958**, *1*, 175–180.

(18) Case, D. A. *J. Biomol. NMR* **1995**, *6*, 341–346.

(19) Boyd, J.; Skrynnikov, N. R. *J. Am. Chem. Soc.* **2002**, *124*, 1832–1833.

(20) Berman, J.; Westbrook, H. M.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. *Nucleic Acids Res.* **2000**, *106*, 16972–16977.

(21) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. *Nucleic Acids Res.* **2007**, *35*, 522–525.

(22) Dolinsky, T. J.; Nielsen, J.; McCammon, J. A.; Baker, N. A. *Nucleic Acids Res.* **2004**, *32*, 665–667.

(23) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.

(24) Becke, A. D. *J. Phys. Chem.* **1993**, *98*, 5648–5652.

(25) Stephens, P.; Devlin, F.; Chabalowski, C.; Frisch, M. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(26) Moon, S.; Case, D. A. *Nucleic Acids Res.* **2004**, *32*, 665–667.

(27) Kupka, T.; Ruscic, B.; Botto, R. E. *J. Phys. Chem. A* **2002**, *106*, 10396–10407.

(28) Kupka, T.; Ruscic, B.; Botto, R. E. *Solid State Nucl. Magn. Reson.* **2003**, *23*, 145–167.

(29) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(30) Marquardt, D. W. *J. Appl. Math.* **1963**, *11*, 431–441.

(31) Levenberg, K. *Q. Appl. Math.* **1944**, *2*, 164–168.

(32) Jensen, F. *J. Chem. Theory Comput.* **2008**, *5*, 719–727.

(33) Jameson, K. A.; Jameson, C. J. *Chem. Phys. Lett.* **1987**, *134*, 461–466.

(34) Vaara, J.; Kaski, J.; Joksaari, J.; Diehl, P. *J. Phys. Chem. A* **1997**, *101*, 5069–5081.

(35) Ruud, K.; Astrand, P.-O.; Taylor, P. R. *J. Am. Chem. Soc.* **2001**, *123*, 4826–4833.

(36) Sauer, S. P. A.; Spirko, V.; Paidarová, I.; Kraemer, W. P. *Chem. Phys.* **1997**, *214*, 91–102.

(37) Wigglesworth, R. D.; Raynes, W. T.; Sauer, S. P. A.; Oddershede, J. *Mol. Phys.* **1999**, *96*, 1595–1607.

(38) Wigglesworth, R. D.; Raynes, W. T.; Kirpekar, S.; Oddershede, J.; Sauer, S. P. A. *J. Chem. Phys.* **2000**, *112*, 736–746.

(39) Kern, C. W.; Matcha, R. L. *J. Phys. Chem.* **1968**, *49*, 2081–2092.

(40) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, A. J., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.

(41) Ahlrichs, R.; Baer, M.; Haeser, M.; Horn, H.; Koelmel, C. *Chem. Phys. Lett.* **1989**, *162*, 165.

(42) Stanton, J. F.; Gauss, J.; Harding, M. E.; Szalay, P.; Auer, A. A.; Bartlett, R. J.; Benedikt, U.; Berger, C.; Bernholdt, D. E.; Bomble, Y. J.; Christiansen, O.; Heckert, M.; Heun, O.; Huber, C.; Jagau, T.-C.; Jonsson, D.; Juselius, J.; Klein, K.; Lauderdale, W. J.; Matthews, D. A.; Metzroth, T.; O'Neill, D. P.; Price, D. R.; Prochnow, E.; Ruud, K.; Schiffmann, F.; Stopkowicz, S.; Tajti, A.; Vazquez, J.; Wang, F.; Watts, J. D.; Almlöf, J.; Taylor, P. R.; Taylor, P. R.; Helgaker, T.; Jensen, H. J. A.; Jorgensen, P.; Olsen, J.; Mitin, A. V.; van Wüllen, C. CFOUR, a quantum chemical program package. For the current version, see http://www.cfour.de (accessed June 2010).

(43) London, F. *J. Phys. Radium* **1937**, *8*, 397–409.

(44) Ditchfield, R. *Mol. Phys.* **1974**, *27*, 789–807.

(45) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. *J. Chem. Phys.* **1996**, *104*, 5497–5509.

(46) Rablen, P. R.; Pearlman, S. A.; Finkbiner, J. *J. Phys. Chem. A* **1999**, *103*, 7357–7363.

(47) Ösapay, K.; Case, D. A. *J. Am. Chem. Soc.* **1991**, *111*, 9436–9444.

(48) Perkins, S. J. *Applications of ring current calculations to proton NMR of proteins and transfer RNA*; Plenum Press: New York, 1982; Vol. 4, Chapter 4, pp 193–336.

PHAISTOS: A Framework for Markov Chain Monte Carlo Simulation and Inference of Protein Structure

Wouter Boomsma,[a,b]* Jes Frellsen,[a] Tim Harder,[a,c] Sandro Bottaro,[d,e] Kristoffer E. Johansson,[a] Pengfei Tian,[f] Kasper Stovgaard,[a] Christian Andreetta,[a,g] Simon Olsson,[a] Jan B. Valentin,[a] Lubomir D. Antonov,[a] Anders S. Christensen,[h] Mikael Borg,[a,i] Jan H. Jensen,[h] Kresten Lindorff-Larsen,[a] Jesper Ferkinghoff-Borg,[e] and Thomas Hamelryck[a]

We present a new software framework for Markov chain Monte Carlo sampling for simulation, prediction, and inference of protein structure. The software package contains implementations of recent advances in Monte Carlo methodology, such as efficient local updates and sampling from probabilistic models of local protein structure. These models form a probabilistic alternative to the widely used fragment and rotamer libraries. Combined with an easily extendible software architecture, this makes PHAISTOS well suited for Bayesian inference of protein structure from sequence and/or experimental data. Currently, two force-fields are available within the framework: PROFASI and OPLS-AA/L, the latter including the generalized Born surface area solvent model. A flexible command-line and configuration-file interface allows users quickly to set up simulations with the desired configuration. PHAISTOS is released under the GNU General Public License v3.0. Source code and documentation are freely available from http://phaistos.source-forge.net. The software is implemented in C++ and has been tested on Linux and OSX platforms. © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23292

Introduction

Two methods dominate the field of molecular simulation: molecular dynamics (MD) and Markov chain Monte Carlo (MCMC). The main difference between the methods lies in the way the system is updated in each iteration. MD involves iterating between calculating the forces exerted on each particle in the system and using Newton's equations of motion to update their positions. In contrast, MCMC is a statistical approach, where the goal is to generate samples from a probability distribution associated with the system, typically a Boltzmann distribution. MD has generally been regarded as best-suited for exploring dense molecular systems such as the native ensemble of proteins, while MCMC methods can be more efficient for longer time scale simulations involving large structural rearrangements.[1] Using optimized move sets it has, however, been demonstrated that even in the densely packed native state, MCMC can serve as an efficient alternative to MD.[2–4] In addition, the statistical nature of MCMC methods make them particularly well-suited for Bayesian inference of protein structure from experimental data.[5]

The freedom in the choice of moves in Monte Carlo simulations means that there is potential progress to be made in designing new, improved move types, thereby further increasing the time scales and molecular sizes amenable to simulation. In this article, we present a software framework designed with this goal in mind. The PHAISTOS framework contains implementations of recently developed tools that increase the

[a] W. Boomsma, J. Frellsen, T. Harder, KE. Johansson, K. Stovgaard, C. Andreetta, S. Olsson, JB. Valentin, L. D. Antonov, M. Borg, K. Lindorff-Larsen and T. Hamelryck
Department of Biology, University of Copenhagen, Copenhagen, 2200, Denmark
E-mail: wb@bio.ku.dk

[b] W. Boomsma
Department of Astronomy and Theoretical Physics, University of Lund, Lund, SE-223 62, Sweden

[c] T. Harder
Center for Bioinformatics, University of Hamburg, Hamburg, 20146, Germany

[d] S. Bottaro
Scuola Internazionale Superiore di Studi Avanzati, Trieste, 34136, Italy

[e] S. Bottaro and J. Ferkinghoff-Borg
Department of Biomedical Engineering, DTU Elektro, DTU, Kongens Lyngby, 2800, Denmark

[f] P. Tian
Niels Bohr Institute, University of Copenhagen, Copenhagen, 2100, Denmark

[g] C. Andreetta
Computational Biology Unit, Uni Computing, Uni Research, Norway

[h] AS. Christensen and JH. Jensen
Department of Chemistry, University of Copenhagen, Copenhagen, 2100, Denmark

[i] M. Borg
BILS, Science for Life Laboratory, Box 1031, Solna, 171 21, Sweden

efficiency and scope of MCMC-based simulations. Through a modular design, the software can easily be extended with new move types and force-fields, making it possible to experiment with novel Monte Carlo strategies. Finally, using flexible configuration file and command line options, users can quickly set up simulations with any combination of moves, energy terms, and other simulation settings.

By making our methods available in an easily extendible, open source framework, we hope to further encourage the use of MCMC for protein simulations and promote the development of new MCMC methodologies for the simulation, prediction, and inference of protein structure.

Methodology

The framework is split into four main types of components: moves, energy terms, observables, and Monte Carlo methods. For each of these types, a number of algorithms are available. Moves and energies are normally used in sets: a weighted set of moves is referred to as a move collection, while an energy function is composed of a weighted sum of energy terms. Observables are similar to energy terms, but are typically only evaluated at certain intervals to extract statistics during a simulation. In the following description, each algorithm is annotated with its corresponding command line option name in a monospace font.

Moves

One of the main distinguishing features of the PHAISTOS package is efficient sampling, obtained through an elaborate set of both established and novel Monte Carlo moves. Each move stochastically modifies a protein chain in a specific way. Weighted sets of these moves can be selected from the command line, allowing the user to easily experiment and fine-tune the set of moves for a given simulation scenario. All moves in PHAISTOS can be applied such that detailed balance is obeyed, which ensures, if the sampling is ergodic, that simulations sample from a well-defined target distribution (e.g., the canonical or multicanonical ensemble).

The framework contains many of the established moves from the literature, including various pivot moves (`move-pivot-uniform`, `move-pivot-local`), the crankshaft/backrub local move (`move-crankshaft`),[6,7] the CRA local move (`move-cra`),[8] and the semilocal biased Gaussian step (BGS) (`move-semilocal`).[9] Side-chain conformational sampling can be done either from Gaussian distributions given by rotamer libraries (`move-sidechain-rotamer`)[10] or through Gaussians centered around the current side-chain conformation (`move-sidechain-local`).

Moves using Probabilistic Models. PHAISTOS has broad support for sampling using biased proposals. Usually, if an MCMC simulation were to be conducted without the presence of a force-field, a uniform distribution in configurational space would be obtained. In the case of biased sampling, moves are instead allowed to follow a specific distribution during the

simulation. This bias can then, optionally, be divided out, so that it does not influence the final statistical ensemble, but only serves to increase sampling efficiency by focusing on the most important regions of conformational space. If the bias is left in, it corresponds to an implicit extra term in the energy function.

Typically, the bias is chosen to reflect prior knowledge about the local structure of the molecule. A good example is the common use of fragment and rotamer libraries for structure prediction.[11,12] These methods are used strictly for sampling, and the introduced bias is not easily quantifiable, which also makes it difficult to ensure detailed balance for the Markov chain. In contrast, PHAISTOS includes a number of moves based on probabilistic models, which support both sampling of conformations and the evaluation of the bias introduced with those moves. This makes them uniquely suited for use in MCMC simulations.

Four different structural, probabilistic models are available: FB5HMM models the Cα trace of a protein,[13] COMPAS models a reduced single-particle representation of amino acid side-chains, whereas TORUSDBN and BASILISK, respectively, model backbone and side-chain structure in atomic detail.[14,15] All models can be applied both as proposal distributions in the form of Monte Carlo moves (`move-backbone-dbn`, `move-sidechain-basilisk`, `move-sidechain-compas`) and as probabilistic components of an energy function (`energy-backbone-dbn`, `energy-basilisk`, `energy-compas`). Figure 1 illustrates how dihedral angles are sampled from a TORUSDBN-like model of the protein backbone. The practical details on how probabilistic models can be incorporated in an energy function are discussed in the "Energies" section.
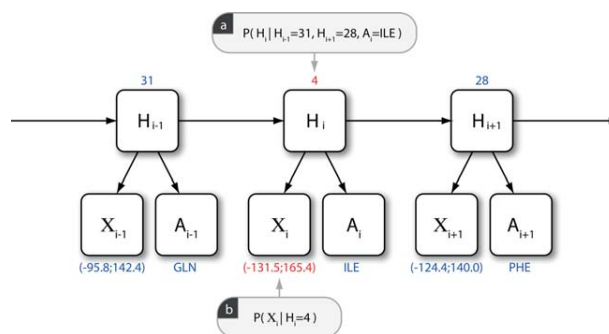
Figure 1. An illustration of a simplified version of the TORUSDBN model of backbone local structure, showing the architecture of the dynamic Bayesian network (DBN) and an example of values for the individual nodes. Each A node is a discrete distribution over amino acids, whereas each X node is a bivariate distribution over (ϕ, ψ) angle pairs. The hidden node (H) sequence is a Markov chain of discrete states, representing the sequence of residues in a protein chain. Each hidden node state corresponds to a particular distribution over angle pairs and amino acid labels. The values highlighted in red are the result of a single resampling step of the (ϕ, ψ) angle pair at some position i in the chain: a) The hidden node state H_i is resampled based on the current values of the values of neighboring H values and the amino acid label at position i ($P(H_i|H_{i-1}, H_{i+1}, A_i) \propto P(H_i|H_{i-1})P(H_{i+1}|H_i)P(A|H_i)$); b) A (ϕ, ψ) value is drawn from the bivariate angular distribution corresponding to the sampled H value ($P(X_i|H_i)$). A full description of the TORUSDBN model can be found in the original publication.[14]

Efficient Local Updates. An important challenge in Monte Carlo simulations is to ensure efficient sampling in dense states where proposed conformational changes will have a high probability of containing self-collisions. In particular, pivot moves will typically have very poor acceptance rates in this scenario. The solution is typically to expand the move set to include local moves, which only change the atom positions within a small segment of the chain.

In addition to various established local move methods from the literature, PHAISTOS includes a novel method, called CRISP[4] (`move-crisp`), which is particularly well-suited for this problem. Unlike other local move approaches,[7,8] CRISP is able to generate updates to a segment of the chain without disrupting its local geometry. Often, local move algorithms are designed as a two-step process, where some angular degrees of freedom are modified stochastically (prerotation), whereas others are modified deterministically to bring the chain back to a closed state (postrotation). From the work of Gō and Scheraga,[16] it is known that in general, six degrees of freedom are required for the postrotation step. CRISP distinguishes itself from previous methods by merging these two steps, modifying the stochastic prerotation step so that it takes the resulting postrotation step into account. More precisely, for each application of the move, a random segment of the protein is selected, and a multivariate Gaussian distribution is constructed over the angular change in the $n-6$ prerotation degrees of freedom $\delta\bar{\chi}_{\text{pre}}$

$$P\left(\delta\bar{\chi}_{\text{pre}}\right) \propto \exp\left(-\frac{1}{2}\delta\bar{\chi}_{\text{pre}}^{T}\lambda\left(\mathbf{C}_{n-6}+\mathbf{S}^{T}\mathbf{C}_{6}\mathbf{S}\right)\delta\bar{\chi}_{\text{pre}}\right). \quad (1)$$

Here, $\mathbf{C}$ is an inverse diagonal covariance matrix specifying the desired fluctuations for the individual angular degrees of freedom, $\mathbf{S}$ is a linear transformation mapping the prerotational degrees of freedom to the corresponding postrotational values, and λ is a scaling parameter determining the size of the move. In effect, to first order, the method samples from a distribution of closed chain structures, ensuring high-quality local structure in all samples. We have recently shown that this has a dramatic impact on simulation performance, in particular for dense molecular systems.[4]

Low acceptance rates are also sometimes observed in side-chain moves. When using a fine-grained force-field such as OPLS-AA/L, we have experienced that the standard resampling of side-chains can be overly intrusive. Particularly in the case of side-chains involved in several simultaneous hydrogen bonds, traditional moves tend to break all hydrogen bonds at once, typically leading to the rejection of such updates. To avoid this problem, PHAISTOS includes a novel move (`side-chain-local`) that, for a given side-chain, randomly selects an atom that potentially participates in hydrogen bonds, and constrains its position using a technique similar to that of the semilocal BGS backbone move.[4,9]

Energies

Two established force-fields are currently implemented within the framework: the PROFASI force-field[17] and the

OPLS-AA/L[18] force-field in combination with the generalized Born surface area (GB/SA) implicit solvent model.[19] These represent two extremes in the range of force-fields available in the literature: an ultrafast force-field modeling effective interactions in the presence of a solvent and a classic fine-grained molecular mechanics force-field combined with a more accurate implicit solvent model. The two force-fields were selected to provide support for a broad range of simulation tasks. The efficiency of the PROFASI force-field makes it possible readily to conduct reversible folding simulations of peptides and small proteins.[17] The OPLS-AA/L force-field in combination with the GB/SA solvent model is more accurate, but also significantly slower, and is typically used for exploring the details of native ensembles. It can also be used for structure refinement, for instance of structures obtained in a reversible folding simulation using PROFASI. For increased efficiency, all nonbonded force-field terms in both force-fields have been implemented using the chaintree data structure,[20] which avoids recalculation of energy contributions that are not modified in a given iteration of the simulation. Together with effective local moves, this can result in a considerable computational speed-up.

PROFASI. The PROFASI force-field consists of four terms[17]

$$E = E_{\text{ev}} + E_{\text{hb}} + E_{\text{sc}} + E_{\text{loc}} \quad (2)$$

where E_{ev} captures excluded volume effects, E_{hb} is a hydrogen bond term, E_{sc} is a side-chain interaction term, and E_{loc} concerns the local interactions along the chain. The excluded volume potential is a simple r^{-12} interaction between all atom pairs, where r denotes the distance between the atoms. The strength of a hydrogen bond in PROFASI depends on the detailed geometry of the bond, parameterized through the N—H—O and H—O—C angles. The side-chain potential consists of a charge–charge and a hydrophobicity contribution. For each residue pair, these consist of a product between a conformation-dependent contact strength and an energy that depends on the specific amino acid types involved in the bond. Finally, the local energy term captures interactions between partial charges in neighboring peptide units along the chain, with a correction term for improved consistency with the Ramachandran plot, and a side-chain torsion potential. As bond angles and bond lengths are assumed fixed during PROFASI simulations, no further local interactions are included.

A distinguishing feature of the PROFASI force-field is the presence of a global interaction cutoff of 4.5 Å. Although this necessarily excludes various long-range interactions, it is also one of the main reasons behind the efficiency of the force-field. Despite this restriction, the force-field has been demonstrated to successfully fold a range of peptides and small proteins,[17] while still being fast enough for many-body aggregation simulations.[21,22]

OPLS-AA/L. In contrast to PROFASI, the OPLS-AA/L force-field includes local terms for bond angles, bond lengths, and torsions. The bond angle and bond length potentials are simple harmonic terms, whereas the torsion term has the form[18]

$$E_{\text{torsion}} = \sum_i \sum_{j=1}^{3} w_j \left(1 + (-1)^{j+1} \cos(j\theta_i) \right) \qquad (3)$$

where the outer sum iterates over all dihedrals θ_i. The non-bonded interactions include standard Lennard-Jones and Coulomb potentials

$$E_{\text{nb}} = \sum_{i>j} w_{ij} \left(4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) + \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}} \right) \qquad (4)$$

where r_{ij} is the distance between atoms i and j, q_i and q_j are the corresponding partial charges, ε_0 is the vacuum permittivity, and σ_{ij} and ϵ_{ij} are calculated using the combination rules $\sigma_{ij} = \sqrt{\sigma_i \sigma_j}$ and $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$, respectively. Finally, w_{ij} works to exclude interactions between atoms that are separated by only a few covalent bonds. Thus, $w_{ij} = 0.0$ for direct neighbors (1,2) and pairs separated by a single other atom (1,3), $w_{ij} = 0.5$ for pairs separated by two atoms (1,4), and $w_{ij} = 1.0$ for all others.

Our implementation of OPLS-AA/L follows that of the Tinker simulation package.[23] We ensured that energies produced by our program match those obtained when running Tinker.

GB/SA. The PROFASI force-field is parameterized to capture effective interactions in the presence of a solvent. In contrast, OPLS-AA/L should be combined with a suitable solvent model to reproduce physiological conditions. To model the effect of the solvent on hydrophobic interactions and electrostatics, we use the OPLS-AA/L force-field in combination with the GB/SA implicit solvent model.[19]

Many implicit solvent models express the solvation free energy G_{solv} as a sum of nonpolar and electrostatic contributions

$$G_{\text{solv}} = G_{\text{npol}} + G_{\text{pol}} \qquad (5)$$

Here, G_{npol} is the free energy of solvating the molecule with all the partial charges set to zero, and G_{pol} is the reversible work required to increase the charges from zero to their full values.[24] In GB/SA, the nonpolar contribution G_{npol} is assumed to be proportional to the solvent accessible surface area, while the generalized Born approximation is used to calculate the electrostatic solvation energy using the pairwise summation[25]

$$G_{\text{pol}} = -\frac{1}{8\pi\varepsilon_0} \left(1 - \frac{1}{\varepsilon} \right) \sum_{i,j}^{n} \frac{q_i q_j}{f_{\text{GB}}} \qquad (6)$$

where ε is the dielectric constant of the solvent, and q_i is the partial charge of atom i. $f_{\text{GB}} = \sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp\left(-r_{ij}^2/4\alpha_i\alpha_j\right)}$ is a function of the distance r_{ij} and of the so-called Born radii α, which reflects the average distance of the charged atom to the dielectric medium. For our implementation, the Born radii are calculated using an analytical expression proposed by Still and coworkers.[19]

Incorporating Probabilistic Models in the Energy Function. When using moves that are based on probabilistic models such as TORUSDBN and BASILISK, it gives rise to a bias in the simulation, which can be regarded as an implicit energy term. In PHAISTOS, the energy contributions of these probabilistic models can also be evaluated explicitly, by adding them as a term to the energy function. This makes it possible to use the probabilistic models as energies in a simulation with a standard set of unbiased moves, or to compensate for the bias of a move by adding the corresponding energy term with negative weight. When used as energies, the values are reported in minus log-probabilities. To facilitate the combination of classic energy terms with probabilistic terms, the energies of physical force-fields such as PROFASI and OPLS-AA/L are likewise reported as minus log-probabilities: they are multiplied by $-1/kT$, where T is the simulation temperature and k is the Boltzmann constant.

As an example, for the TORUSDBN-like model in Figure 1, the log-likelihood for a given state is

$$LL(\bar{X}, \bar{A}) = \ln \sum_{\bar{H}} P(\bar{X}, \bar{A}, \bar{H})$$
$$= \ln \sum_{\bar{H}} P(X_1|H_1)P(A_1|H_1)P(H_1) \prod_{i=2}^{N} P(X_i|H_i)P(A_i|H_i)P(H_i|H_{i-1})$$

$$(7)$$

where $\bar{X}$, $\bar{A}$, and $\bar{H}$ are the sequences of angle pairs, amino acid labels, and hidden node labels, respectively, i is the residue index, and N is the sequence length. Each hidden node label is the index of a component of the emission distributions of the model. For instance, the Ramachandran distribution is modeled as a weighted sum of bivariate von Mises distribution components.[14] The hidden nodes are "nuisance" parameters and are therefore summed out in the evaluation of the likelihood. Note that the sum runs over all possible hidden node *sequences*, a calculation that can be done efficiently using dynamic programming.[14]

Observables

Observables in PHAISTOS allow a user to extract information about the current state of a simulation. Examples of observables include root-mean-square-deviation (RMSD) (`observable-rmsd`) and radius of gyration (`observable-rg`). In addition, all energy terms are also available as observables. A user can specify a selection of observables from the command line or settings file, choosing how frequently they should be registered and in which format. While most observables will return a single value, others have more elaborate outputs, such as dumping of complete structural states to PDB files (`observable-pdb`) or to a molecular trajectory file in the Gromacs XTC format (`observable-xtc-trajectory`).[26] Finally, observables can be dumped to the header or as *b*-factors in outputted PDB-files. The latter makes it possible to annotate structures with residue-specific information, such as the number of contacts, degree of burial, and more sophisticated evaluations of the environment of each residue.[27]

Monte Carlo

Although PHAISTOS can be used for Monte Carlo minimization, the primary focus of the framework is MCMC simulation,

where the goal is to produce samples from the Boltzmann distribution corresponding to a given force-field at a specified temperature. All moves in PHAISTOS are, therefore, designed to be compatible with the property of detailed balance, in the sense that their proposal probabilities can be evaluated. That is, for a move from state x to state x'

$$\pi(x)P(x \rightarrow x') = \pi(x')P(x' \rightarrow x) \tag{8}$$

where $\pi(x)$ is the stationary distribution, and $P(x \rightarrow x')$ is the probability of moving from state x to x' using a given move. Factoring $P(x \rightarrow x')$ into a "selection" probability P_s and an "acceptance" probability P_α, we have

$$\frac{P_\alpha(x \rightarrow x')}{P_\alpha(x' \rightarrow x)} = \frac{\pi(x')P_s(x' \rightarrow x)}{\pi(x)P_s(x \rightarrow x')} \tag{9}$$

Most of the moves are symmetric, in the sense that $P_s(x' \rightarrow x)/P_s(x \rightarrow x') = 1$. However, for moves such as the local and semilocal moves, this is not the case, and it is important that this bias be correctly compensated for. Implementation-wise, each Move object is responsible for calculating the bias that it introduces, and the Monte Carlo class will then compensate for it when necessary.

Metropolis-Hastings.   The most common way to ensure that eq. (9) is fulfilled is to use the Metropolis-Hastings (MH) acceptance criterion

$$P_\alpha(x \rightarrow x') = \min\left(1, \frac{\pi(x')P_s(x' \rightarrow x)}{\pi(x)P_s(x \rightarrow x')}\right) \tag{10}$$

This is the default simulation method used in PHAISTOS (`monte-carlo-metropolis-hastings`). It is useful for exploring near-native ensembles and can be efficient when simulating at the critical temperature of a system. However, for more complicated systems, MH simulations tend to spend excessive periods of time exploring local minima, leading to poor mixing, and, therefore, slow convergence.

Generalized Ensembles.   To avoid the mixing problems associated with standard MH simulations, PHAISTOS includes support for conducting simulations in generalized ensembles.[28] Rather than sampling directly from the Boltzmann distribution, the central idea is to generate samples from a modified distribution, and subsequently reweight the obtained statistics to the Boltzmann distribution at a desired temperature. The acceptance criterion becomes

$$P_\alpha(x \rightarrow x') = \min\left(1, \frac{w(x')P_s(x' \rightarrow x)}{w(x)P_s(x \rightarrow x')}\right) \tag{11}$$

for a given weight function $w(x)$. The typical choice is the multicanonical ensemble,[29] corresponding to a flat distribution over energies. That is, $w(x) = 1/g(E(x))$, where $E(x)$ is the energy associated with conformational state x, and g is the number of states associated with a given energy (density of states). Another example is the $1/k$ ensemble, which attempts to provide ergodic sampling while maintaining primary focus on the low energy states.[30] In this case, the weight function is $w(x) = 1/k(E(x))$,

where $k(E(x)) = \int_{-\infty}^{E(x)} g(\hat{E}) d\hat{E}$. It can be shown that this is approximately equivalent to a flat histogram over $\ln(E(x))$.[30]

We have recently developed an automated method, MUNINN, for estimating the weights w in generalized ensemble simulations (http://muninn.sourceforge.net/). It employs the generalized multihistogram equations,[31] and uses a non-uniform adaptive binning of the energy space, ensuring efficient scaling to large systems. In addition, MUNINN allows weights to be restricted to cover a limited temperature range of interest. The MUNINN functionality is seamlessly integrated into PHAISTOS and can be activated by selecting the corresponding Monte Carlo engine (`monte-carlo-muninn`).

Monte Carlo Minimization.   PHAISTOS contains a few simulation algorithms that are directed at optimization, rather than sampling. These include a simulated annealing class (`monte-carlo-simulated-annealing`) and a greedy Monte Carlo optimization class (`monte-carlo-greedy-optimization`), which are useful in cases where the user is interested in a single low-energy structure, rather than a full structural ensemble.

Program design

The framework is designed to be modular, both in software design and in the choices exposed to the user from the command line or settings file. As illustrated in the UML diagram in Figure 2, all energy terms are derived from the same base class and implement the same interface. Energy functions can easily be constructed from the command line or settings file by including the energy terms of interest. Moves and Monte Carlo simulation algorithms are structured in a similar way. This design makes it straightforward to implement new energy terms, moves or simulation algorithms with little knowledge of the overall code. Iterators are provided for easy iteration over atoms or residues in a molecule. In addition, caching and rapid determination of interacting atom pairs is made possible by an implementation of the chaintree algorithm.[20]

Finally, through a modular build-system, developers can readily write their own modules utilizing the library. Modules are separate code entities that are autodetected by the build system when present and can be enabled and disabled at compile time, making it easy to share code among collaborators.

Example

We include a step-by-step walk-through of the PHAISTOS simulation process. The goal is to conduct a reversible folding simulation of the 20-residue beta3s peptide,[32] demonstrating several of the features described earlier.

The user interface of PHAISTOS is designed to make it as easy as possible to set up simulations. Almost all options have default values, and it is, therefore, usually sufficient to supply only a few input options to the program. The program behavior can then gradually be fine-tuned using additional options in the configuration file later on. For this particular example, we use the following command from the command line
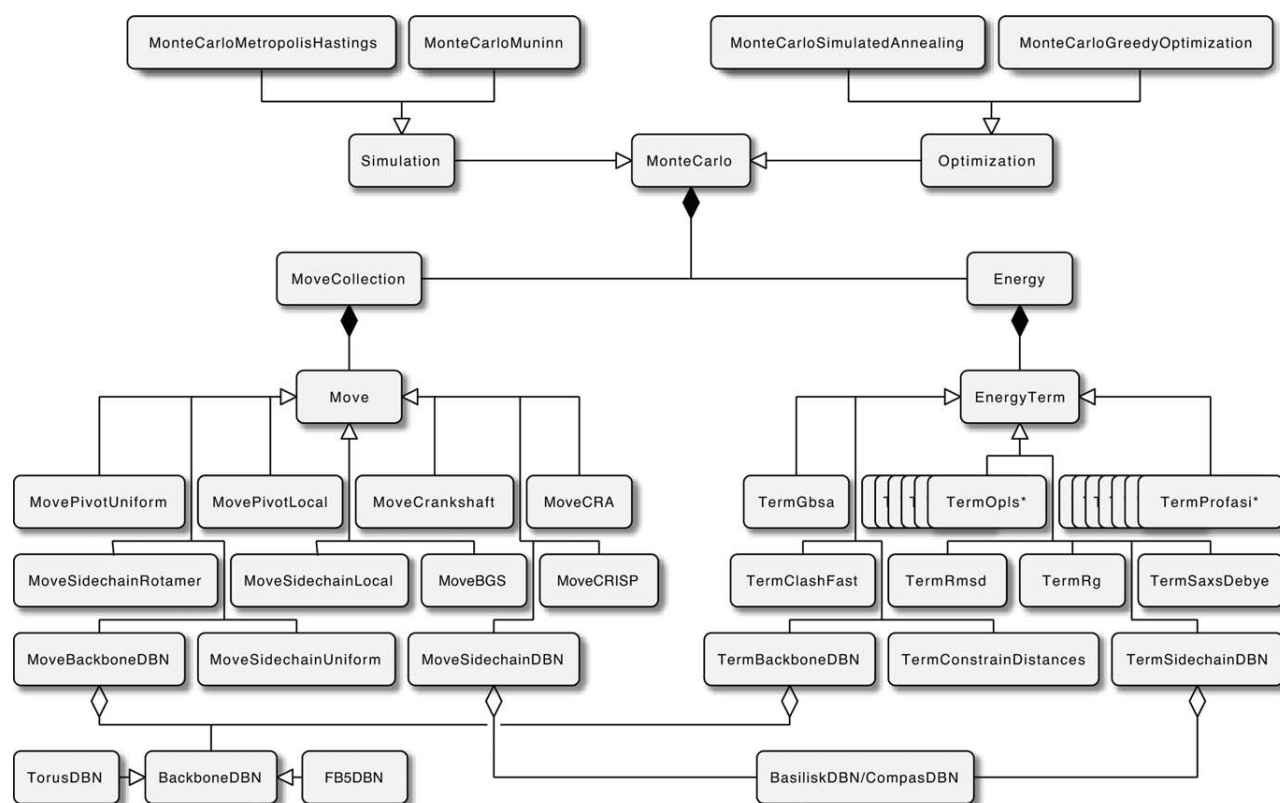
Figure 2. A UML-diagram of the major classes in the PHAISTOS library (black diamond: composition, white diamond: aggregation, arrow: inheritance). A Monte Carlo simulation object contains a MoveCollection object, which consists of a selection of moves, and an Energy object, which comprises a number of energy terms (TermOpls* and TermProfasi* denote the entire set of OPLS and PROFASI energy terms, respectively). Note that the probabilistic models (BackboneDBN/BasiliskDBN/CompasDBN) are available both as energy terms and as moves. A detailed description of all classes can be found in the Doxygen documentation on the PHAISTOS web site.

```
$ ./phaistos --aa-file beta3s.aa \
          --energy profasi-cached backbone-dbn [weight:-1] \
          --move backbone-dbn sidechain-uniform semilocal-dbn-eh \
          --threads 8 --temperature 283 \
          --monte-carlo muninn[min-beta:0.6,max-beta:1.1] \
          --observable backbone-dbn rmsd[reference-pdb-file:beta3s. pdb] \
          --observable xtc-trajectory
```

The aa-file argument specifies that we are reading an amino acid sequence from a file. The energy and move options select the relevant energy terms and moves, respectively. In this case, we use a cached version of the PROFASI force-field, and sample using TORUSDBN moves, uniformly distributed side-chain moves, and BGS moves using the TORUSDBN as a prior. We specify that the simulation should be conducted in eight parallel threads and set the temperature to 283 K. MUNINN is chosen to be the Monte Carlo engine, using a β (inverse temperature) range of [0.6; 1.1]. These β factors are unit-less, specified relative to the inverse temperature, and thus correspond to a temperature range of $\left[(1.1 \cdot 1/283K)^{-1}; (0.6 \cdot 1/283K)^{-1}\right] = [257K; 472K]$. Finally, we specify that we wish to record observables about the backbone-dbn energy, the RMSD to the native state, and dump structures to an XTC trajectory file. Apart from the RMSD observable, we do not provide the program with any information about the structure of the protein, and the simulation will, therefore, start in a random extended state.

To illustrate the framework's support for various types of biased sampling, this example uses a variant where the TORUSDBN bias is included in the sampling, and explicitly subtracted in the energy (i.e., the weight: -1 option of backbone-dbn). This means that the bias cancels out when extracting statistics at $\beta=1$, thus producing unbiased estimates at 283K. The simulation will produce a flat histogram over the expected energy range corresponding to the specified temperature range ($[\langle E \rangle_{257K}; \langle E \rangle_{472K}]$).

We ran PHAISTOS with the settings above on an eight-core 3.4 GHz Intel Xeon processor for 1 week. Figure 3 gives an overview of the results. The free energy plot in Figure 3a shows the distribution of energy versus RMSD extracted directly from the samples dumped during the simulation. As the samples were generated using a generalized ensemble
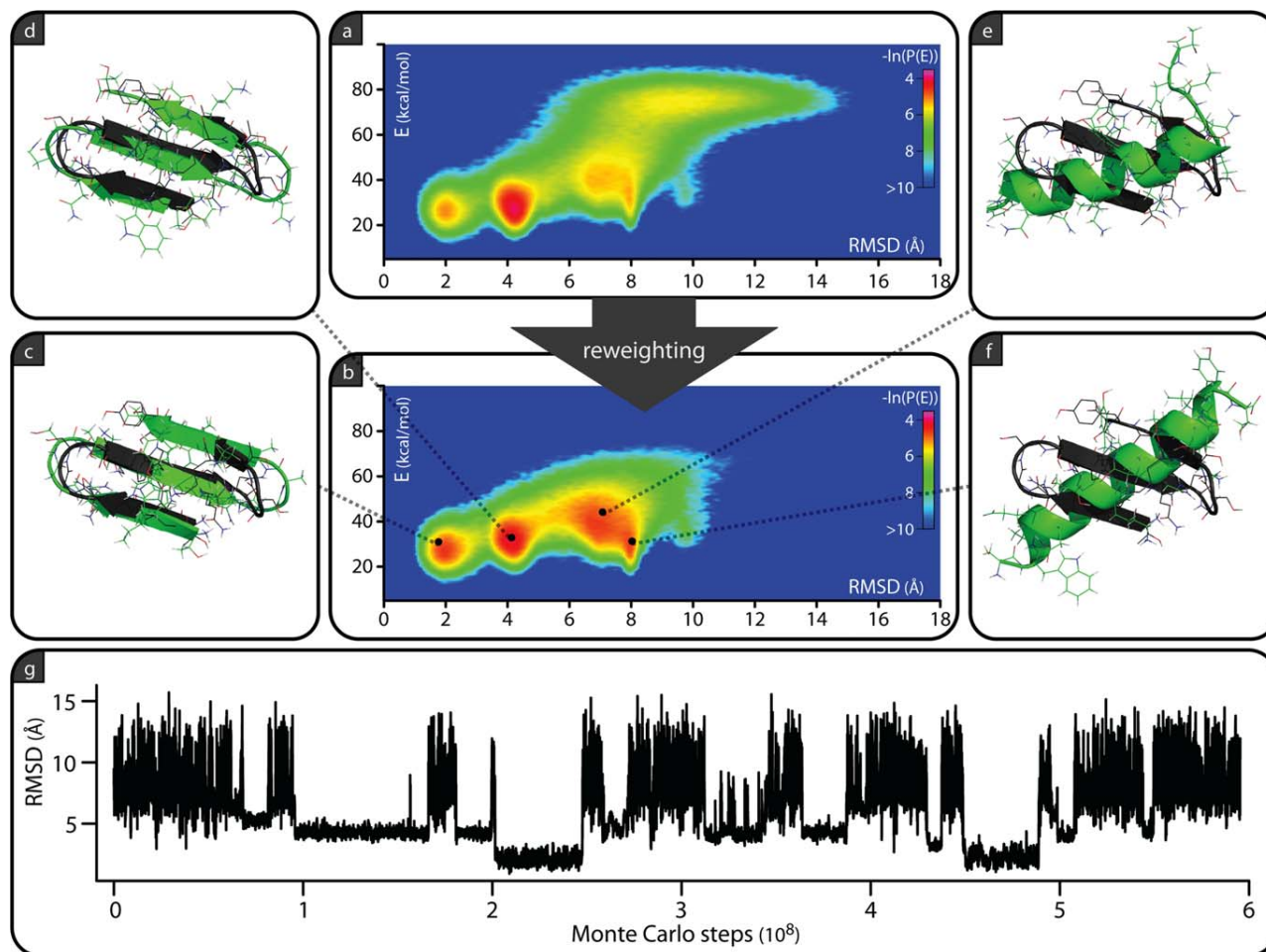
Figure 3. Illustration of a reversible folding simulation of the beta3s peptide in PHAISTOS. The simulation was conducted with the PROFASI force-field, using the MUNINN multihistogram method and a set of moves including TORUSDBN as a dihedral proposal distribution. The bias introduced by TORUSDBN is compensated for to ensure correctly distributed samples. a) Free energy plot as a function of energy and RMSD in the multicanonical (flat histogram) ensemble. b) Free energy plot as a function of energy and RMSD, reweighted to the canonical ensemble at 283 K. c–f) Representative cluster medoids found with reweighted clustering using the PLEIADES module, compared to the native structure[32] (shown in black). Figures created using Pymol.[33] g) RMSD versus time of one of the eight threads in the simulation.

technique, they must be reweighted to retrieve the statistics according to the Boltzmann distribution at the specified temperature. This is done using a script included in the MUNINN module, resulting in the plot in Figure 3b.

To find representative structures in the ensemble, we use the PLEIADES clustering module,[34] included in the framework. Again, it is important to remember that the raw data are produced in a generalize ensemble setting and must be reweighted. We ran the RMSD-based weighted k-means method implemented in the PLEIADES module to select the highlighted structures in Figure 3.

From the analysis above, we conclude that at 283 K, the protein is marginally stable in the PROFASI force-field, with native-like populations at 2 and 4 Å RMSD, but also a significant population of unstructured or helical conformations. These results are in approximate agreement with experiments, which suggest a folded population of between 13 and 31%.[32] This result is compatible with a previously published simulation of the same protein using an unbiased simulation technique.[17]

Results

To illustrate the versatility of PHAISTOS, we highlight several recently published applications of the framework.

Structure prediction and inference

An example of the applicability of PHAISTOS in the context of protein structure prediction is found in a recent study on potentials of mean force.[35] The study demonstrates how probabilistic models of local protein structure such as TORUSDBN and BASILISK can be combined with probabilistic models of nonlocal features, such as hydrogen bonding and compactness, using a simple probabilistic technique.

The framework has also been applied for inference of protein structure from small-angle X-ray scattering (SAXS) and nuclear magnetic resonance (NMR) experimental data. SAXS data contain low-resolution information on the overall shape of a protein, which can be useful for determining the relative domain positions and orientations in multidomain proteins or complexes.

This can for instance be used to infer structural models of multi-domain proteins connected by flexible linkers, given the atomic structures of the individual domains. Such calculations require efficient back-calculation of SAXS curves, which is made possible through a coarse-grained Debye method.[36,37]

NMR experimental data can provide high-resolution structural information that can improve the accuracy of a simulation. PHAISTOS contains preliminary support for sampling conditional on chemical shift data, which is known to contain substantial information on the local structure of a protein.[38,39] Furthermore, the framework was recently used for inferential structure determination using pair-wise distances obtained from NOE experiments, with TORUSDBN and BASILISK as prior distribution for the protein's backbone and side chains.[40]

Efficient clustering

Efficiently clustering a large number of protein structures is an important task in protein structure prediction and analysis. Typically, clustering programs require costly RMSD calculations for many pairs in the set of structures. PHAISTOS contains a clustering module called PLEIADES that uses a k-means clustering approach[41] to reduce the number of pair-wise RMSD distance calculations. Furthermore, PLEIADES includes support for replacing the RMSD distance computations with distances between vectors of Gauss integrals,[42] which provides dramatic computational speedups.[34]

Native ensembles

The energy landscape around the native state tends to be rugged, making it challenging to sample such states efficiently.[1] For these tasks, the CRISP backbone move is particularly well suited, given its ability to propose subtle, nondisruptive updates to the protein backbone. Monte Carlo simulations using this move were recently shown to explore conformational space with an efficiency on par with MD, outperforming the current state-of-the-art in local Monte Carlo move methods.[4]

The TYPHON module[43] rapidly explores near-native ensembles using the CRISP move in combination with a user-defined set of nonlocal restraints. Local structure is under the control of probabilistic models of the backbone (TORUSDBN) and side chains (BASILISK), while nonlocal interactions such as hydrogen bonds and disulfide bridges are heuristically imposed as Gaussian restraints. TYPHON can be seen as a "null model" of conformational fluctuations in proteins: it rapidly explores the conformational space accessible to a protein given a set of specified restraints.

Discussion

The relevance of a new software package should be assessed relative to already existing packages in the literature. We acknowledge that in our case, there are a number of such alternatives already available. We describe the most important ones here, focusing on the differences to the framework presented in this article.

Of the available Monte Carlo software packages, the ROSETTA package[11] is perhaps the most widely used and has an impressive track record for protein structure prediction and design.[44] The package focuses primarily on structure/

sequence prediction (optimization) rather than simulation, and consequently, many of the moves in ROSETTA are not compatible with the property of detailed balance.

PHAISTOS also has some overlap with the PROFASI simulation package,[45] in the sense that both implement the BGS move[9] and the PROFASI energy function.[17] The PROFASI simulation program was designed as a tool for studying protein aggregation and is thus highly optimized for many-chain simulation using their lightweight force-field and under the assumption of fixed bond angles. PHAISTOS aims to provide a greater flexibility in the choice of energies and a wider selection of moves and is not limited to a fixed bond-angle representation.

The closest alternatives to PHAISTOS are perhaps the CAMPARI software package[1] and the Monte Carlo package in CHARMM,[2] which both provide functionality for conducting MCMC simulations using various force-fields and moves. Compared with PHAISTOS, the selection of force-fields and moves differ, and the focus is different. For instance, PHAISTOS has a strong focus on sampling using probabilistic models of local structure, which is not supported by either of the two alternatives.

The current version of the PHAISTOS framework has several limitations. To a user familiar with MD software, the primary limitation will presumably be the lack of explicit solvent models in the framework. The large conformational moves that provide the sampling advantage of Monte Carlo simulations are difficult to combine with an explicit solvent representation. In line with other Monte Carlo simulation packages, PHAISTOS is, therefore, currently limited to implicit solvent simulations. Another limitation is that PHAISTOS can currently only simulate a single polypeptide at a time. This restriction will be removed in the next release of the software, which will also include implementations of several new force-fields.

As the list of applications demonstrates, even in its current form, the framework provides the necessary tools for conducting relevant MCMC simulations of protein systems. The framework incorporates generalized ensembles and novel Monte Carlo moves, including moves that incorporate structural priors as proposal distributions. These features are unique to this framework and have been shown to increase sampling efficiency considerably.

The software is freely available under the GNU General Public License v3.0. All source code is fully documented using the Doxygen system (http://www.doxygen.org), and a user manual is available for detailed descriptions on how to set up simulations. Both sources of information are accessible via the PHAISTOS web site, http://phaistos.sourceforge.net.

[1] A. Vitalis, R. V. Pappu, *Annu. Rep. Comput. Chem.* **2009**, *5*, 49.

[2] J. Hu, A. Ma, A. R. Dinner, *J. Comput. Chem.* **2006**, *27*, 203.

[3] J. P. Ulmschneider, M. B. Ulmschneider, A. Di Nola, *J. Phys. Chem. B* **2006**, *110*, 16733.

[4] S. Bottaro, W. Boomsma, K. E. Johansson, C. Andreetta, T. Hamelryck, J. Ferkinghoff-Borg, *J. Chem. Theory Comput.* **2012**, *8*, 695.

[5] M. Habeck, M. Nilges, W. Rieping, *Phys. Rev. Lett.* **2005**, *94*, 18105.

[6] M. R. Betancourt, *J. Chem. Phys.* **2005**, *123*, 174905.

[7] C. Smith, T. Kortemme, *J. Mol. Biol.* **2008**, *380*, 742.

[8] J. P. Ulmschneider, W. L. Jorgensen, *J. Chem. Phys.* **2003**, *118*, 4261.

[9] G. Favrin, A. Irbäck, F. Sjunnesson, *J. Chem. Phys.* **2001**, *114*, 8154.

[10] R. L. Dunbrack, F. E. Cohen, *Protein Sci.* **1997**, *6*, 1661.

[11] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y.-E. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popovíc, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker and P. Bradley, *Methods Enzymol.* **2011**, *487*, 545.

[12] T. Przytycka, *Proteins* **2004**, *57*, 338.

[13] T. Hamelryck, J. T. Kent, A. Krogh, *Plos Comput. Biol.* **2006**, *2*, e131.

[14] W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, T. Hamelryck, *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 8932.

[15] T. Harder, W. Boomsma, M. Paluszewski, J. Frellsen, K. E. Johansson, T. Hamelryck, *BMC Bioinform.* **2010**, *11*, 306.

[16] N. Gō, H. A. Scheraga, *Macromolecules* **1970**, *3*, 178.

[17] A. Irbäck, S. Mitternacht, S. Mohanty, *PMC Biophys.* **2009**, *2*, 2.

[18] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, W. L. Jorgensen, *J. Phys. Chem. B* **2001**, *105*, 6474.

[19] D. Qiu, P. S. Shenkin, F. P. Hollinger, W. C. Still, *J. Phys. Chem. A* **1997**, *101*, 3005.

[20] I. Lotan, F. Schwarzer, D. Halperin, J.C. Latombe, *J. Comput. Biol.* **2004**, *11*, 902.

[21] A. Irbäck, S. Mitternacht, *Proteins* **2007**, *71*, 207.

[22] D. W. Li, S. Mohanty, A. Irbäck, S. Huo, *PLoS Comput. Biol.* **2008**, *4*, e1000238.

[23] J. W. Ponder, F. M. Richards, *J. Am. Chem. Soc.* **1987**, *8*, 1016.

[24] B. Roux, T. Simonson, *Biophys. Chem.* **1999**, *78*, 1.

[25] W. C. Still, A. Tempczyk, R. C. Hawley, T. Hendrickson, *J. Am. Chem. Soc.* **1990**, *112*, 6127.

[26] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, *J. Chem. Theory Comput.* **2008**, *4*, 435.

[27] K. E. Johansson, T. Hamelryck, *Proteins*, doi: 10.1002/prot.24277.

[28] U. H. E. Hansmann, Y. Okamoto, *J. Comput. Chem.* **1997**, *18*, 920.

[29] B. A. Berg, T. Neuhaus, *Phys. Rev. Lett.* **1992**, *68*, 9.

[30] B. Hesselbo, R. B. Stinchcombe, *Phys. Rev. Lett.* **1995**, *74*, 2151.

[31] J. Ferkinghoff-Borg, *J. Eur. Phys. J. B* **2002**, *29*, 481.

[32] E. De Alba, J. Santoro, M. Rico, M. Jimenez, *Protein Sci.* **1999**, *8*, 854.

[33] Schrödinger, LLC., The PyMOL Molecular Graphics System, Version 0.99rc6.

[34] T. Harder, M. Borg, W. Boomsma, P. Røgen, T. Hamelryck, *Bioinformatics* **2012**, *28*, 510.

[35] T. Hamelryck, M. Borg, M. Paluszewski, J. Paulsen, J. Frellsen, C. Andreetta, W. Boomsma, S. Bottaro, J. Ferkinghoff-Borg, *PLoS ONE* **2010**, *5*, e13714.

[36] K. Stovgaard, C. Andreetta, J. Ferkinghoff-Borg, T. Hamelryck, *BMC Bioinform.* **2010**, *11*, 429.

[37] N. G. Sgourakis, O. F. Lange, F. DiMaio, I. André, N. C. Fitzkee, P. Rossi, G. T. Montelione, A. Bax, D. Baker, *J. Am. Chem. Soc.* **2011**, *133*, 6288.

[38] A. Cavalli, X. Salvatella, C. M. Dobson, M. Vendruscolo, *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9615.

[39] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker, A. Bax, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 4685.

[40] S. Olsson, W. Boomsma, J. Frellsen, S. Bottaro, T. Harder, J. Ferkinghoff-Borg, T. Hamelryck, *J. Magn. Reson.* **2011**, *213*, 182.

[41] S. Lloyd, *IEEE Trans. Inf. Theory* **1982**, *28*, 129.

[42] P. Røgen, B. Fain, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 119.

[43] T. Harder, M. Borg, S. Bottaro, W. Boomsma, S. Olsson, J. Ferkinghoff-Borg, T. Hamelryck, *Structure* **2012**, *20*, 1028.

[44] R. Das, D. Baker, *Annu. Rev. Biochem.* **2008**, *77*, 363.

[45] A. Irbäck, S. Mohanty, *J. Comput. Chem.* **2006**, *27*, 1548.

Protein Structure Validation and Refinement Using Amide Proton Chemical Shifts Derived from Quantum Mechanics

Anders S. Christensen[1]*, Troels E. Linnet[2], Mikael Borg[3], Wouter Boomsma[2], Kresten Lindorff-Larsen[2], Thomas Hamelryck[3], Jan H. Jensen[1]

1 Department of Chemistry, University of Copenhagen, Copenhagen, Denmark, **2** Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Copenhagen, Denmark, **3** Structural Bioinformatics Group, Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen, Denmark

Abstract

We present the ProCS method for the rapid and accurate prediction of protein backbone amide proton chemical shifts - sensitive probes of the geometry of key hydrogen bonds that determine protein structure. ProCS is parameterized against quantum mechanical (QM) calculations and reproduces high level QM results obtained for a small protein with an RMSD of 0.25 ppm ($r = 0.94$). ProCS is interfaced with the PHAISTOS protein simulation program and is used to infer statistical protein ensembles that reflect experimentally measured amide proton chemical shift values. Such chemical shift-based structural refinements, starting from high-resolution X-ray structures of Protein G, ubiquitin, and SMN Tudor Domain, result in average chemical shifts, hydrogen bond geometries, and trans-hydrogen bond ($^{h3}J_{NC'}$) spin-spin coupling constants that are in excellent agreement with experiment. We show that the structural sensitivity of the QM-based amide proton chemical shift predictions is needed to obtain this agreement. The ProCS method thus offers a powerful new tool for refining the structures of hydrogen bonding networks to high accuracy with many potential applications such as protein flexibility in ligand binding.

Introduction

Chemical shifts hold valuable structural information that is being used increasingly in the determination of protein structure and dynamics[1]. This is made possible primarily by empirical chemical shift predictors such as SHIFTS, SPARTA, SHIFTX, PROSHIFT, and CamShift [2–7]. While these methods generally offer quite accurate predictions, the predicted chemical shifts of backbone amide protons (δ_H) tend to be significantly less accurate than, for example, the proton on the α-carbon [8,9]. This is unfortunate since ^{15}N-HSQC forms a large fraction of all protein NMR studies and δ_H holds valuable information about the hydrogen bond geometry of the ubiquitous amide-amide hydrogen bonds that are key to protein secondary structure. Parker, Houk and Jensen [10] have proposed a δ_H-predictor that was shown to offer significantly more accurate predictions, although this was only demonstrated for 13 δ_H-values. The method suggests that there is an exponential dependence of δ_H in the NH··O=C bond length (as suggested by Barfield [11] and Cornilescu et al. [12]) as well as a non-negligible contribution from cooperative effects in hydrogen bonding networks. This exponential dependence makes empirical parameterizations of δ_H-predictors challenging since even small discrepancies between the structure used in the

parameterization (usually an X-ray structure without explicitly represented hydrogens) and the solution-phase structural ensemble that gives rise to the experimentally observed δ_H-values can have a significant effect. The method by Parker et al. addresses this problem by parameterization against δ_H-values obtained by quantum mechanical (QM) calculations, and is similar in spirit to the QM-based α-carbon chemical shift predictor CheShift developed by Vila et al. [13,14]. Both studies noted that the QM-based chemical shift predictors tend to be more sensitive to small structural changes compared to popular empirical chemical shift predictors and therefore promises to be valuable tools in protein structure validation and refinement. Here we present several key advances in the use of backbone amide proton chemical shifts to refine and validate the geometry of the amide-amide hydrogen bonding network in proteins. First we present and validate the ProCS method which extends the QM-based backbone amide proton chemical shift predictor proposed by Parker et al. [10]. Second we present a computational methodology for using ProCS and experimental δ_H-values to refine the hydrogen bond-geometries of proteins. This is accomplished by implementing ProCS in the Markov chain Monte Carlo (MCMC) protein simulation framework PHAISTOS [15], and using this in

combination with a molecular mechanics (MM) force field. Third, we show for a number of small proteins that structural refinement against experimental δ_H values using ProCS leads to hydrogen bond geometries that are in closer agreement with high-resolution X-ray structures and experimental trans-hydrogen bond spin-spin coupling constants ($^{h3}J_{NC'}$) compared to using an energy function based on the empirical chemical shift predictor CamShift [7] or solely using a force field (OPLS-AA/L [16] with the GB/SA continuum solvent model [17]).

Results and Discussion

The ProCS method

The ProCS program uses a modified implementation of the formula developed by Parker et al.[10] where the amide proton chemical shift is approximated by a sum of additive terms:

$$\delta_H = \delta_{BB} + \Delta\delta_{1°HB} + \Delta\delta_{2°HB} + \Delta\delta_{3°HB} + \Delta\delta_{RC} \quad (1)$$

Here, δ_{BB} is a backbone term that depends on the (ϕ,ψ) torsion angles of the residue, $\Delta\delta_{1°HB}$ is due to a primary hydrogen bond directly to the amide proton in question, $\Delta\delta_{2°HB}$ is due to a secondary hydrogen bond to the carbonyl oxygen in the amide group, $\Delta\delta_{3°HB}$ is a small term that incorporates further polarization due to hydrogen bonding at the primary and/or secondary bonding partner and $rDelta\delta_{RC}$ describes magnetic perturbations due to ring currents in nearby aromatic side chains. ProCS calculates amide proton chemical shift values referenced to dimethyl-silapentane-sulfonate (DSS).

We have replaced the original δ_{BB} term, which was a crude 3-step function, by a scaled version of the (ϕ,ψ) backbone torsion angle hypersurface parametrized by Czinki and Császár [18]. The δ_{BB} term is given as

$$\delta_{BB} = 0.828 \cdot (ICS(\phi,\psi) + 0.77 \text{ ppm}) \quad (2)$$

where $ICS(\phi,\psi)$ is the n-th order cosine series given in reference [18]. The scaling is necessary to account for differences in choice of basis set and molecular geometry optimization [19].

In the cases described by Parker et al., $\Delta\delta_{RC}$-values are obtained through the SHIFTS web-interface[3]. Since this would be impractical, we implemented the point-dipole [20,21] approximation given by:

$$\Delta\delta_{RC} = i \, B \, \frac{1 - 3\cos^2(\theta)}{|\vec{r}|^3} \quad (3)$$

where i is an intensity parameter which depends on the type of aromatic ring, B is a constant of 30.42 ppm Å^3, $\vec{r}$ is the vector between the amide proton and the center of the aromatic ring and θ is the angle between $\vec{r}$ and the normal to the plane of the aromatic ring located on its center. The values of i and B are obtained from the parameter set by Christensen et al. [22].

The following expression for $\Delta\delta_{1°HB}$ was implemented for primary bonds to backbone amide carbonyl oxygen atoms:

$$\Delta\delta_{1°HB} = [4.81\cos^2(\theta) + \sin^2(\theta)\{3.10\cos^2(\rho)$$

$$-0.84\cos(\rho) + 1.75\}]e^{-2.0 \text{ Å}^{-1}(r_{OH} - 212.760 \, 2Å)} \cdot 1 \text{ ppm} \quad (4)$$

This formula originates from the works of Barfield[11] and is fitted to chemical shifts computed for model systems of hydrogen bonding between two formamide molecules. In order to treat hydrogen bonding to other oxygen atom types (carboxylic acids and alcohols as found in side chains and C-terminal), we carried out similar scans (see Section S2 and Fig. S4 in Supporting Information S1) over bond angles and lengths and stored these in lookup-tables from which the chemical shift perturbation due to any hydrogen bonding geometry can be interpolated. Hydrogen bonding to carboxylic acid oxygen atoms interaction were modeled by N-methylacetamide/acetate dimers, while bonds to alcohols oxygen atoms were modeled by N-methylacetamide/methanol dimers.

For non-hydrogen bonding amide protons, which are found primarily on the protein surface, $\Delta\delta_{1°HB}$ is approximated as the interaction between a water molecule and an N-methylacetamide molecule. In this case, $\Delta\delta_{1°HB}$ is equal to 2.07 ppm for an energy minimized bonding geometry (see Section S3 and Fig. S5 in Supporting Information S1). The functional forms of $\Delta\delta_{2°HB}$ and $\Delta\delta_{3°HB}$ were kept as described in reference [10].

Reproducing QM chemical shifts

ProCS predictions result from several terms [Eq. 1] that are assumed to be additive. To test this additivity assumption we use density functional theory (DFT) and compute chemical shielding values (at the B3LYP/cc-pVTZ/PCM level) for the crystal structure of human parathyroid hormone, residues 1–34 at 0.9 Å resolution, PDB-code 1ET1 [23]. Chemical shift values for amide protons at the termini are excluded from the statistics presented in this section, since they do not participate in any hydrogen bonds in the crystal structure. Using the linear scaling method due to Jain et al. [24] similar DFT calculations reproduce experimental proton chemical shifts of a test set of 80 small to medium sized molecules to an RMSD of 0.13 ppm. [24]

ProCS reproduces the QM calculation with an RMSD of 0.25 ppm (Table 1) based on the same structure. ProCS is parameterized based on a number of DFT calculations (see Methods section) which have been shown to yield proton chemical shifts within 0.16 ppm of experimental values for small organic molecules [19]. Thus, the error from non-additivity is roughly the same as the expected deviation from experiment.

The chemical shifts predicted by empirical methods do not agree well with the DFT results, with RMSD values ranging from 0.56 to 0.70 ppm (see Table 1 and Fig. 1). The DFT chemical shifts span a relatively large range (5.8–9.3 ppm) while the empirically predicted chemical shifts span a very narrow range (up to 6.9–8.9 ppm for SPARTA+) - see Fig. 1. This indicates that the empirical methods are less sensitive to small differences in hydrogen bond geometry found in the X-ray structure.

Reproducing experimental chemical shifts from X-ray structures

The QM method used here reproduces small molecule ^{1}H chemical shifts with an RMSD of 0.13 ppm [24]. The RMSD between the chemical shifts calculated by QM using the static X-Ray structure and the experimental data obtained in solution is 0.66 ppm. The main sources of this discrepancy are likely inaccuracies in the hydrogen bond lengths in the X-ray structure compared to solution, since there is an exponential dependence of the proton chemical shifts on this distance [Eq. 4], and/or the use of a single structure rather than a structural ensemble.

The corresponding RMSD to experimental data for ProCS (0.63 ppm) is similar to the QM RMSD and significantly larger than the 0.25 ppm RMSD between QM and ProCS, indicating

Table 1. Correlation coefficients and RMSD between five chemical shift predictors, chemical shifts derived from quantum mechanics (B3LYP/cc-pVTZ/PCM) chemical shifts and experimental values.

Data source[a]	Exp'tl	Exp'tl	QM	QM
	r	RMSD	r	RMSD
ProCS	0.54	0.63	0.94	0.25
SHIFTS[2]	0.64	0.37	0.59	0.70
SHIFTX[5]	0.69	0.37	0.71	0.62
SPARTA+[40]	0.69	0.42	0.68	0.56
CamShift[7]	0.64	0.32	0.59	0.66

[a]The crystal structure of human parathyroid hormone, residues 1–34 at 0.9 Å resolution (PDB-code 1ET1[23]) is used as input structure in all chemical shift calculations.
doi:10.1371/journal.pone.0084123.t001

Table 2. Reproduction of experimental amide proton chemical shift values based on 13 X-ray structures with a crystallographic resolution of 1.35 Å or less.

Method	$\langle r \rangle$[a]	$\langle RMSD \rangle$[b]
ProCS	0.58	1.13 ppm
SHIFTS[2]	0.56	0.64 ppm
SHIFTX[5]	0.71[c]	0.51 ppmc
SPARTA+[40]	0.79	0.40 ppm
CamShift[7]	0.74	0.46 ppm

$\langle r \rangle$">denotes the average correlation coefficient over the 13 structure.
$\langle$">RMSD$\rangle$ denotes the average root mean square deviation over the 13 structure.
[c]For SHIFTX, three structures displayed over fitting behavior with $r \approx 0.99$. These structures are excluded from the average values.
doi:10.1371/journal.pone.0084123.t002

that ProCS is sufficiently accurate to identify inaccuracies in the X-ray structure, and/or the effect of using a single structure rather than a structural ensemble. A similar comparison to experiment for 13 other proteins is given in Table 2 (PDB-codes: 1BRF, 1CEX, 1CY5, 1ET1, 1I27, 1IFC, 1IGD, 1OGW, 1PLC, 1RGE, 1RUV, 3LZT, 5PTI). The deviation from experiment for the empirical methods are significantly smaller than for ProCS with RMSD values ranging from 0.46 to 0.64 ppm (Table 2). A likely explanation for this is that the empirical methods are parameterized using X-ray structures. In order for these methods to produce low RMSD values relative to experiment they need to be insensitive to errors in protein structure.

Refining protein structures based on chemical shifts

If indeed the difference in experimental and computed chemical shifts reports on inaccuracies in the protein structure, then minimizing this difference can be used for structural refinement. To test this hypothesis we generate structural ensembles that minimizes the difference in computed and observed chemical shifts to the specified uncertainty in the chemical shift model and determine the quality of these structures by comparison to experimental structures and coupling constants (next section).

Refinement is accomplished using a Markov chain Monte Carlo (MCMC) technique described in detail in the Methods section. In short, the method involves Monte Carlo sampling of structural changes using a posterior distribution constructed using the OPLS-AA/L fore field [16] with the GB/SA implicit solvent model [17] (referred to hereafter simply as "OPLS") and amide

proton chemical shifts differences from experiment computed using either CamShift or ProCS. We note that the resulting ensemble is not a dynamic ensemble but an ensemble that reflects experimentally measured amide proton chemical shifts. The simulation lengths are roughly equivalent to 6–10 ns of molecular dynamics simulations [25]. We refine the structure of ubiquitin, Protein G, and SMN Tudor domain each based on three energy functions: OPLS alone, OPLS+ProCS and OPLS+CamShift. Each MC refinement results in an ensemble of 24,000 structural samples for Ubiquitin and 40,000 for Protein G and SMN Tudor Domain, from which average chemical shifts for each amide proton are computed. The results are summarized in Table 3.

The average ProCS chemical shifts are in better agreement with experiment (RMSD 0.81 ppm) compared to using X-ray structures (RMSD 1.10 ppm). The respective RMSD values for amide protons hydrogen bonded to backbone amide groups, other hydrogen bonds, and no hydrogens bonds are 0.31 ppm, 0.78 ppm and 1.09, respectively. These RMSD values reflect the uncertainties defined for each kind of hydrogen bonding situation in the ProCS model (see Methods section) meaning that the simulations have indeed converged to a distribution of structures reflecting the experimental chemical shifts within the accuracy of the ProCS model at the given temperature. A corresponding structural ensemble generated solely from the OPLS force field increases the RMSD from experiment to 1.52 ppm, indicating more inaccurate hydrogen bond geometries (more on this in the next section).

An MC-based structural refinement based on OPLS and chemical shifts derived from CamShift has no substantial effect
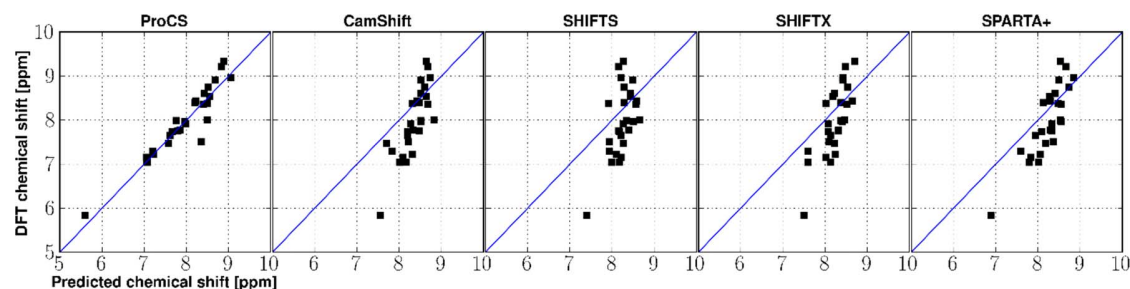
Figure 1. Correlation between chemical shift predictions from five different NMR prediction methods and quantum mechanical chemical shifts for human parathyroid hormone, residues 1–37 (PDB code: 1ET1). Blue lines represent a 1-to-1 correlation.
doi:10.1371/journal.pone.0084123.g001

Table 3. Statistics for three different types of protein simulations.

Structures[a]	ProCS ^{1}H RMSD	CamShift ^{1}H RMSD	⟨Bond length deviation⟩[b]	[h3]$J_{NC'}$ RMSD
Ubiquitin Ensembles: CamShift + OPLS	0.79 ppm	-	0.03 Å	0.17 Hz
Ubiquitin Ensembles: CamShift + OPLS	-	0.50 ppm	0.37 Å	0.17 Hz
Ubiquitin Ensembles: OPLS (no chemical shifts)	1.56 ppm	0.60 ppm	0.41 Å	0.18 Hz
1UBQ X-ray starting structure	1.22 ppm	0.51 ppm	-	0.22 Hz
SMN Tudor Domain Ensembles: ProCS + OPLS	0.93 ppm	-	0.09 Å	0.24 Hz
SMN Tudor Domain Ensembles: CamShift + OPLS	-	0.46 ppm	0.17 Å	0.23 Hz
SMN Tudor Domain Ensembles: OPLS (no chemical shifts)	1.47 ppm	0.61 ppm	0.22 Å	0.23 Hz
1MHN X-ray starting structure	1.09 ppm	0.65 ppm	-	0.24 Hz
Protein G Ensembles: ProCS + OPLS	0.69 ppm	-	0.06 Å	0.14 Hz
Protein G Ensembles: CamShift + OPLS	-	0.52 ppm	0.38 Å	0.18 Hz
Protein G Ensembles: OPLS (no chemical shifts)	1.54 ppm	0.68 ppm	0.37 Å	0.20 Hz
1PGB X-ray starting structure	1.21 ppm	0.55 ppm	-	0.17 Hz

[a]The ensembles are obtained from MCMC simulations using either OPLS-AA/L with the GB/SA solvent model (OPLS) force field energy or OPLS energy plus a chemical shift energy term from from either ProCS or CamShift. Values are calculated over four runs on each of three protein structures, Ubiquitin, Protein G and SMN Tudor Domain, or their static X-ray structure.
[b]The mean bond length deviation denotes the mean absolute difference between the mean hydrogen bond length observed in the sampled structures to the mean hydrogen bond length observed in the corresponding X-ray structure noted below.
doi:10.1371/journal.pone.0084123.t003

on the chemical shift RMSD compared to the X-ray structure (0.50 vs 0.46 ppm). Using the OPLS-derived structural ensemble increases the RMSD by 0.1 ppm compared to using X-ray structures when CamShift is used to calculate chemical shifts. This indicates that an OPLS-based refinement does not improve the hydrogen bonding geometry and that CamShift is less sensitive to a change in structure compared to ProCS.

Hydrogen bond geometries

The H··O distances and H··O = C angles of the backbone amide-amide hydrogen bonds for which [h3]$J_{NC'}$ coupling constants have been measured (see next section) are extracted from the ensembles and compared to the corresponding values found in the experimental X-ray structures with hydrogens added from PDB2PQR [26,27]. The result are shown in Table 3 and Figures 2 and 3.

Fig. 2 shows the distributions of H··O distances from the ensembles computed using the three energy terms described in the previous section. Structural refinement using OPLS and ProCS for ubiquitin results in ensembles with average H··O distances that have an RMSD within 0.02 Å of those found in the X-ray structures 1UBQ and 1UBI (both 1.80 Å X-ray resolution) and 0.04 Å from the ubiquitin structure 1OGW (1.30 Å X-ray resolution) in which the leucine residues 50 and 67 have been replaced by fluoro leucine. For Protein G we note that the resulting ensemble does not have an average H··O distance that agrees well (0.07 Å difference) with the starting structure 1PGB (1.92 Å X-ray resolution). However the difference from the 1PGA structure (2.07 Å X-ray resolution) and the more accurate 1IGD structure (X-ray resolution of 1.1 Å) is much less, 0.02 Å and 0.00 Å, respectively. The 1IGD structure is a close homologue which has 89% sequence identity score and 95% sequence similarity. In the case of the SMN Tudor Domain, ProCS-based refinement results in slightly longer amide-amide hydrogen bond lengths (0.02 Å on average) compared to the X-ray structure 1MHN.

In contrast, structural refinement using CamShift and OPLS or just OPLS leads to increases in average H··O bond lengths of up to 0.15 Å, with a standard deviation 2–3 times larger than that found in the OPLS+ProCS simulation. In all cases use of CamShift has relatively little effect on the ensemble average H··O distance compared to just using OPLS.

In all cases, the use of ProCS leads to a significantly smaller standard deviation in H··O bond lengths: 0.017 Å compared to 0.045 and 0.041 Å for CamShift+OPLS and OPLS, respectively (Fig. 3A). The H··O = C bond angles observed in the ProCS+OPLS simulations are on average within −2.0° of corresponding value observed in the X-ray structures. The same bond angle differences are −6.7° and −7.4° observed in the CamShift+OPLS and OPLS simulations, respectively (Fig. 3B).

Trans-hydrogen bond coupling constants

Better agreement with X-ray structures does not necessarily imply better solution-phase structures. In order to compare the resulting ensembles to solution-phase data we compute average trans-hydrogen bond coupling constants and compare these to experimental values. Experimental trans-hydrogen bond [h3]$J_{NC'}$ spin-spin coupling constants represent a very sensitive measure for solution-phase hydrogen bonding conformations and are known to correlate with amide proton chemical shifts [28]. The coupling constants depend exponentially on the hydrogen bonding distance and on bond angles [11]. Data from ensemble back-calculated [h3]$J_{NC'}$ spin-spin coupling constants are summarized in Fig. 4 and Table 3.

In the ubiquitin simulations, the OPLS force field on its own does not yield ensemble [h3]$J_{NC'}$ averages in good agreement with experimental data. In this simulation, several hydrogen bonds were eventually broken. Calculated [h3]$J_{NC'}$-values for these partly unfolded hydrogen bonds show up close to 0 Hz (see Fig. 4A). The RMSD to experimental values is here 0.18 Hz. Adding the energy term from amide proton chemical shifts via CamShift does not help keeping these hydrogen bonds fixed, but results in a minor improvement in RMSD to 0.17 Hz. Adding the amide proton

chemical shifts energy term via ProCS to the OPLS force field stabilized the hydrogen bonds and also gave an improvement in the RMSD values to 0.14 Hz, which is close to that of the most accurate structural NMR ensembles of ubiquitin (see Table 4). For Protein G we obtained similar RMSD values: 0.20 Hz, 0.14 Hz and 0.18 Hz for the OPLS alone, OPLS+ProCS and the OPLS+CamShift simulations, respectively. In the SMN Tudor Domain simulation, the average $^{h3}J_{NC'}$ value of all three types of simulations were comparably close to experimental values 0.24, 0.24 and 0.23 Hz for OPLS alone, OPLS+ProCS and the OPLS+CamShift simulations, respectively. Thus, overall the coupling constants based on the ProCS refined ensembles are indeed in better agreement with experimental values indicating the refinement led to improved hydrogen bond geometries compared to using OPLS or OPLS+CamShift.

Impact on Q-factor

In this section we investigate how amide proton chemical shifts restraints affect back-calculated $^{1}D_{NH}$ residual dipolar couplings (RDCs) compared to experimental values for ubiquitin. RDCs are attractive in this regard since they report on structural features that are not related to hydrogen bonding conformations as studied intensively in the previous sections. The Q-factor is a qualitative measure for the agreement between back-calculated RDCs and the corresponding experimentally observed values [29].

We find, that for our Ubiquitin ensemble generated using the OPLS force field alone has a Q-factor of 0.29 while inclusion of chemical shifts only gives a very modest improvement of this figure to 0.27 for both CamShift and ProCS as chemical shift model. The same value calculated for the three X-ray structures 1UBQ, 1UBI and 1OGW are 0.22, 0.25 and 0.26, respectively. For six NMR-based ensembles the Q-factor is in the range 0.04–0.38, though in some cases the ensembles were refined against the RDCs (see Table 4). We observe no significant correlation ($P < 0.05$) between RMSDs for predicted chemical shifts or spin-spin couplings constant to their experimental values and the calculated Q-factor for the 12 cases presented in Table 4.

While amide proton chemical shifts have some dependence on the dihedral angles of the backbone, the dependence on the particular hydrogen bonding conformations is much larger in

comparison. This is due to an exponential dependence on the hydrogen bond length.

The distribution from which we sample chemical shifts is constructed from a prior distribution based on the OPLS force field and a likelihood which contains information from experimental chemical shifts. We expect that structural features of the resulting ensemble, which are not local to the hydrogen bond geometry, will largely reflect the prior distribution, i.e. in this our case, the OPLS force field.

Computational efficiency

Executing the simulations on one core of a Intel Xeon X5560 running at 2.80 GHz with the 1UBQ structure, the average evaluation time of the three different energy-terms were OPLS-AA/L: 27 ms, CamShift 1.35: 4.7 ms, ProCS: 0.74 ms. Similar evaluation times were observed for the 1MHN and 1PGB simulations. Note that, in our implementation, the CamShift term calculates chemical shifts for six atoms per residue, even if those chemical shifts are not a used to evaluate the corresponding energy term. The OPLS and CamShift terms were implemented with a caching algorithm, so only the subset of parts of the chemical shift terms that change after a local Monte Carlo move were recomputed. This approach was not implemented for ProCS since the OPLS force field energy evaluation is by far the most computationally expensive step. Running on four cores, we obtained between 10 to 16 mio Monte Carlo iteration steps total *per* day, depending on the protein size and combination of energy terms.

Methods

### Monte Carlo refinement of protein structure

We employ Markov chain Monte Carlo sampling from a Bayesian posterior distribution to perform protein structure refinements and simulations. MCMC simulations are attractive because no gradient expressions need to be derived for ProCS. Bayesian inference[30] provides a rigorous mathematical framework for the inference of protein structure from experimental data. It involves the construction of a posterior distribution, which consists of a prior distribution and a likelihood. The former brings in general information on protein structure, and in our case is
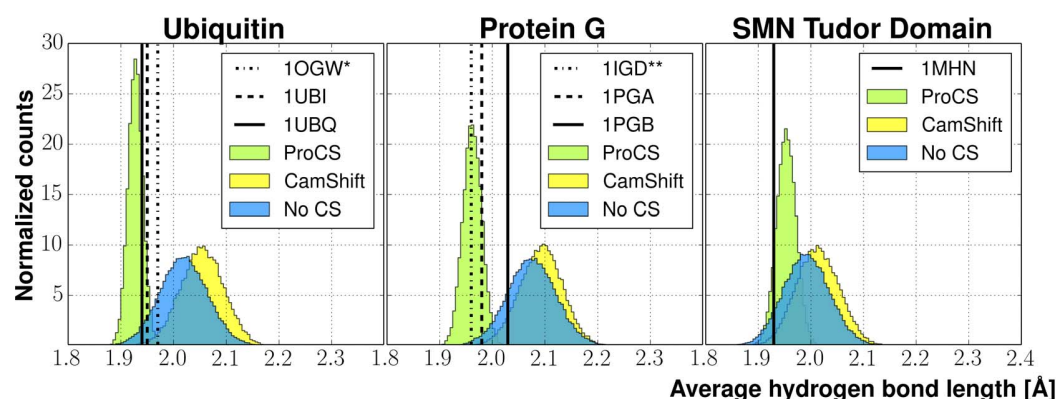
Figure 2. Distribution of average hydrogen bond lengths throughout Monte Carlo simulations on Ubiquitin, Protein G and SMN Tudor Domain. Histograms are normalized (to an area of 1) to fit identical axes. Vertical lines indicate average values obtained from experimental X-ray structures (PDB-codes are noted in the figure legends). The blue histogram represents the simulation with only the molecular mechanics energy from the OPLS-AA/L force field with the GB/SA solvent model (but no chemical shift energy term). Green and yellow histograms indicate the use of OPLS force field plus an additional chemical shift energy term from ProCS or CamShift, respectively. *1OGW contains fluoro leucine at residues 50 and 67. **1IGD is a closely related homologue (see text).
doi:10.1371/journal.pone.0084123.g002
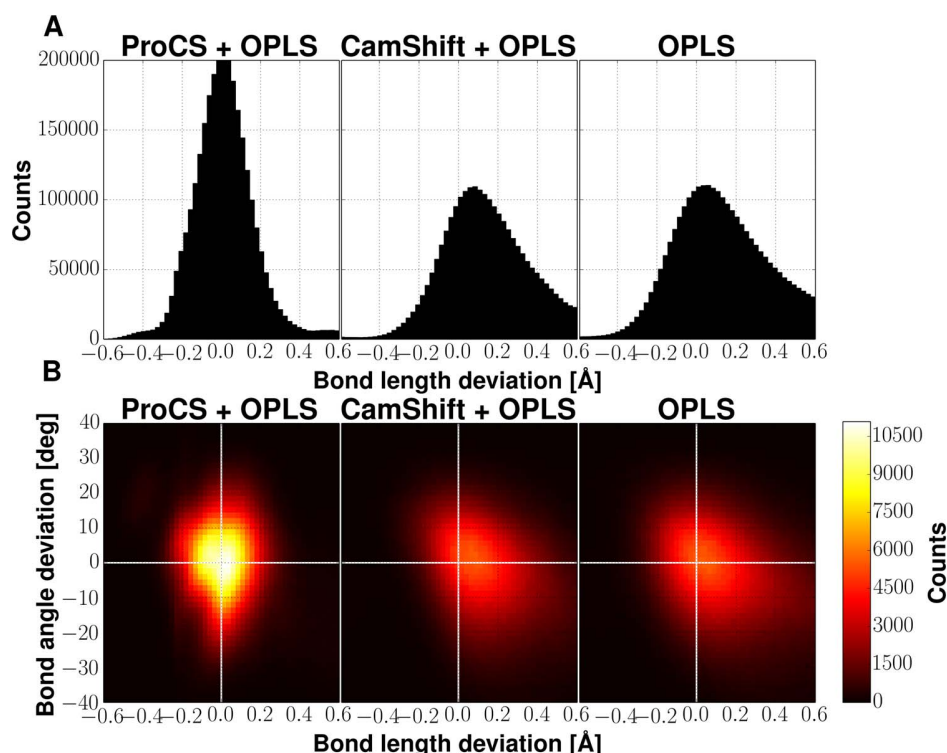
Figure 3. Deviation in hydrogen bonding geometries between the experimental X-ray structure and samples obtained from Markov Chain Monte Carlo (MCMC) simulations using the OPLS-AA/L force field with the GB/SA solvent model with either no chemical shift energy term or a chemical shift energy from either ProCS or CamShift. Data is calculated over all amide-amide bonding pairs for which experimental $^{h3}J_{NC'}$ spin-spin coupling constants were present. (A) shows the distribution of the deviations found in the MCMC ensembles from the experimental hydrogen bond length found in the X-ray structure. (B) shows the correlation of deviations in hydrogen bond lengths and H··O=C bond angles from the experimental X-ray structures.
doi:10.1371/journal.pone.0084123.g003

based on the OPLS energy function. The latter brings in the experimental data, and is based on the difference between the back-calculated data from a simulated structure and the experimental data. Using PHAISTOS, we draw samples from the joint probability distribution, which is given by:

$$p\left(X|\{\delta_i^{\exp}\},I\right) \propto p\left(\{\delta_i^{\exp}\}|X,I\right)p(X|I) \quad (5)$$

where X represents a protein structure, $\{\delta_i^{\exp}\}$ is experimental chemical shift data and I denotes prior information, such as sequence and knowledge about the uncertainties in the prediction model. The prior distribution $p(X|I)$ is proportional to $\exp(-\beta E_{FF})$, where E_{FF} is the molecular mechanics force field potential energy and $\beta = 1/k_B T$. $p\left(\{\delta_i^{\exp}\}|X,I\right)$ denotes the probability of observing experimental data given a trial structure. Under the assumption that the error in the chemical shift prediction model follows a Gaussian distribution with some set of standard deviations $\{\sigma_i\}$, the expression for $p\left(\{\delta_i^{\exp}\}|X,I\right)$ is:

$$\text{p2}(\{\delta_i^{2exp}\}|X, \{\sigma_i\}) = \Pi_{i=1}^n\left[\sqrt{\frac{1}{2\pi\sigma_i^2}}\exp\left\{-\frac{(\Delta\delta_i)^2}{2\sigma_i^2}\right\}\right] \quad (6)$$

where $\Delta\delta_i$ is the discrepancy between predicted and experimental data for the i-th nucleus of the data set in the trial structure, X. This formulation of the posterior distribution assumes that the prior distribution on X is also a good prior distribution for the

chemical shift differences, $\Delta\delta_i$, otherwise an additional term would be required[31]. The set of standard deviations, $\{\sigma_i\}$ was assigned based on the primary bond type, since, for instance, the model for solvent exposed amide protons is much cruder than the amide-amide bonding model. σ_i was set to 0.3 ppm, for primary bonds to another backbone amide, 0.5 ppm to a side chain amide group, 0.8 ppm to a side chain alcohol or carboxylic acid group and 1.2 ppm for solvent exposed amide protons and other types of bond not included in the prediction model.

Protein Structures and NMR data

All protein structures used in this study were downloaded from the RCSB Protein Data Bank[32] (PDB) and protonated using PDB2PQR 1.5, [26,27] with PROPKA[33] to determine protonation states at the pH at which NMR data was recorded. Chemical shift data were obtained from the RefDB[34] or the Biological Magnetic Resonance Bank[35], and subsequently re-referenced through Shiftcor[34]. $^{h3}J_{NC'}$ spin-spin coupling constants for 1PGB, 1UBQ and 1MHN were obtained from references [28], [12] and [36], respectively.

MCMC simulations

MCMC simulations were carried out in PHAISTOS v1.0-rc1 (rev. 335) using the Metropolis-Hastings algorithm at 300 K. The simulations are initialized from the experimental crystal structures. Four independent trajectories were simulated for each protein structure. A total of 100 mio MC steps were taken for each trajectory for Protein G and the SMN Tudor Domain simulation
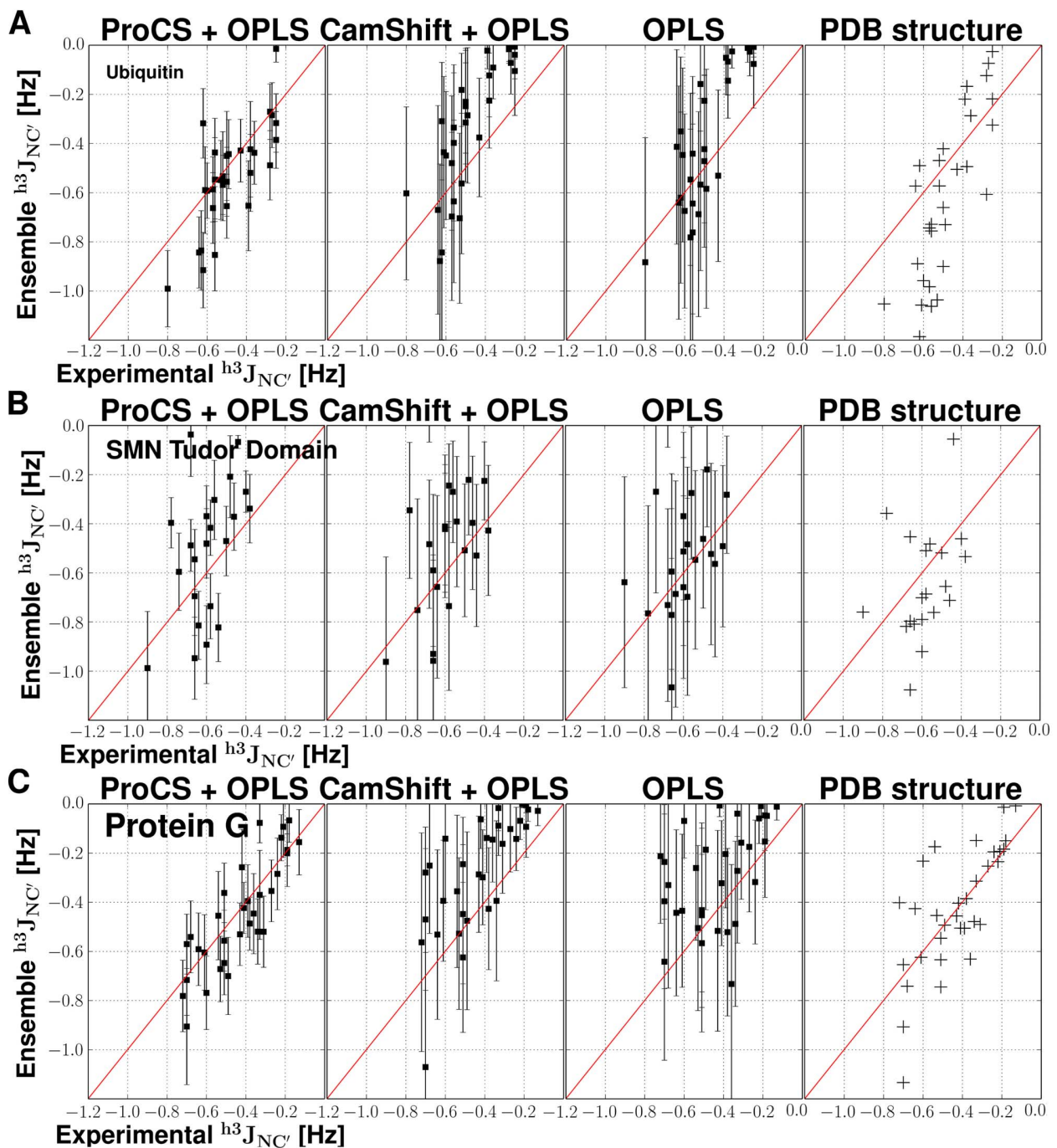
Figure 4. Reproducing experimental $^{h3}J_{NC'}$ spin-spin coupling constants via different structural ensembles and experimental X-ray structures. Squares denote the average coupling constant observed for that hydrogen bond in the ensemble and error bars represent the standard deviation observed throughout the simulations. Crosses represent the spin-spin coupling constants calculated using the static experimental X-ray structure. Results from simulations on ubiquitin is displayed in A, SMN Tudor domain in B and Protein G in C. Left column displays simulations only the OPLS-AA/L force field with the GB/SA solvent model (OPLS) and the ProCS energy term; second column is from OPLS plus the CamShift energy term; thrid column is for the simulation with only the OPLS force field energy. In the rightmost column $^{h3}J_{NC'}$ are computed from the corresponding X-ray structure.
doi:10.1371/journal.pone.0084123.g004

and 85 mio MC steps for the Ubiquitin simulation. Structures were saved every 10,000 Monte Carlo step. The Monte Carlo move-set was composed of 25% CRISP backbone moves[25] and 75% uniform side chain moves. The force field energy was

calculated using the OPLS-AA/L force field [16] with the GB/SA continuum solvent model [17]. The following crystal structures obtained from the PDB were used as starting structures in the simulations: 1PGB (Protein G), 1UBQ (Ubiquitin) and 1MHN

Table 4. Statistics for selected ubiquitin ensembles and X-ray structures.[a]

PDB-ID	(CamShift) ^{1}H RMSD	(CamShift) r	(ProCS) ^{1}H RMSD	(ProCS) r	$^{h3}J_{NC'}$ RMSD	$^{h3}J_{NC'}$ Q-factor
[b]2KOX	0.29	0.84	0.68	0.86	0.12	0.04
[c]2K39	0.34	0.82	0.98	0.77	0.13	0.07
[d]2KN5	0.23	0.91	0.71	0.82	0.12	0.22
[e]2NR2	0.44	0.74	1.35	0.64	0.14	0.25
[f]1XQQ	0.38	0.81	0.92	0.77	0.14	0.38
[g]1D3Z	0.41	0.79	1.00	0.71	0.30	0.06
[h]1UBQ	0.40	0.77	0.92	0.72	0.22	0.22
[i]1UBI	0.40	0.77	0.97	0.73	0.33	0.25
[j]1OGW	0.36	0.73	0.84	0.73	0.17	0.26
[k]OPLS + ProCS	0.32	0.79	0.17	0.98	0.14	0.27
[k]OPLS + CamShift	0.32	0.90	1.15	0.86	0.17	0.27
[k]OPLS	0.48	0.78	1.11	0.78	0.18	0.29

[a]Chemical shifts RMSD and r values are calculated for the residues for which $^{h3}J_{NC'}$ spin-spin coupling constants have been measured. [12]
[b]ERNST method/CHARMM27 + NOE + RDC [41]
[c]OPLS/AA-L + NOE + RDC [42]
[d]Backrub method/Rosetta all-atom energy + RDC [42]
[e]MUMO method/CHARMM22 + NOE + RDC [43]
[f]DER method/CHARMM22 + NOE + S^2 [44]
[g]NOE + RDC [45]
[h]X-ray 1.80 Å structure [46]
[i]X-ray 1.80 Å structure [47]
[j]X-ray 1.32 Å structure (synthetic protein with fluoro-LEU at residues 50 and 67) [48]
[k]The methods presented here
doi:10.1371/journal.pone.0084123.t004

(SMN Tudor Domain). Time evolution of Monte Carlo energy and chemical shift RMSDs are available in the Supplementary Information (Section S1, Figures S1–S3 of Supporting Information S1).

Back calculation of spin-spin coupling constants

$^{h3}J_{NC'}$ spin-spin coupling constants were calculated using the approximation by Barfield[11].

$$^{h3}J_{NC'}(\theta, \rho, \gamma_{OH}) = [-1.31 \cos^2(\theta) + \{0.62 \cos^2(\rho) + 0.92 \cos(\rho) + 0.14\} \sin^2(\theta)] \, e^{-3.2A} - 1(r_{2OH} - 1.760A).1 \; Hz \quad (7)$$

Here, the coupling depend on the $\angle$ N-H··O = C angle, ρ, $\angle$ H··O = C, θ, and the hydrogen bonding distance, r_{OH}. From the MCMC ensembles, the mean $^{h3}J_{NC'}$ spin-spin coupling constant was calculated via Eqn. 7 and the standard deviation was calculated as the root mean square deviation from the mean. The $^{h3}J_{NC'}$ RMSD to experiment is then given as

$$^{h3}J_{NC'} RMSD = \sqrt{\frac{\sum_i \left(^{h3}J_{NC'}^{exp,i} - \langle ^{h3}J_{NC'}^{calc,i} \rangle \right)^2}{N}} \quad (8)$$

where $\langle ^{h3}J_{NC'}^{calc,i} \rangle$ is the average value over the ensemble for the i'th coupling constant.

QM NMR calculations

All density functional theory (DFT) calculations of NMR isotropic shielding constants involved in the parametrization of ProCS were carried out in Gaussian 03[37]. Data was obtained at the GIAO/B3LYP/6-311++G(d,p)//B3LYP/6-31+G(d) level of theory using the scaling technique by Rablen *et al.* [19].

The NMR calculation on the 1ET1 protein structure was carried out at the B3LYP/cc-pVTZ/PCM level of theory with a water-like dielectric constant of 78.3553. In this case shielding constants were converted to chemical shifts using the scaling factor obtained by Jain *et al.* [24], assuming that the value of the dielectric constant has a negligible contribution to the scaling factors.

Calculation of ubiquitin Residual Dipolar Couplings

Residual dipolar couplings were back-calculated from the structural ensembles using singular value decomposition to fit the alignment tensor [38]. Ensemble averaging was taken into account so that all structures simultaneously were fitted to a single alignment tensor [39]. The agreement to experimental values was calculated via the Q-factor: [29]

$$Q = \frac{\sqrt{\Sigma(RDC^{exp} - RDC^{calc})^2}}{\sqrt{\Sigma(RDC^{calc})^2}} \quad (9)$$

Conclusions

ProCS is a QM-based backbone amide proton chemical shift (δ_H) predictor that can deliver QM quality chemical shift predictions for a protein structure in a millisecond. δ_H-values predicted using X-ray structures are in worse agreement with experiment, compared to those of the popular empirical chemical shift-predictors CamShift, SHIFTS, SHIFTX, and SPARTA+.

However the agreement with experiment can be significantly improved by refining the protein structures using an energy function that includes a force field and a solvation term (OPLS-AA/L with the GB/SA continuum solvent model) and a chemical shift term in the program PHAISTOS. This refinement also results in structures with predicted trans-hydrogen bond coupling constants ($^{h3}J_{NC'}$) in good agreement with experiment indicating that the refined protein structures reflect the structures in solution. Comparison of average hydrogen bond geometries to those of high-resolution (<1.35 Å) X-ray structures reveals that the structural refinement improves the predicted δ_H-values through relatively small changes in the hydrogen bond geometry distribution.

Structural refinement without chemical shifts (i.e. using only the OPLS-AA/L + Generalized Born solvation energy) or combined with CamShift has relatively little effect on the predicted δ_H-values, while the predicted $^{h3}J_{NC'}$ values are in slightly worse agreement with experiment compared to using X-ray structures or ProCS-refined structures. This is not surprising given the fact that CamShift and similar empirical methods were designed to be insensitive to relatively small changes in protein structure in order to offer robust chemical shift predictions based on X-ray structures of varying accuracy. Structural refinement based on other empirical shift predictors, such as SHIFTS, SHIFTX, and SPARTA+, were not tested mainly because an efficient interface to PHAISTOS requires a complete re-implementation of the method. However, based on our comparison to the QM-calculations (Table 1 and Fig. 1) we do not think the conclusions will be substantially different. Our data, and that of Vila *et al.* [14], suggests that QM-derived chemical shift predictors are sufficiently accurate to extract small changes in structure and dynamics from experimentally measured protein chemical shifts.

We are currently working on implementing a QM-based chemical shift prediction method for the remaining H, C, and N nuclei in a protein in ProCS (unfortunately, the source code of the CheShift method developed by Vila *et al.* for QM-based C chemical shift prediction is not available). The resulting ProCS/PHAISTOS interface should provide a powerful tool for chemical shift-based protein structure refinement.

The ensembles resulting from the simulations can be downloaded from DOI: http://dx.doi.org/10.5879/BILS/p000001

Implementations of ProCS and CamShift can be downloaded as separate modules for PHAISTOS under the terms of the GNU General Public License v3 from: http://github.com/jensengroup/

Supporting Information

Supporting Information S1   Section S1: Time evolution of energies and chemical shift RMSDs during MCMC simulation. Figures S1–S3: Details of Monte Carlo energies and chemical shift RMSDs over time for the presented simulations. **Section S2: Parametrization of chemical shift contributions due to hydrogen bonding interactions to carboxylic acids and alcohols.** Figure S4: Sketches showing the geometric parameters and the systems used in the modeling of chemical shift contributions due to hydrogen bonding. **Section S3: Model for solvent exposed amide protons.** Table S1: Chemical shift contributions due to hydrogen bonding to water molecules. Figure S5: Local minima of NMA-water dimer. (PDF)

Author Contributions

Conceived and designed the experiments: ASC KLL TH JHJ. Performed the experiments: ASC TEL MB WB. Analyzed the data: ASC TEL JHJ. Wrote the paper: ASC JHJ.

References

1. Mulder FAA, Filatov M (2010) Ab initio NMR chemical shift data and shielding calculations: Emerging tools for protein structure determination. Chem Soc Rev 395: 578–590.
2. Moon S, Case DA (2001) A new model for chemical shifts of amide hydrogens in proteins. J Biomol NMR 38: 139–150.
3. Xu XP, Case DA (2001) Automated prediction of 15N, 13C$^\alpha$, 13C$^\beta$ and 13C chemical shifts in proteins using a density functional database. J Biomol NMR 21: 321–333.
4. Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J Biomol NMR 38: 289–302.
5. Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein 1h and 13C and 15N chemical shifts. J Biomol NMR 26: 215–240.
6. Meiler J (2003) PROSHIFT: Protein chemical shift prediction using artificial neural networks. J Biomol NMR 26: 25–37.
7. Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate pre-dictions of protein NMR chemical shifts from interatomic distances. J Am Chem Soc 131: 13894–13895.
8. Wishart D, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338: 3–34.
9. Case DA (2013) Chemical shifts in biomolecules. Curr Opin Struct Biol 23: 172–176.
10. Parker LL, Houk AR, Jensen JH (2006) Cooperative hydrogen bonding effects are key determinants of backbone amide proton chemical shifts in proteins. J Am Chem Soc 128: 9863–9872.
11. Barfield M (2002) Structural dependencies of interresidue scalar coupling ^{h3}J nc' and donor 1H chemical shifts in the hydrogen bonding regions of proteins. J Am Chem Soc 124: 4158–4168.
12. Cornilescu G, Ramirez BE, Frank MK, Clore MG, Gronenborn AM, et al. (1999) Correlation between ^{h3}J nc' and hydrogen bond length in proteins. J Am Chem Soc 121: 6275–6279.
13. Vila JA, Scheraga HA (2009) Assessing the accuracy of protein structures by quantum mechanical computations of 13C(alpha) chemical shifts. Acc Chem Res 42: 1545–1553.
14. Vila JA, Arnautova YA, Martin OA, Scheraga HA (2009) Quantum-mechanics-derived 13Ca chemical shift server (cheshift) for protein structure validation. Proc Natl Acad Sci 106: 16972–16977.

15. Boomsma W, Frellsen J, Harder T, Bottaro S, Johansson KE, et al. (2013) PHAISTOS: a framework for markov chain monte carlo simulation and inference of protein structure. J of Comp Chem 00: 000-000, DOI: 10.1002/jcc.23292.
16. Kaminski GA, Friesner RA (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105: 6474–6487.
17. Qiu D, Shenkin PS, Hollinger FP, Still WC (1997) The GB/SA continuum model for solvation: A fast analytical method for the calculation of approximate born radii. J Phys Chem A 101: 3005–3014.
18. Czinki E, Császár AG (2004) On NMR isotropic chemical shift surfaces of peptide models. J Mol Struct (THEOCHEM) 675: 107–116.
19. Rablen PR, Pearlman SA, Finkbiner J (1999) A comparison of density functional methods for the estimation of proton chemical shifts with chemical accuracy. J Phys Chem A 103: 7357–7363.
20. Pople JA (1956) Proton magnetic resonance of hydrocarbons. J Chem Phys 24: 1111.
21. Pople JA (1958) Molecular orbital theory of aromatic ring currents. Mol Phys 1: 175–180.
22. Christensen AS, Sauer SPA, Jensen JH (2011) Definitive benchmark study of ring current effects on amide proton chemical shifts. J Chem Theory Comput 7: 2078–2084.
23. Jin L, Briggs SL, Chandrasekhar S, Chirgadze NY, Clawson DK, et al. (2000) Crystal structure of human parathyroid hormone 1-34 at 0.9 å resolution. J Biol Chem 275: 27238–27244.
24. Jain R, Bally T, Rablen PR (2009) Calculating accurate proton chemical shifts of organic molecules with density functional methods and modest basis sets. J Org Chem 74: 4017–4023.
25. Bottaro S, Boomsma W, Johansson KE, Andreetta C, Hamelryck TW, et al. (2011) Subtle monte carlo updates in dense molecular systems. J Chem Theory Comput 8: 695–702.
26. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup, execution, and analysis of poisson-boltzmann electrostatics calculations. Nucl Acids Res 32: W665–W667.
27. Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, et al. (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucl Acids Res 35: W522–W525.

28. Cordier F, Grzesiek S (1999) Direct observation of hydrogen bonds in proteins by interresidue 3hJNC' scalar couplings. J Am Chem Soc 121: 1601–1602.

29. Bax A (2003) Weak alignment offiers new nmr opportunities to study protein structure and dynamcs. Prot Sci 12: 1–16.

30. Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. Science 308: 303–306.

31. Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, et al. (2010) Potentials of mean force for protein structure prediction vindicated, formalized and generalized. PLoS ONE 5: e13714.

32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucl Acids Res 28: 235–242.

33. Li H, Robertson AD, Jensen JH (2005) Very fast empirical prediction and rationalization of protein pKa values. Proteins 61: 704–721.

34. Zhang H, Neal S, Wishart D (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25: 173–195.

35. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, et al. (2008) Biomagresbank. Nucl Acids Res 36: 402–408.

36. Markwick PRL, Sprangers R, Sattler M (2003) Dynamic effects on j-couplings across hydrogen bonds in proteins. J Am Chem Soc 125: 644–645.

37. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, et al. (2004) Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT.

38. Losonczi JA, Andrec M, Fischer MW, Prestegard JH (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. J Magn Reson 138: 334342.

39. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433: 128–132.

40. Shen Y, Bax A (2010) SPARTA+: a modest improvement in empirical NMR chemical shift pre- diction by means of an artificial neural network. J Biomol NMR 48: 13–22.

41. Fenwick RB, Esteban-Martn S, Richter B, Lee D, Walter KFA, et al. (2011) Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. J Am Chem Soc 133: 10336–10339.

42. Lange OF, Lakomek NA, Fars C, Schrder GF, Walter KFA, et al. (2008) Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. Science 320: 1471–1475.

43. Richter B, Gsponer J, Várnai P, Salvatella X, Vendruscolo M (2007) The mumo (minimal under-restraining minimal over-restraining) method for the determi- nation of native state ensembles of proteins. J Biomol NMR 37: 117–135.

44. Lindorff-Larsen K, Best R, DePristo M, Dobson C, Vendruscolo M (2004) Simultaneous determination of protein structure and dynamics. Nature 433: 128–132.

45. Cornilescu G, Marquardt J, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 120: 6836–6837.

46. Vijay-Kumar S, Bugg C, Cook W (1987) Structure of ubiquitin refined at 1.8 a resolution. J Mol Biol 194: 531–544.

47. Ramage R, Green J, Muir T, Ogunjobi O, Love S, et al. (1994) Synthetic, structural and biological studies of the ubiquitin system: the total chemical synthesis of ubiquitin. Biochem J 299: 151–158.

48. Alexeev D, Barlow PN, Bury SM, Charrier JD, Cooper A, et al. (2003) Synthesis, structural and biological studies of ubiquitin mutants containing (2s, 4s)-5- uoroleucine residues strategically placed in the hydrophobic core. ChemBio- Chem 4: 894–896.

Bibliography

[Rieping et al., 2005] Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science*, 308:303–306.