PhD thesis

SIMON OLSSON

Inference problems in structural biology: protein structures and protein structure ensembles
Cover image: 500 samples from the native ensemble of GB3 in free form superimposed onto GB3 in its binding site on Fab. The samples were generated using a method proposed in section 2.2 details about the simulation may be found in section 2.3.
Preface

This thesis is a summary of the research I have conducted between April 2010 to April 2013 which is in relation to the thesis title. The work has been carried out at the Bioinformatics center, Department of Biology, University of Copenhagen as part of the fulfillment of the PhD degree. Work carried out during a stimulating stay in Douglas Theobalds lab at Brandeis University in the summer of 2012 is not presented in this thesis.

The key goals of the work carried in this thesis was to, firstly, improve and, secondly, generalize existing methodology of probabilistic structure inference from experimental data. The first aspect was motivated by the considerable computational demands of previous implementations. This limitation could prohibit the practical tractability of any generalizations. The second part was motivated by an increasing interest in distributions, or ensembles, of protein structure that are in accord with noisy and averaged experimental data.

I present a general introduction to the topics essential to the understand the original manuscripts. This introduction also contains a discussion of previous work in the field. As the literature on these topics is quite substantial, I have made a considerable effort to prioritize the presented background material.
Abstract

The structure and dynamics of biological molecules are essential for their function. Consequently, a wealth of experimental techniques have been developed to study these features. However, while experiments yield detailed information about geometrical features of molecules, this information is often incomplete and subject to averaging through both space and time. In addition, experimental noise often is significant. These facts complicate the use of the information in the construction of models representing the conformational properties of the molecules.

In this thesis I review the current literature in the field of protein structure and structural ensemble determination from experimental data. Following this I present three original research papers which address a number of current shortcomings in the field. Firstly, a method to increase the efficiency of probabilistic structure determination is presented. Second, a generalization of this structural inference framework is presented, to account for flexibility in the molecules. Finally, I apply the generalization to restrain a simulation of the native fluctuations of a protein using experimental data.

Resume


# Contents

0.1 List of Abbreviations ........................................ 7  
0.2 Notation .................................................. 7  

1 General introduction ........................................ 9  
1.1 Protein structure and dynamics ............................. 10  
1.2 Biophysical experiments .................................... 12  
1.2.1 Nuclear magnetic resonance spectroscopy .............. 15  
1.3 Probability theory ........................................... 21  
1.3.1 Bayesian Networks .................................... 22  
1.3.2 Prior distributions ...................................... 24  
1.3.3 Expectation Maximization ............................... 26  
1.3.4 Kullback-Leibler divergence ............................ 28  
1.3.5 The reference ratio method ............................. 28  
1.4 Statistical sampling in high-dimensional spaces ......... 29  
1.4.1 The Metropolis-Hastings algorithm .................... 32  
1.4.2 Generalized ensembles ................................ 33  
1.4.3 Proposal strategies for protein simulations .......... 34  
1.5 Methods for inference of protein structures and ensembles 36  
1.5.1 General considerations. ................................ 37  
1.5.2 Hybrid energy minimization heuristics ................ 40  
1.5.3 Inferential structure determination .................... 41  
1.5.4 Restrained molecular simulations ..................... 42  
1.5.5 Post hoc ensemble selection ........................... 44  
1.5.6 Analysis of methodology .............................. 44  

2 Scientific Papers .............................................. 59  
2.1 Generative probabilistic models increase the scope of inferential structural determination ..................... 59  
2.2 Inferential determination of protein structural ensembles from sparse, averaged data. .................... 70
2.3 Construction of protein structure ensembles in continuous space through minimal biasing of a simple physical force field. 86

3 Closing

3.1 Conclusion and Outlook ........................................... 97
3.2 Acknowledgments .................................................. 101

4 Appendix

4.1 Appendix A: Technical note on adaptive kernel proposals . 103
4.2 Appendix B: TYPHON .............................................. 104
4.3 Appendix C: Phaistos ............................................. 117
0.1 List of Abbreviations

PDF probability density function
EM Expectation Maximization
NMR nuclear magnetic resonance spectroscopy
PDB protein databank
RDC residual dipolar coupling
SAXS Small-angle X-ray scattering
NOE Nuclear Overhauser enhancement
MCMC Markov chain Monte Carlo
MH Metropolis Hastings
ISD Inferential Structure Determination

0.2 Notation

There will be some use of mathematical formulas throughout this thesis. This section will briefly describe the notation used. Lowercase italic latin letters, such as $x$, will represent scalar values. Uppercase and lowercase boldface latin letters will be used to represent matrices (e.g. $A$) and vectors (e.g. $x$), respectively. Probability density functions (PDFs) will generally be represented by $p(\cdot)$ when no particular distribution is specified. Similarly, $P(\cdot)$ will be used for discrete probability mass functions. Though, prior probabilities will usually be represented by $\pi(\cdot)$. When normalized probability functions are specified, calligraphic lettering (e.g. $\mathcal{N}$, for the Normal distribution) will be used. Since I take a Bayesian stand in this thesis, I will not explicitly distinguish between random and non-random variables.
Chapter 1

General introduction

Analysis of data obtained in experiments remains an essential part of many scientific studies. Models obtained through such analyses may provide evidence for the synthesis of novel scientific hypothesis. Thus, the accuracy and validity of such hypotheses are strictly dependent upon the rigor of the analysis method employed.

Here, I consider a series of biophysical experiments related to nuclear magnetic resonance spectroscopy (NMR). In these experiments inferred models will here correspond to determining a biomolecular structure or an ensemble of such. Proteins will be the biomolecules discussed herein.

To establish methods that rigorously yield sound structural models of the studied proteins we need to map the properties of the employed experiments. Firstly, experimental noise is intrinsic to most experimental measurements. Secondly, the information which can be derived from experimental measurements is often incomplete with respect to full specification of the system [1]. For biomolecular systems this corresponds to observing very sparse data or data redundant in information [2, 3]. Finally, in the methods considered herein measurement is carried out on samples of near molar concentrations, during timescales which exceed that of most motion on atomic scale. Consequently, our observations are temporal and spatial averages of systems possibly undergoing complex motions [4].

These facts prohibits us from "letting the data speak for itself", in the vast majority of cases. That is, in a context where we wish to do inferences of models our data cannot stand alone. We may, however, introduce useful assumptions about the data and/or the models that we wish to consider. In a statistical formalism this corresponds to explicitly introducing prior information – this branch of statistics is known as Bayesian statistics.

Prior information may amount to a range of different things. Firstly,
the choice of error models may be motivated by some known properties of the data and some desired properties of the model [5, 6, 7]. Second, forward models, or structural relationships, are typically motivated by physical theory, or approximations thereof [8, 9, 10, 11, 12]. These, provide the means to relate experimental observables to geometrical properties of the biomolecules, through a number of model parameters. Model parameters of the forward- and error-models may again be subject to other prior assumptions. Finally, prior knowledge may be used to limit the space on which we wish to make inferences. In the case of structure determination, this may correspond to applying a forcefield or a knowledge-based potential in the inference process [13, 14, 15].

In this thesis I will present a number of new methods for inference of protein structure, protein structural ensembles as well as associated model parameters. Probability theory provides a natural mathematical language to formulate models for rigorous inference – the basic principles for formulation of and inference from such models will be introduced in this synopsis. The experimental methods providing the data on which inferences are made, will also be outlined, along with existing forward-models, which relates the data to structures. Finally, there will be a introduction to protein structure. In particular, there will be an emphasis on existing methodology for inference and prediction of protein structure and protein structural ensembles.

1.1 Protein structure and dynamics

Proteins are one of the four major classes of macromolecules ubiquitous in living systems – lipids, carbohydrates and nucleic acids being the other three. Proteins are often envisaged as the molecular machines of life. Indeed, they do carry many of the functions essential to the longevity and proliferation of their hosts including catalysis, transport, regulation and structure [16].

Amino acids are the basic constituents of proteins. There exist 20 common ones, each of which are characterized by the physiochemical properties of their side-chains. A protein is formed by a sequence of these joining to form a linear heteropolymer, through a condensation reaction. The sequence of amino acids in a protein is commonly referred to as its primary structure. When put in an aqueous environment many proteins will spontaneously adapt local (secondary) and non-local (tertiary) structures (see figure 1.1), while other require the presence of molecular chaperones or co-factors [17, 18]. This observation has lead to the hypothesis that the native state is encoded in the primary sequence of amino acids, as originally for-
1.1. PROTEIN STRUCTURE AND DYNAMICS

mulated by Nobel laureate Christian Anfinsen and his collaborators. The process of the protein reaching its native state is known as protein folding. Much effort has gone into understanding this problem, both by experimental and theoretical means. In spite of much progress, a full understanding of this process remains one of the most important problems in molecular biology.

![Figure 1.1: A raytraced rendering of a cartoon representation of the third immunoglobulin binding domain of protein G (pdb accession code: 2oed). The secondary structure elements are highlighted by their black (coil), blue ($\beta$-strands) and cyan ($\alpha$-helix) colors. The tertiary structure is illustrated by the spatial arrangement of the secondary structure elements. [Figure prepared using PyMOL (DeLano Scientific LCC).]](image)

The native state has classically been regarded as a well-folded rigid structure. However, proteins are intrinsically dynamic entities [19], covering many spatial and temporal orders of magnitude, see Figure 1.2. In fact, an increasing amount evidence has suggested that the dynamic character of proteins is essential to biological functions [20, 21] such as ligand-binding [22], allostery regulation [23], signal-transduction [24] and catalysis [25]. Many of these functions occur on timescales which inherently difficult to study: the microsecond to millisecond window. Although the experimental techniques discussed herein in principle cover this window, successful studies of these timescales remain the exception rather than the rule. Ultra long molecular dynamics simulations still require special-purpose super computers [26].
In NMR, relaxation dispersion experiments depend on a considerable perturbation of chemical environment between exchanging states [27] and may require experimental equipment to be pushed to its limits [28]. Alternatively, enormous amounts complementary data be acquired and combined with simulations [29, 30]. Finally, diffraction experiments provide rather low-resolution information, limiting its application to study dynamic events which have substantial spatial rearrangements. Consequently, any progress which will make the study of this time window on an atomic more accessible will be of great interest.

More recently, proteins apparently devoid of well-defined folds have emerged [31]. These are frequently referred to as intrinsically disordered proteins and their functional span in biology is broad [32, 33]. In particular, their flexibility has been suggested to be important to function [34]. Some studies suggest that some of these proteins are well-folded \textit{in vivo} and suggests that the observation of disorder is rather an artifact of \textit{in vitro} experimentation [35]. While, this may indeed be valid for some particular examples, any presence of endogenously disordered systems has a significant impact of our understanding of how function arises in structural biology [34].

Representing protein structure  There are a number of ways to parameterize protein conformations. For instance may we explicitly represent all \( N \) atomic positions as degrees of freedom in the Euclidean space \( \mathbb{R}^{3N} \), or the conformational space. Alternatively, we may use coarse grained representations, where a reduced set of degrees of freedom aim at representing the protein – there exists a wide variety of methods to accomplish this [36]. One example is to represent the protein by its dihedral angles, \( \phi, \psi \) and \( \omega \) for the backbone (see figure 1.3) and \( \chi \)-angles for the side-chains [37]. This representation allows us to fully reconstruct all atomic positions by assuming bond-angles and bond-lengths are constant at some idealized values, for instance those specified by Engh and Huber [38]. This representation is what will be employed throughout this thesis.

1.2 Biophysical experiments

A broad range of biophysical techniques exist to study the structure in biomolecules. However, due to intrinsic insensitivity of many of these, the samples studied usually contain a large number \( N \) of identical molecules, typically concentrations above those typically expected \textit{in vivo}. Our observed signal, \( \bar{d} \) is thus a spatial average of all the microscopic contributions
Figure 1.2: A qualitative overview of the time-scales of motions in proteins and their connection some biological events and the capabilities of experimental techniques. Figure loosely adapted from [19].
CHAPTER 1. GENERAL INTRODUCTION

Figure 1.3: Dihedral angles ($\omega$, $\phi$ and $\psi$) of an alanine residue in a polypeptide backbone of a protein. Atoms (black dots) interconnected by covalent bonds (lines). Figure adapted from [14].

\[
\bar{d} = \frac{1}{N} \sum_{i=1}^{N} \delta_i \tag{1.1}
\]

Each molecule, represented theoretically by $\delta_i$, individually undergoes fluctuations. In addition, many of these motions occur on time-scales faster than the time of acquisition of a biophysical experiment, $\tau_e$. Thus these $\delta_i$s are themselves temporal averages,

\[
\delta_i = \frac{1}{\tau_e} \int_{t=0}^{\tau_e} dt f(x_t), \tag{1.2}
\]

where $f(\cdot)$ is a forward model which relates a conformational microstate $x_t$ at time $t$ to its corresponding contribution to the experimental observation. Consequently, the signals observed are time- and ensemble-averaged. Here I will use the common assumption that all $N$ molecule passes through the same conformational space during $\tau_e$. In this case the spatial and temporal averages are equivalent. This assumption is called the ergodic assumption, [39] and allows us to recast eqs. 1.1 and 1.2 to the expectation of $f(\cdot)$,

\[
\bar{d} = \mathbb{E}[f(x) \mid p(x)] = \int_{x \in \chi} dx \ p(x) \ f(x) \tag{1.3}
\]

where $p(\cdot)$ is a normalized probability density function of the conformational space $\chi$. The probability density function for closed physical systems at
constant temperature is the canonical ensemble \[40\]

\[ p(x) = \frac{\exp(-\beta E(x))}{Z} \]  

where \( E(\cdot) \) is the Hamiltonian of the system and \( \beta = 1/(kT) \) with \( k \) is Boltzmann’s constant, \( T \) is the temperature and \( Z \) is the normalization constant, or the partition function. Consequently, biophysical measurements carried out on bulk solutions are expectations of the forward model with respect to the canonical ensemble (Boltzmann distribution). The implications of this in the context of structure and structural ensemble inference will be discussed in section 1.5.

1.2.1 Nuclear magnetic resonance spectroscopy

NMR (Nuclear magnetic resonance spectroscopy) is a versatile biophysical technique which can be used to study system both in, and away from, chemical equilibrium. Thus, the application of NMR may provide a range of structural as well as kinetic data. The resolution of the data spans a range from atomic detail to overall shape. This thesis will be exclusively concerned with data coming from experiments on samples in chemical equilibrium, and thus the discussion herein will be limited to such \[41\].

In NMR we measure magnetic resonance signals arising when a sample containing NMR active nuclei (where spin quantum number \( S \geq \frac{1}{2} \)) placed in a strong magnetic field, is irradiated by a sequence of radio-frequency (rf) pulses. The pulses perturb the magnetic equilibrium state of the spins. A suitable combination of rf-pulses can isolate a specific phenomenon, reflecting a specific piece of information. These sequences of rf-pulses are commonly referred to as pulse sequences or NMR experiments.

The signal acquired following a given pulse sequence is a linear mixture of complex waveforms, each mixture component corresponds to a nuclei, in a specific physiochemical environment. The signal decays, or relaxes, with a rate depending on a broad range of physical effects. This signal is referred to as the free-induction decay, or the FID. In order to distinguish resonances more easily, the FID is usually transformed from the time-domain into the frequency-domain using the Fourier transform, or alternatively maximum entropy and Bayesian reconstruction techniques \[42, 43\]. This results in an NMR spectrum. In the spectrum, peaks appear at the resonance frequencies corresponding to specific nuclei. The resonance frequencies are usually normalized to the magnetic field strength and referenced yielding chemical shifts. Like their corresponding decaying wave-forms, the chemical shifts
are affected by a broad range of physical effects both properties intrinsic to the nuclear spin and its physical environment. The assignment of peaks to atomic spins typically follows a series of NMR experiments in a procedure which may in some cases be automated [44, 45, 46]. This assignment is important in most subsequent analyses of the data.

**Nuclear Overhauser enhancements**

Nuclear Overhauser enhancements (NOEs) are the most important (semi-)quantitative measurements in protein structure inference. Cross relaxation is the physical basis of the NOE. This results from the transfer of magnetization through dipolar interactions in between excited nuclear spins. The efficiency of magnetization transfer is inversely related to the inter-atomic distance as the process is driven by dipolar interactions. Measuring this process therefore yields direct structural information. However, the indirect transfer of magnetization via other spins (spin diffusion) often hampers the interpretation of the NOEs. Thus, to extract structural information approximations are often used. This effectively results in the semi-quantitative nature of the NOE where the power lies in obtaining a large amount of imprecise pair-wise distance restraints [41, 47].

The isolated spin-pair approximation (ISPA) is the most widely used approach to extract information from NOEs. Here, a pair of spins, in a rigid body, proximate in space are thought to be isolated and in this way solely affecting each others cross-relaxation. This yields a simple relationship (a forward-model) between the NOE, $I$, and the inter atomic distance, $r$ as

$$I = \gamma r^{-6}$$

where $\gamma$ is known as the *equilibration parameter*, which relates the arbitrary scale of the NOE intensity to a distance scale. The equilibration parameter may either be estimated using a reference experiment [8] or treated as a nuisance parameter and subject to statistical inference [7]. A formal derivation of the approximation, and a description of the experimental conditions where it is valid may be found elsewhere [8].

More recently some methods have been developed attempting to model the contribution of spin diffusion to the NOE using a simulated relaxation matrix [48] or an auxiliary mean-field spin [49], yielding some more quantitative NOEs, facilitating the determination of ensembles of protein structures [50].

Ambiguously assigned NOEs, $I_a$, often constitute a considerable amount of the available data. Thus it is not exactly known which atomic spin pairs
are giving rise to an observed NOE. However, this information may often be salvaged by considering all possible assignments, \( \mathcal{A} \), and simply using a ‘\( r^{-6} \) summed NOE’ [51],

\[
I_a = \gamma \left( \sum_{i \in \mathcal{A}} r_i^{-6} \right). \tag{1.6}
\]

This relationship is of course only valid if the equilibration factor of the different spin pairs is the same – this assumption include a range of physical effects, perhaps most significantly that the spin-spin correlation times and order parameters are identical for all spin-pairs, \( \mathcal{A} \).

I stress that the ISPA forward-model is an approximation, and is highly degenerate. That is, there will be many pairs of \( \gamma \) and \( r \) that will yield a correct NOE – this degeneracy becomes even more pronounced when assignment of the NOEs is ambiguous. Consequently, these short-comings are ideally accounted for when used for structure calculation.

**Residual dipolar couplings**

The residual dipolar coupling (RDC) is another parameter which is measurable by NMR. From a set of these, we may obtain information about relative orientational geometry of inter-atomic vectors in the frame of the molecular system. In many cases, RDCs may be used for validation or refinement of determined structures [52] or in some cases even structure determination [53].

The geometrical information of the RDCs is averaged on time-scales up to 100ms. Consequently, any dynamic processes happening on time-scales shorter than this may be uncovered, provided sufficiently accurate measurements [54].

A dipolar coupling between spins \( i \) and \( j \) in a body undergoing rotational diffusion is commonly expressed in the laboratory frame, [41]

\[
D_{ij} = D_{\text{max}} \left\langle \frac{3 \cos^2 \theta_{ij} - 1}{2} \right\rangle \tag{1.7}
\]

with

\[
D_{\text{max}} = -\frac{\mu_0 \gamma_i \gamma_j \hbar}{8\pi^3 r_{ij}^3}. \tag{1.8}
\]

\( \theta_{ij} \) is the angle between an external magnetic field and the vector between spins \( i \) and \( j \), \( r_{ij} \) with length \( r_{ij} \), \( \mu_0 \) is the magnetic permittivity of vacuum, \( \gamma_X \) is the gyromagnetic ratio of nuclei \( X \) and \( \hbar \) is Planck’s constant. This
corresponds to the time-independent form of the secular part of the dipolar interaction Hamiltonian which approximates the full Hamiltonian well at high magnetic fields \([54]\). We have assumed that the bond-length is time-independent.

The laboratory frame is, however, often of limited practical use, as it is difficult to deconvolute the angular average as it depends on a range of effects including the internal and overall motions of the molecule and the molecular structure. Consequently, it is often more convenient to express the dipolar coupling in the \textit{molecular frame} of a rigid body,

\[ D_{ij} = D_{\text{max}} r_{ij}^T A_{ij}. \] (1.9)

The Saupe, or \textit{alignment}, tensor, \(A\), is a \(3 \times 3\) traceless matrix with five independent components. It represents the alignment (effective average orientation) of the molecule with respect to an external magnetic field \([9, 5]\).

In an isotropic solution, where molecules are allowed to undergo free rotational diffusion, dipolar couplings average to zero. Consequently, the measurement of RDCs crucially depends on partial alignment, that is, an effective average orientation of the molecule with respect to the external magnetic field. Using a nematic solvent, such as liquid crystals, may yield a partial alignment \([55]\). Alternatively, systems that spontaneously align in an external magnetic field may be used \([56]\). I will here limit the discussion to alignment in nematic phase solvents.

The alignment, and therefore also \(A\), is determined by the physiochemical properties (e.g. steric and electrostatic) of molecular system as well as interplay with a nematic solvent. Additionally, the degree of alignment may be dependent on the concentration of the alignment media. Typically, all of these effects are summarized in the alignment tensor. From equation (1.9) it follows that detailed structural characterization depends on an accurate determination of \(A\).

Methods for the the determination of alignment tensors can be put in to two general groups: data-based and data-free. The data-based methods include the histogram method \([57]\) and tensor fitting procedures (see below) \([58]\). Many of these procedures may be seen as special cases of a general probabilistic framework formulated by Habeck and co-workers \([5]\). The data-free, methods employ simple physical models of the alignment event and result in \textit{de novo} prediction of alignment tensors \([60, 61, 62]\). However, both approaches have their distinctive pros and cons. Data-based methods are phenomenologically motivated, and thus physical realism of the tensor is not guaranteed – it may however be imposed \textit{post hoc}. Conversely, data-free
methods are typically mechanistically motivated – however, their predictive power is typically limited by the precision of the physical models used. I will give examples of each of such types of alignment tensor models below.

**The Almond-Axelsen tensor** The Almond-Axelsen alignment tensor model is mechanically motivated: it assumes that the hydrodynamic shape of a molecule is the main contributor to alignment. This assumption has been supported by results in their paper presenting the method [61] and a series of similar mechanistic tensor models [60, 62]. Unlike the more widely used method PALES (prediction of alignment from structure) explicit simulation of the alignment event is avoided, greatly increasing the computational efficiency of the method.

Almond and Axelsen exploit that the $3 \times 3$ hydrodynamic (or gyration) tensor,

$$
G = \frac{1}{N} \sum_{i=1}^{N} x_i^T x_i
$$

(1.10)

of a protein structure, $x$, situated in its center of mass, shares its eigenbasis with the alignment tensor, $A$. While, the eigenvectors of the two matrices are shared, the eigenvalues are not. However, the following relations for the eigenvalues, $\alpha_i$, of $A$ to the eigenvalues, $\gamma_i$, of $G$ were observed to provide good results:

$$
(\alpha_1, \alpha_2, \alpha_3) \propto \left(1 - \frac{\delta}{2}, \delta - \frac{1}{2}, -\frac{1}{2} - \frac{\delta}{2}\right)
$$

(1.11)

with,

$$
\delta = \frac{\sqrt{\gamma_2^2 - \gamma_3^2}}{\sqrt{\gamma_1^2 - \gamma_3^2}}
$$

(1.12)

where $\gamma_1 > \gamma_2 > \gamma_3$ and $\alpha_1 > \alpha_2 > \alpha_3$. Thus, determination of the alignment tensor reduces to the determination of a gyration tensor, an eigenvalue decomposition (or similar) and a few matrix-matrix multiplications. Note, that the alignment tensor obtained is only known up to a scaling constant, as absolute degree of alignment is not known.

While the tensor model of Almond and Axelsen is very efficient it does currently not allow for modeling of non-steric alignment and is not trivially extended to do so. One example is when alignment is due to electrostatic interactions between the alignment-media and the aligning molecule. In this case simple steric obstruction does not accurately capture the molecular alignment [62, 63]. However, the lack of generally successful mechanistic models of electrostatic alignment, suggests that it is an intrinsically difficult problem.
CHAPTER 1. GENERAL INTRODUCTION

**Tensor fitting** Unlike the Almond-Axelsen tensor discussed above the tensor fitting procedure is phenomenologically motivated. That is, it is independent of any underlying physical model of the alignment event, but rather aims at determining the alignment tensor which explains the data best. As suggested by its name, the procedure amounts to fitting the an alignment tensor, $\mathbf{A}$, given some structure, or set of structures, and some observed data. This is achieved by considering the problem as a system of linear equations, and solving it [58].

We may reformulate eq. 1.9 above to

$$\frac{D_{ij}}{D_{\text{max}}} = \mathbf{p}_j^T \mathbf{a},$$

(1.13)

where $\mathbf{a}$ is a $5 \times 1$ vector independent components of $\mathbf{A}$ and $\mathbf{p}$ is $5 \times 1$ vector of projections of the normalized inter-nucleic vector $\mathbf{r}_{ij}$. We have,

$$\mathbf{p}_{ij} = \begin{bmatrix} x^2 - y^2, y^2 - z^2, 2xy, 2xz, 2yz \end{bmatrix},$$

(1.14)

where $x, y, z$ are the cartesian coordinates of the inter-nucleic vector $\mathbf{r}$. Eq. 1.13 is a linear equation with five unknowns, however, if we observe $N$ data points, we may recast the equation to,

$$\frac{\mathbf{D}}{D_{\text{max}}} = \mathbf{P}^T \mathbf{a},$$

(1.15)

where $\mathbf{D}$ is a $N \times 1$ vector of the observed RDCs, $\mathbf{P}$ is an $N \times 5$ matrix of projections and $\mathbf{a}$ is our vector of independent alignment tensor components. This system of equations is overdetermined (iff $N > 5$) any may readily be solved by using the Moore-Penrose pseudo-inverse, singular-value decomposition or other methods.

A set of structures, or an ensemble, may also be used in the fitting procedure. Here, the projections are replaced by average projections,

$$\bar{\mathbf{p}}_{ij} = \begin{bmatrix} \bar{x}^2 - \bar{y}^2, \bar{y}^2 - \bar{z}^2, 2\bar{xy}, 2\bar{xz}, 2\bar{yz} \end{bmatrix},$$

(1.16)

where the averages are over the ensemble of structures. In turn we may use these averaged projections to form a new projection matrix, $\bar{\mathbf{P}}$. It is important to ensure that the structures are appropriately superpositioned prior to the calculation of the projections.

Both models for the determination of alignment tensors may be considered structural relationships or forward models – as may the ISPA. They will have the respective advantages in different applications. The Almond-Axelsen tensor will be useful when studying ensembles of structures where
the structural heterogeneity is large where meaningful superpositioning is
difficult [64]. Examples of such may be unfolded protein states or intrinsi-
cally disordered proteins – here we may assume that each of the members
of the ensemble align independently and does not change shape during the
alignment event [65, 66]. When superpositioning is not needed or no prob-
lem, the fitting of tensors may be desirable, in particular when alignment is
non-steric [30].

1.3 Probability theory

From experimental noise and data sparsity arises one of key concepts of the
inference problems: uncertainty. The theory of probability provides a nat-
ural framework for the quantification of uncertainties [67]. Consequently,
the application of probability theory to the problems of inferring protein
structures and protein structural ensembles from data allows for rigorous
assessment of data and models [2]. This section will outline the basic con-
cepts essential to probabilistic protein structure inference.

There are two fundamental rules that govern probabilistic models: the
sum and product rules. For the variables $X$ and $Y$, we have:

**Sum rule:** \( P(X) = \sum_Y P(X, Y) \)

**Product rule:** \( P(X, Y) = P(X \mid Y)P(Y) \).

The summation runs over all possible outcomes of the discrete variable,
$Y$, in a process often referred to as *marginalization*. Marginalization, or
integrating out, can be generalized to continuous variables ($x$ and $y$) by
replacing the sum by an integral:

\[
p(x) = \int_y dy p(x, y).
\]

The product rule reads: the joint probability of $X$ and $Y$ is equal to the
probability of $X$ *given* $Y$ times the probability of $Y$. Please note, the product
rule takes the same form for continuous variables.

Since the probability of $X$ and $Y$ is the same as the probability $Y$ and
$X$, we may use the product rule to obtain,

\[
P(Y \mid X)P(X) = P(X \mid Y)P(Y)
\]

which is equivalent to

\[
P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}.
\] (1.17)
This relationship between the conditional probabilities of $X$ and $Y$ is the much celebrated Bayes’ theorem. In a Bayesian view probabilities quantify our degrees of belief, and Bayes theorem has a very particular interpretation [68, 67]. For example, let $Y$ be a protein structure and $X$ be some experimental data. The conditional probability $P(Y \mid X)$ is called the posterior which quantifies our belief of a certain protein structure $Y$ having observed our data, $X$. $P(X \mid Y)$, is called the likelihood of the data, $X$, given a protein structure, $Y$. Finally, $P(Y)$ is called the prior distribution of $Y$, and specifies our belief, or knowledge, about protein structures ($Y$) before having observed the data ($X$). The denominator, $P(X)$ is the typically called the evidence. Using the sum and product rules, we note the evidence can take the form 

$$P(X) = \sum_Y P(X \mid Y)P(Y),$$

which is exactly the sum of quantities in the nominator.

While this section has primarily been considering discrete variables, the principles discussed also apply to continuous variables.

### 1.3.1 Bayesian Networks

The process of constructing Bayesian models usually involves repeated application of the sum and product rules to joint probability functions. This can occasionally lead to intricate expressions which in turn make it difficult to parse out statistical independencies amongst variables. However, it is possible to establish diagrams, or graphs, of probabilistic models. These yield a visualization of the model and determine the independencies of the model. These two properties of graphs may be used to make visual adjustments and expert assessments, while at the same time formulating the mathematical model [69].

Graphs are composed of two basic components: nodes which represent variables or sets thereof. The nodes are interconnected by edges which represent their statistical relationships. To illustrate Bayesian networks consider the joint probability function $p(d, e, f)$ of variables $d$, $e$ and $f$ in the factorization

$$p(d, e, f) = p(d \mid e)p(f \mid e)p(e).$$

(1.18)

The corresponding Bayesian network of this factorization is shown in figure 1.4. Each of the variables are represented by a node and for each for each of the conditional distributions, we add directed edges (arrows) representing the conditioning.
From the graph we can directly read out conditional independence of $d$ and $f$ given $e$ – such information can greatly ease estimation and inference from such models. If we consider all paths between $d$ and $f$ in the network and check whether these are blocked or not. A path between two nodes is any sequence of edges which connect these. A given path from $d$ to $f$ is said to be blocked if arrows meet head-to-tail or tail-to-tail at $e$ or if arrows meet head-to-head at a node which is not $e$ nor either of its descendants are $e$. By descendants we here mean all nodes which can be reached from a given node following the direction of the edges – $d$ and $f$ are descendants of $e$ in figure 1.4. If all paths from $d$ to $f$ are blocked these two are said to be d-separated [69]. D-separation is a necessary and sufficient condition to conclude conditional independence of $d$ and $f$ given $e$. In our example we observe only one path connecting the nodes $d$ and $f$: the path meets tail-to-tail at $e$ which implies it is blocked and thus that $f$ and $d$ are d-separated [70].

![Figure 1.4](image.png)

Figure 1.4: An example of a Bayesian network representing a joint probability distribution of three variables, $d$, $e$ and $f$, according to the right hand side factorization of equation (1.18).

More generally, any factorization of a joint probability distribution can be represented by a graph. D-separation also applies to disjoint sets of nodes on a graph – in the example above the disjoint sets only contain one node each.

**Latent variables.** Often Bayesian network models are used to model joint distributions of multiple parameters through latent or hidden variables which, for instance, are on two different manifolds. As an example, I will turn to figure 1.4 again: let $e$ be a latent variable which describes
the relationship between the number of amino-acid residues in a protein \(f\) and its iso-electric point \(d\). The parameter \(f\) would be modeled using a discrete multinomial distribution, while the continuous Normal distribution appears appropriate to model \(d\). In this example the latent variable, \(e\), does not have any apparent physical interpretation, but serves to model any correlation structure between these two types of experimental data. The analysis of the latent variable states often reveal physically sensible features [14, 71]. As the latent variables are unobserved they are to be estimated estimated from training data using the Expectation-Maximization algorithm (see section 1.3.3).

There are a range of examples of Bayesian network models of structure in biological macromolecules including RNA [72] and proteins [73, 14, 74]. These latter models will be discussed below.

### 1.3.2 Prior distributions

The construction of the prior distribution is one of the key aspects of the Bayesian view of probability theory. Ideally, a given prior distribution on a particular variable should quantify our ignorance about that same variable. Such a prior typically referred to as being non-informative, or ‘objective’. In the one-dimensional case there are a number of ways to construct such priors including, the principle of maximum entropy (and the related principle of indifference) and Jeffreys principle of invariance. The former readily apply to discrete variables, in finite intervals, whereas the latter apply to bounded or unbounded one-dimensional continuous variables [75]. This in this thesis I exclusively work with continuous variables. Consequently, I will only introduce the concept of the Jeffreys prior.

Jeffreys principle of invariance states that any prior density on a parameter, \(\theta\), should yield an equivalent result if applied to a bijective transformation of that parameter [76]. From this principle, we may get a prior of \(\pi(\theta)\) of a one-dimensional parameter \(\theta\) is given by,

\[
\pi(\theta) \propto \sqrt{\mathcal{J}(\theta)} \tag{1.19}
\]

where \(\mathcal{J}(\theta)\) denotes the Fisher information of \(\theta\). The Fisher information is given by,

\[
\mathcal{J}(\theta) = -\int_{x \in \chi} dx \left[ \frac{d^2}{d\theta^2} \log p(x \mid \theta) \right] p(x \mid \theta) \tag{1.20}
\]

which is exactly minus the expectation of second derivative of the log-likelihood of \(x\) given \(\theta\) with respect to \(\theta\). Equation 1.19 may be extended
1.3. PROBABILITY THEORY

to construct multidimensional prior distributions – however, the results are then more controversial [76].

Prior distributions for protein structure

In the opening of the chapter and in section 1.3 I introduced the prior distribution $P(Y)$ of protein structures, however, no details of its origin was given. I will make extensive use of such priors below, thus the final part of this section will introduce a few examples of such. Protein structure is ideally described using a continuous, multidimensional descriptors, for example all atomic position of the given protein. Thus to conform to my standard notation, $\pi(x)$, will from now on refer to prior distribution of protein structure.

We may generally classify the existing prior distributions of protein structure in two different classes: mechanically motivated and phenomenologically motivated. The former typically refers to the Boltzmann distribution $[\pi(x) \propto \exp(-\beta E(x))]$ of a physical forcefield ($E(x) = E_{pf}(x)$) based upon, supposedly bona fide physical interactions within a protein [77]. The latter, commonly known as ’knowledge-based potentials’, is based upon learning statistical models of protein structure using experimentally solved structures deposited in the Protein DataBank (PDB) [14, 74]. However, the boundary between these two classes of model is soft as there exists many examples of combinations of these approaches.

Bayesian network models for protein structure

As a pair the knowledge-based probabilistic models, TorusDBN[14] and Basilisk [74], cover the entire protein conformational space. Both are dynamic Bayesian networks which essentially corresponds to generalized hidden Markov models. Such models are simply Bayesian network models used to model sequential data. The strategy in TorusDBN is to model the sequences of amino acids, $\langle \phi, \psi\rangle$-angle pairs, $\omega$ configuration and secondary structure, jointly. This is done along the backbone yielding a model of local protein structure. Similarly, Basilisk models amino acids, $\langle \phi, \psi\rangle$-angles and side-chain $\chi$-angles as dynamic Bayesian network. Both models are summarized in Figure 1.5. A key strength of these models is the continuous representation of the dihedral angles using uni- and bi-variate von Mises distributions: the Normal distributions on the circle or the torus, respectively. This is is in contrast to the widely used fragment libraries and rotamer libraries, which effectively discretize the conformational space [78]. Additionally, these models allow for sampling and conditioning on any of the variables. Thus, if we assume
idealized bond lengths and bond angles we may for example sample protein structures with realistic local structure (dihedral angles), given an amino acid sequence.

**Physical forcefields for protein structure** As an example of an all-atom protein forcefield I will turn to Profasi, $E_{\text{profasi}}(x)$. This forcefield is parameterized in torsional space, similarly to TorusDBN and Basilisk – that is, torsional angles within a protein, such as $\phi$ and $\psi$ of the backbone and $\chi$-angles of the sidechain. Thus, to be able to specify all the positions of atoms geometrical parameters such as bond lengths and bond angles are assumed to be fixed [79]. The mathematical expression for the interaction potential consists of four terms

$$E_{\text{profasi}}(x) = E_{\text{loc}}(x) + E_{\text{ev}}(x) + E_{\text{hb}}(x) + E_{\text{sc}}(x)$$

where, $E_{\text{loc}}(x)$, describes *local* interactions, that is, interactions between atoms separated through only a few covalent bonds. The remaining three terms account for *non-local* interactions, such as excluded-volume effects ($E_{\text{ev}}(x)$) and hydrogen-bonds ($E_{\text{hb}}(x)$). Finally, charge-charge and hydrophobic interactions between sidechains is contained in the term $E_{\text{sc}}(x)$. The simplicity of Profasi makes its evaluation extremely fast, and thus amenable for use on non-supercomputers. In spite of the simplicity, the forcefield has been shown to exert physical realism on smaller protein systems.

### 1.3.3 Expectation Maximization

The Expectation Maximization (EM) algorithm is a general technique to estimate latent variables in statistical models [80]. It can be shown that the algorithm is guaranteed to yield a maximum likelihood estimate of the latent variables conditioned on the observed data, even if some data is missing [70]. This section will broadly outline the technique.

Let $x$ be a set of experimental observations, $h$ be a set of latent variables and $m$ represent the missing data. From a Bayesian standpoint there is no conceptual difference between missing data and the unobserved latent variables. Here I will however make the distinction, as it makes the discussion of the EM algorithm more clear.

Now, the EM algorithm proceeds as follows. Firstly, the missing values $m$ are filled in using the expected values from $p(m \mid h, x)$ where $h$ is initialized at random. Secondly, the latent variables $h$ are estimating assuming
Figure 1.5: Two dynamic Bayesian network models of backbone and side chain protein conformational space, respectively. Top: TorusDBN models a sequence of hidden node states which which emits probability distributions of $(\phi, \psi)$-angle pairs, $\omega$ configuration, amino acids and secondary structure. Bottom: Basilisk models $(\phi, \psi)$-angles, amino acids and $\chi$ angles through a sequence of hidden nodes.
the missing data is equal to the expectation from the previous step and maximizing \( p(h \mid x, m) \). Use the newly estimated \( h \) in step 1 and repeat the sequence until convergence [76]. In some cases evaluating the expectation is not trivial – here a variation of the algorithm called Stochastic EM may be used [81].

Validity and convergence of an EM algorithm can be assess by considering the change in the expected marginal log-likelihood \( p(x \mid l, m) \) as a function of iteration. A valid algorithm will alway result in a change in the log-likelihood which is either equal to or larger than zero [80].

1.3.4 Kullback-Leibler divergence

The Kullback-Liebler divergence \( KL(p \mid q) \) is a common measure used to compare probability distribution functions \( p(\cdot) \) and \( q(\cdot) \). Although it is not strictly a distance measure, it is often used as an intuitive analogue. The reason for this is that it does not fulfill the axiomatic requirements for a metric – for instance, \( KL(p \mid q) \neq KL(q \mid p) \) nor is \( KL(a \mid c) \leq KL(a \mid b) + KL(b \mid c) \). Rather the divergence more precisely measures the information lost when \( p(\cdot) \) is approximated by \( q(\cdot) \) [82].

For two continuous probability density functions \( p(\cdot) \) and \( q(\cdot) \) of \( x \) the Kullback-Liebler divergence is given by,

\[
KL(p \mid q) = \int_{-\infty}^{\infty} dx \ln \left[ \frac{p(x)}{q(x)} \right] p(x).
\]  

(1.21)

1.3.5 The reference ratio method

Below I will make use of the so-called reference ratio method (RRM). This method allows for seamless combination of coarse grained information and fine grained information [83, 75]. That is, a probability distribution in a detailed representation of a system, for instance a full atomic model of protein structure, \( x \), may be combined with a probability distribution of a reduced representation of the same system. For instance, let \( f(\cdot) \) be a function which maps the fine grained space \( \chi \) into a reduced space in \( \mathbb{R}^N \), \( f : \chi \to \mathbb{R}^N \). This function could for instance be the calculation of the radius of gyration of a protein structure given its coordinates,

\[
R_g = f(x) = \frac{1}{M} \sum_{i=1}^{M} (x_i - \bar{x})^T (x_i - \bar{x}) \in \mathbb{R}
\]  

(1.22)
where \( x_i \) represent mass-weighed atomic coordinate of atom \( i \) and \( \mathbf{x} \) represents the mass-weighed average atomic position, of the \( M \) atoms in the protein.

In the following we will assume that we have obtained a probability distribution of the radius of gyration, \( p(R_g) \), for instance a knowledge-based potential as discussed above. We would like to combine this distribution with some fine-grained distribution of the conformational space, \( q(\mathbf{x}) \), to a new distribution \( \bar{q}(\mathbf{x}) \), such that \( f(\bar{q}(\mathbf{x})) = p(R_g) \). The fine-grained distributions may for instance be TorusDBN discussed above. The RRM provides the result,

\[
\bar{q}(\mathbf{x}) = q(\mathbf{x}) \frac{p(R_g)}{q_r(R_g)}
\]

(1.23)

where \( q_r(R_g) \) is called the reference distribution and is the distribution of \( R_g \) according to the distribution \( q(\mathbf{x}) \). Note, that the distribution \( q_r(\cdot) \) rarely is of a standard form, in particular when more complicated functions for \( f(\cdot) \) are considered. Such cases may often be handled through approximation which provide sufficiently accurate results. In other cases is the distribution \( q_r(\cdot) \) close to uniform (\( \frac{p(\cdot)}{q_r(\cdot)} \approx p(\cdot) \)). Here we may safely ignore \( q_r(\cdot) \) to good approximation.

Although, I’ve here given a simple example from structural bioinformatics, the RRM is a general statistical result. Effectively, it yeilds a Kullback-Leibler optimal modification of a distribution, \( q(\cdot) \), in the light of a distribution in a transformed space, \( p(\cdot) \) [75]. Kullback-Liebler optimality refers to methods property of minimizing the divergence measure of the original probability distribution \( q(\cdot) \) with respect to the resulting one \( \bar{q}(\cdot) \). In information theory this corresponds to preserving as much as the original information as possible and is closely linked to the principle of maximum entropy of statistical physics [82].

1.4 Statistical sampling in high-dimensional spaces

The determination of protein structure and protein structural ensembles frequently amounts to inference based upon a probability density defined for the protein system of interest. For instance, a probability distribution of a physical system may typically be the Boltzmann distribution. The Boltzmann distribution is defined from mathematical expressions of the potential energy of the system – a force field – and specifies the probability of a given micro-state, or molecular configuration, \( \mathbf{x} \). In the context of inference of protein structure and protein structural ensembles, simplistic force fields are
Figure 1.6: An illustration of the reference ratio method where a distribution of local protein structure (TorusDBN) is combined with a coarse-grained probability distribution of the radius of gyration. This figure is based on a figure appearing in [75].
often combined with expression bringing in experimental observations, in a hybrid energy [1]. These probability distributions are of non-standard form and very high-dimensional: they typically are expressed using all atomic positions in protein system of interest. Moreover are these distributions often only known up to a normalization constant. For these reasons, direct sampling is often infeasible and analytical treatment is generally intractable.

This section introduces a set of methods which allows us to draw statistical samples from such high-dimensional distributions – Markov chain Monte Carlo (MCMC) methods. MCMC methods have a long history in most branches of quantitative, computational sciences and engineering [84]. Samples obtained using an MCMC method facilitate the estimation of normalization constants and evaluations of arbitrary expectations, such as experimental observables [3]. In addition, they may be used to determine protein structures and ensembles of such.

The goal of MCMC methods is to construct a Markov chain which has the desired probability distribution as its stationary distribution. A Markov chain is a stochastic process which undergoes transitions in between states. It is characterized by possessing the Markov property: the next state of the process depends only on the current state. The Markov property is sometimes called ‘memorylessness’. Here, I will not distinguish between discrete and general-state space Markov chains, as the results used here are analogous [84].

To guarantee that a given target probability distribution, \( p(\mathbf{x}) \), is stationary with respect to a Markov chain, \( \mathcal{M} \), we must ensure that the transition probabilities \( T_M(\mathbf{x} \rightarrow \mathbf{x}') \) of going from state \( \mathbf{x} \) to \( \mathbf{x}' \) fulfills the property of detailed balance. The detailed balance criterion for our target distribution, \( p(\mathbf{x}) \) is given by

\[
T_M(\mathbf{x} \rightarrow \mathbf{x}') p(\mathbf{x}) = T_M(\mathbf{x}' \rightarrow \mathbf{x}) p(\mathbf{x}').
\] (1.24)

Please note that this is a sufficient but not necessary criterion. For a more thorough discussion of Markov chain theory in relation to statistical sampling I refer to the work of Roberts and Tierney [84, Chapters 3 & 4].

There exists much theory on how to construct Markov chains with a desired distribution, and a complete discussion of this work is beyond the scope of this thesis. Instead I will limit the discussion to the celebrated Metropolis-Hastings (MH) algorithm and the use of generalized ensembles to overcome some of the technical challenges arising when applying the MH algorithm to high-dimensional probability distributions.
1.4.1 The Metropolis-Hastings algorithm

While the discussion above may appear terse, it is fortunately simple to construct a Markov chain with the desired properties. One of the most well-known algorithms to do this is the Metropolis-Hastings (MH) algorithm. The MH algorithm was formulated by Hastings [85] as a generalization of the algorithm proposed by Metropolis and co-workers [86]. Before I proceed to the MH algorithm I will briefly describe the Metropolis algorithm.

The Metropolis algorithm proceeds as follows: given the current state, \( x \), a new state, \( x' \), is proposed from a symmetrical proposal distribution, \( q(x' \mid x) \): that is, \( q(x' \mid x) = q(x \mid x') \). The new state \( x' \), is accepted according to the Metropolis criteria [86]

\[
\mathcal{U}(0, 1) < \min \left( 1, \frac{p(x')}{p(x)} \right),
\]

where \( \mathcal{U}(0, 1) \) represents a uniformly distributed number on the unit interval. If the criterion is not fulfilled the new state is rejected and current state is repeated. On the other hand, if the criterion is fulfilled the new state becomes the current state. We note that the evaluation of the Metropolis criterion will not require calculation of the normalization constant of the target distribution \( p(\cdot) \), as it cancels out in the ratio. This may in many cases be convenient as computation of normalization constants is intractable in many complex systems.

The MH algorithm relaxes the requirement of having a symmetrical proposal distribution \( q(x' \mid x) \). This may be done by simply accounting for the bias introduced by the proposal distribution in the Metropolis criterion,

\[
\mathcal{U}(0, 1) < \min \left( 1, \frac{p(x')q(x \mid x')}{p(x)q(x' \mid x)} \right),
\]

(1.26)

It has been shown that the Metropolis and Metropolis-Hastings algorithms indeed fulfill the criterion of detailed balance, and thus generate samples from the desired target distribution, \( p(x) \) [70].

The choice of proposal distributions is application dependent and often of decides the efficiency of the algorithm. In particular, in complex problems such as protein structure prediction, the curvature of the target probability density changes drastically as a function of state, yielding general proposal difficult. Consequently, such simulations often suffer from either low acceptance or poor mixing. Poor mixing often constitutes to the Markov chain ‘getting stuck’ in unrepresentative modes of the target distribution, or high correlation in between samples, leading the slow convergence.
There has been a lot of work done to overcome the shortcomings of the MH algorithm outlined in the section above. One approach is to replace the target distribution, $p(x)$, by a distribution which may be sampled with higher efficiency. The distributions usually chosen based upon their ability increase connectivity in between states, which otherwise are poorly connected. Collectively, these algorithms are referred to as extended ensembles, and include replica-exchange Monte carlo or parallel-tempering and the generalized ensembles. This section will only be concerned with the latter. A thorough discussion of extended ensembles in general can be found elsewhere [87].

The extended ensembles have their origin in statistical physics, it will therefore be useful to discuss these using the appropriate semantics. Let $\chi$ be the state-space of all configurations, $x$, and let $E(\cdot)$ be a function (an energy function) which maps this state-space, $\chi$, to a real number – that is, $E(\cdot): \chi \rightarrow \mathbb{R}$. A probability density function over all states, $x \in \chi$, may now be defined as,

$$p_w(x) = \frac{w(E(x))}{Z_w}, \quad Z_w = \int_{x \in \chi} \text{d}x \, w(E(x))$$

(1.27)

where $w(\cdot)$ is a weighing function, and $Z_w$ is its corresponding partition function. From physics, a classical choice for the weighting function of some energy, $E_0$, could be $w(E_0) = \exp(-\beta E_0)$, with the inverse temperature, $\beta = (kT)^{-1}$, where $k$ is Boltzmann's constant and $T$ is the temperature. This weighing function corresponds to the Boltzmann distribution, or the canonical ensemble, introduced previously. We note that any (unnormalized) probability density function, $\tilde{p}(x)$, may be expressed in this ensemble with $\beta = 1$ and $E_0 = -\log \tilde{p}(x)$.

Generalized ensembles rely on a reduced representation of the state-space. That is, instead of parameterizing the state-space as $x \in \chi$, a transformation is used. One example could be to use an energy, $E_0$, from an energy function $E(x)$, as introduced above. Using this example, we may obtain the marginal distribution of the the energy $E_0$ as,

$$p(E_0) = \int_{x \in \chi} \text{d}x \, \frac{w(E(x)) \delta(E(x) - E_0)}{Z_w} = \frac{w(E_0)g(E_0)}{Z_w}$$

(1.28)

where $\delta(\cdot)$ denotes Diracs delta-function and $g(E_0) = \int_{x \in \chi} \text{d}x \, \delta(E(x) - E_0)$ is called the density of states in physics. The density of states, $g(E_0)$, may
be interpreted as being a measure of degeneracy of a particular energy, \( E_0 \).
In the generalized ensembles we are still allowed to propose new states using our proposal distribution from above, \( q(x' \mid x) \).

The generalized ensembles distinguish themselves from other extended ensemble methods by changing the weighing function, \( w(E_0) \), the distribution of 'energies' (or \( -\log \tilde{p}(x) \)) away from the Boltzmann distribution. Typically, the weighing function is chosen as to improve traversal of the state-space, that is, to facilitate transitions between low- and high-probability states. I will here outline two of the classical weighing functions: the multicanonical [88] and \( 1/k \) ensembles [89].

In the multicanonical ensemble the weight function is chosen in such a way as to have a uniform marginal distribution over the energies. This is achieved by using the weighing function,

\[
w_{GE}(E_0) = \frac{1}{g(E_0)}. \tag{1.29}
\]

While this weighing function facilitate transitions in between high and low probability regions it may tend to spend much time in low-probability regions, of little interest in the context of structure inference. As an alternative, the \( 1/k \)-ensemble assigns a smooth distribution over the energies with special emphasis on low-energy (high-probability) regions [89]. The weighing function is then given by,

\[
w_{1/k}(E_0) = \left[ \int_{-\infty}^{E_0} \text{d}E \, g(E) \right]^{-1}
\]

in this ensemble.

One of the key challenges in using extended ensembles in practice is that \( g(E_0) \) is rarely known \textit{a priori}. This renders it difficult to evaluate the weight function. However, some algorithms have been proposed to overcome this [90].

### 1.4.3 Proposal strategies for protein simulations

When simulating high-dimensional probability density functions an appropriate proposal strategy is essential to sustain reasonable average acceptance probabilities. This section will introduce methods which allows us to do this for both updating protein conformations, as well as auxiliary parameters.
1.4. STATISTICAL SAMPLING IN HIGH-DIMENSIONAL SPACES

Gaussian kernel proposals  The perhaps simplest strategy to propose new states of a variable is to naively draw a new state from a uniform distribution. This may in some cases allows for efficient sampling. However, typically it will not, particularly in dense system where even minute changes may cause large changes in probability. The use of Gaussian kernel distributions is a simple, yet effective alternative to this naive strategy. A change to the current state is proposed using a normal distribution, where the standard deviation may be chosen to optimize the efficiency of the update. This approach has been discussed in the context of updating protein conformation [91] and we used this strategy to sample nuisance parameters in the scientific papers included in this thesis.

We devised a method to adaptively updating the standard deviation of a given nuisance parameter during the simulation. We make use of the gradient of the ‘energy’ of the current state with respect to the parameter, and use it as to optimize the proposal efficiency over the entire span of the simulation, while obeying the detailed balance criterion. I provide a technical derivation of the procedure in section 4.1.

CRISP  While the Gaussian kernel strategies generally perform well on the auxiliary parameters, they still tend to introduce too radical changes when applied to protein structures. In particular, when updating dihedral angles, a minute angular change in one dihedral angle may result in dramatic conformational changes downstream. One approach to reduce the impact of an angular update is to split the conformational update into two step concerning pre-rotation and a post-rotation, respectively. In this manner a selected local segment of the chain may be selectively updated. The first step involves a stochastic update of a few angles, while the second deterministically restores the local geometry. This is a strategy which has been extensively employed in a local move type called concerted-rotation [92].

A recent development called CRISP joins the two rotational steps such that the pre-rotation explicitly takes into account the deterministic post-rotation. In other words, one constructs a probability distribution over both pre- and post-rotational degrees of freedom at once, making it possible to sample closed structures while maintaining high quality local geometry. This yields more efficient moves which leads to improved sampling, in particular in dense states, such as near native protein conformations [92].
1.5 Methods for inference of protein structures and ensembles

As mentioned in the opening of this thesis the inference of protein structures and protein structural ensembles is pivotal to the analysis of many types of biophysical data.

Since the earliest X-ray crystal diffraction experiments of biomolecules scientists have aimed at building structural models to facilitate an interpretation of the raw data. Approaches to construct these models started with building physical models by hand [93]. Later, computer algorithms took over this task [1]. Over the years these algorithms have become increasingly sophisticated, automating much of the process [94].

This section will introduce the most widely used, existing methodology for the inference of protein structures and protein structural ensembles and discuss their respective assumptions as well as any crucial technical aspects. I put a special focus on methods used in inference from NMR data, although similar approaches are used elsewhere.
1.5. INFERENCE OF PROTEIN STRUCTURES AND ENSEMBLES

1.5.1 General considerations.

Common to all the methods discussed below is the goal to reach a better understanding of the underlying system that gives rise to our data. We saw in section 1.1 that the data are expectations of the corresponding forward model with respect to the Boltzmann distribution of system. I will here outline a few common representations of the Boltzmann ensemble in structure and ensemble inference contexts. I will use $\mathbf{d}$ to represent a set of $N$ experimental observations $\{\bar{d}_i\}_{i=1}^N$ when necessary. Similarly, I will use $\mathbf{r}(x)$ to for the representations of the corresponding datapoints $\{f(x_i)\}_{i=1}^N$.

**Representations.** In the *instantaneous* representation, we assume that the conformational variation is modest and can be approximated by a single state, $x'$, In eq. 1.3 this corresponds to choosing the Boltzmann distribution equal to the Dirac delta function,

$$\bar{d} \approx \int_{x \in \chi} dx \, f(x) \delta(x - x') = f(x').$$  \hfill (1.30)

Similarly we may extend this representation to a discrete set of conformations with their respective weights, $\{x_i, w_i\}_{i=1}^N$, that is, a *discrete ensemble* representation. In this case eq. 1.3 becomes,

$$\bar{d} \approx \sum_{i=1}^N w_i \int_{x \in \chi} dx \, f(x) \delta(x - x_i) = \sum_{i=1}^N w_i f(x_i),$$  \hfill (1.31)

where $\sum_{i=1}^N w_i = 1$. These two representations are common in structure and ensemble inference, respectively. However, these representations intrinsically are approximations. In addition, the forward model may be subject to approximations and the experimental data subject to uncertainty. Consequently, we will not expect a perfect correspondence between the experimental data and the back-calculated data. To overcome this *error models* are used.

**Error models.** Error models consider the discrepancy between experimental data and a back-calculated representation, such as those discussed above. Auxiliary parameters of error models may for instance be an experimental uncertainty or bounds specifying systematic errors. A thorough
discussion of all error models is beyond the scope of this thesis. However, I
will here introduce two commonly used ones, see Figure. 1.8.

The harmonic or Gaussian error model is widely used as it only uses a
single auxiliary parameter, the experimental uncertainty. The probability
of a dataset, \( d \), given a set of representations, \( r(x) \), and an experimental
uncertainty, \( \sigma \) may be expressed as

\[
\mathcal{N}(d \mid r(x), \sigma \mathbf{I}) = \frac{1}{(2\sigma\pi)^{N/2}} \exp \left[ -\frac{1}{2}(d - r(x))^T (\sigma^{-2} \mathbf{I})(d - r(x)) \right].
\]

(1.32)

\( \mathbf{I} \) is the \( N \times N \) identity matrix where \( N \) is the number of data points in \( d \)
and \( r(x) \). This error-model is, however, often used in a unnormalized form
and expressed as its negated logarithm (corresponding to an 'energy'),

\[
E_{\text{harmonic}}(x) = \left[ \frac{(d - r(x))}{\sigma} \right]^2.
\]

(1.33)

Thus, we see the energy is a scaled sum of squares.

Another commonly used error-model is called the square-well potential,
which may be considered as a generalization of the harmonic potential. Here,
bounds define a range of values with the highest probability, unlike the
model above where this value is unique. The error-model is referred to as
a potential, as it is formulated as an 'energy' \((-\log \text{ probability})\) primarily.
The expression for the square-well potential is given by [13],

\[
E_{\text{sqwell}}(x) = w \Delta^e
\]

(1.34)

where \( \Delta \) is given by:

\[
\Delta = \begin{cases} 
    r(x) - (d + d_{\text{plus}} - d_{\text{off}}) & \text{if } d + d_{\text{plus}} - d_{\text{off}} < r(x) \\
    0 & \text{if } d - d_{\text{minus}} < r(x) < d + d_{\text{plus}} - d_{\text{off}} \\
    d - d_{\text{minus}} - r(x) & \text{if } r(x) < d - d_{\text{minus}}
\end{cases}
\]

(1.35)

where \( d_{\text{minus}}, d_{\text{plus}} \) and \( d_{\text{offset}} \) are parameters specifying the bounds of the
flat bottom, and \( e \) is a parameter specifying the steepness of the potential
when outside these bounds. Finally, \( w \) is an overall scale of the potential,
which is \( w = \sigma^{-2} \), above [95]. The square-well potential is typically used
when only bounds of values are obtained from experiments. For instance,
NOEs data is often cast into a number of classes based upon the intensity
of the signal. Subsequently these classes are associated with some distance
bounds.
Figure 1.8: Illustration of two error models: harmonic or Gaussian (dashed) and the square-well potential (solid). In both cases $d = 3$ and the weight is equal to one. For the flat square-well potential $d_{\text{minus}} = d_{\text{plus}} = 1$, $d_{\text{off}} = 0$ and $e = 2$. 
1.5.2 Hybrid energy minimization heuristics

Experimental evidence, as for instance the NOE and RDC restraints discussed above, is always complemented by some prior, as even very dense data-sets have insufficient information to calculate the three-dimensional structures. Classically, this has been achieved by employing empirical force fields $E_{\text{emp}}(x)$ [96, 97], akin to Profasi presented above, to complement an energy derived from the experimental data as introduced in the previous section. We get a hybrid energy

$$E_{\text{hybrid}}(x) = w E_{\text{data}}(x) + E_{\text{emp}}(x) \quad (1.36)$$

where $w$ is the scale parameter which weighs the experimental evidence relative to the empirical forcefield. This general form of the hybrid energy is ubiquitous. However, many different flavors of both $E_{\text{data}}(x)$ and $E_{\text{emp}}(x)$ exists, as discussed previously. In a Bayesian view, this corresponds to having different models for the likelihood ($E_{\text{data}}$) and prior distributions ($E_{\text{emp}}$).

Similar to a likelihood, the data-energy measures the agreement between a candidate representation of a system, and some observed experimental data. For example, in the case of structure determination from NOEs (see section 1.2.1) an experimentally observed NOE is related to an instantaneous pair-wise distance in $x$, through the forward model ISPA. The better the agreement, the lower the error-model energy (or, the higher the probability). The weight, $w$, is typically adjusted such that the agreement of the experimental data is within some expected experimental uncertainty. Nuisance parameters, such as the equilibration constant $\gamma$ in the ISPA forward model of NOE intensities are often fixed to empirical point estimates [8].

Inference from such models is predominantly done through minimization of eq. 1.36. In this context, the hybrid energy is seen as a representation of the free-energy of the protein. The minimization is then a straightforward application of Anfinsens 'thermodynamic hypothesis' where the native state of a protein is thought to be the unique conformation that minimizes the free energy of the system. Again, many methods exist to perform this minimization including Monte Carlo and molecular dynamics method (in both cartesian and angular coordinates) typically combined with a simulated annealing heuristic [98, 13, 99]. During the minimization, scheduling schemes allow for gradual adjustment of auxiliary parameters. Typically, the methods are applied in an iterative manner; adjustments to other parameters such as peak-assignments are carried out in between trial minimizations.
It is common practice to repeat the minimization multiple times to ensure consistent structures are obtained. A set of structures (a bundle) is selected for publication. However, a series of validation and refinement protocols are typically employed before this happens [100, 101, 102, 103].

**Ensemble minimization and refinement** As mentioned at the start, experimental data obtained from biomolecules in solution is often subject to temporal and spatial averaging. Thus the experimental data may be more appropriately represented by a discrete ensemble. This corresponds to changing the representation of the Boltzmann ensemble from an instantaneous one to a weighed average of a set of $N > 1$ conformations. This approach emerged more than twenty years ago and remains popular [104, 105, 106, 107, 108, 109, 50]. The type of hybrid energies used in these approaches may be trivially derived from general considerations discussed above. I will return to the subject of modeling ensembles in section 1.5.4 below.

The energy minimization heuristics have a number of advantages, primarily of a technical or practical nature. The methods are by far the most widely used and consequently have been optimized and tested extensively—in particular single conformer techniques. Additionally, many benchmarks and tests have been carried out comparing the results of different implementations and methods. However, these methods generally suffer from a number of important shortcomings. Firstly, there is no systematic way to deal with the auxiliary parameters such as the weight and the NOE equilibration constant. Consequently, these values are often fixed to empirical point estimates which in addition to empirical bias does not allow for handling of potential uncertainties. In a rigorous application, experimental data should be left out to cross-validate obtained structures and parameters. However, in cases where data is sparse, this may prohibit structure determination all together. Secondly, minimization typically underestimates the degeneracy of the hybrid energy as it inherently aims at finding a unique optimal solution [110, 111, 112]. This in turn makes it difficult to assess the precision and validity of the obtained bundles or ensembles in an objective way [2].

### 1.5.3 Inferential structure determination

Inferential structure determination aims at putting structure determination into a full Bayesian probabilistic framework [2]. That is, to formulate a posterior distribution of structures $\mathbf{x}$ and auxiliary parameters $\mathbf{n}$, given the
observed data $\mathbf{d}$ and prior information $\mathcal{I}$,
\begin{equation}
p(\mathbf{x}, \mathbf{n} \mid \mathbf{d}, \mathcal{I}) \propto p(\mathbf{d} \mid \mathbf{x}, \mathbf{n}, \mathcal{I})\pi(\mathbf{x} \mid \mathcal{I})\pi(\mathbf{n} \mid \mathcal{I}). \tag{1.37}
\end{equation}
If we take the negated logarithm of this posterior probability distribution we immediately obtain an expression analogous to the hybrid energy shown in eq. 1.36 – however, an additional term emerges that accounts for the prior of the auxiliary parameters.

As an illustration I will here give a the posterior distribution for structure inference from NOEs. The likelihood is a combination of a log-normal distribution (the distribution of a variable whose log is Normal distributed) to model the error of distances, or intensities, and an instantaneous representation using the ISPA forward model. This introduces two auxiliary, or nuisance, parameters: the equilibration parameter $\gamma$ and experimental uncertainty $\sigma$. We have, $p(\mathbf{d} \mid \mathbf{x}, \mathbf{n}, \mathcal{I}) = \log \mathcal{N}(\mathbf{d} \mid f(\mathbf{x}, \gamma), \sigma)$, where the function $f(\cdot)$ is ISPA and $\mathbf{n} = \{\gamma, \sigma\}$. Now we only need to specify the priors of the parameter-space of $\mathbf{x}, \gamma$ and $\sigma$. For $\mathbf{x}$ we may chose any structural prior, for instance one of those discussed above. Similarly, for $\gamma$ and $\sigma$ we may chose any suitable prior. In ISD the simple improper priors $\pi(\gamma) \propto \gamma^{-1}$ and $\pi(\sigma) \propto \sigma^{-1}$ were used. The priors on the nuisance parameters put special emphasis on small values and are defined on $\mathbb{R}^+$ only, which is compliant with the domain of the parameters. Both of these priors are referred to as Jeffreys priors in original manuscript.

Like hybrid energies, the negative log of the posterior may also be subject to minimization. This was done on a large-scale recently [113]. However, carrying out the minimization in this case will have similar drawbacks as those outlined above. Instead, Nilges and co-workers carried out statistical inference by sampling from the posterior distribution using extended ensemble MCMC techniques [114, 2]. In this manner, marginal posterior probability distributions are obtained directly. These allows for direct objective assessment of validity, precision and quality of the obtained structures, through well-established statistical tests. Similarly, may statistical assessments regarding the nuisance parameters $\sigma$ and $\gamma$ be made. Such properties have long been lacking in the protein structure determination from NMR data.

1.5.4 Restrained molecular simulations

The methodology discussed above focuses on refining single structure models to agree with experimental data while fulfilling basic prior knowledge of biomolecules. This may be a reasonable first approximation for some systems, however, it will be too crude for other, perhaps even the majority
of cases. In fact, the motion in biomolecules has long been recognized to be one of the key sources of systematic errors in models subject to such assumptions, as touched upon above [4, 1, 115].

In other contexts, methods have been developed to directly infer the probability densities of the geometrical properties represented in experimental observations [116, 117] – however, these require that complete information of all degrees of freedom is represented in the data. This is very rare when biological macromolecules are considered. I will here discuss an alternative approach where the incomplete information from experimental observations is complemented by structural prior information as embodied either in forcefields or knowledge-based potentials. Similar to the structure determination and refinement techniques above hybrid energy potentials are constructed as to ensure agreement with experimental data. However, instead of minimization the goal is to simulate samples from the underlying distribution and thus obtaining a representation of the accessible conformational states [118, 119, 64, 63].

For completeness, I will give an example of a data energy where an $N$-conformer representation with uniform weights is combined with a reduced harmonic error-model,

$$E_{\text{mc-data}}(\mathbf{x}) = \left[ \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}_i, \theta_i) - \mathbf{d} \right]^2 \sigma^2$$

with $\theta$ being the nuisance parameters associated with the forward model $f(\cdot)$. Note, it is common to multiply this restraining term by $N$ to keep the magnitude of the restraining force approximately independent of $N$ [120].

Recently, Pitera and Chodera presented a different approach to restraining molecular simulations through the principle of maximum entropy [120]. In this manner, they obtained a potential which takes the ensemble nature of the data into account, without introducing the notion of multiple replicas. For a single experimental observation without experimental uncertainty, they arrived at the hybrid energy,

$$E_{\text{hybrid-\text{pc}}}(\mathbf{x}) = \alpha f(\mathbf{x}) + E_{\text{emp}}(\mathbf{x})$$

where $\alpha$ is a constant which may be estimated such that the expectation of $f(\mathbf{x})$ according to $\frac{\exp(-\beta E_{\text{hybrid-\text{pc}}}(\mathbf{x}))}{Z}$ is equal to the experimental data $\mathbf{d}$. They describe a procedure to achieve this in their manuscript. They further demonstrate that the method works well on a one dimensional example in absence of experimental uncertainty. Finally, they show that their potential...
is closely related to eq. 1.38 in the limit where $N$ goes to infinity. A similar result was recently reported by others [121].

Another approach represents the data by an average of structures along the simulation trajectory [122, 123]. However, this approach has a number of intrinsic limitations and a thorough discussion is thus beyond the scope here.

### 1.5.5 Post hoc ensemble selection

The final class of methods I will consider here are the *post hoc* ensemble selection techniques. The methods emerged less than ten years ago and has since then been used predominantly to refine ensembles obtained from knowledge-based statistical coil libraries to agree with experimental data of intrinsically disordered proteins [124, 65]. In these methods, a pool of $K$ structures is sampled using either statistical coil models or high-temperature molecular dynamics simulations. Subsequently, either the pool is weighed [125], a subset of the pool is selected [126, 127] or a combination of these (Olsson, Unpublished work). This is done to optimize the agreement with experimental data. Through an increasing number of auxiliary heuristic modifications of forward models [128], these methods now provide excellent agreements to experimental data using a relatively small number of conformations [129, 130].

### 1.5.6 Analysis of methodology

In the previous section I’ve given a brief outline of common methods in structure and structural ensemble determination. This final part is devoted to comparing and analyzing the methods in more detail.

**Reference ratio distribution** In all of the methodologies introduced above, the aim is to infer conformational models in atomic detail. However, the experimental data we observe only reflect some projection of the actual underlying distribution of conformations. Consequently, when we use conformational prior information to account for the incomplete information, we also induce prior on the manifold of the data (section 1.3.5). This prior may in some cases be both strong and inappropriate, and can have a strong negative influence on the posterior distribution. This topic is discussed further in section 2.2.
1.5. INFERENCE OF PROTEIN STRUCTURES AND ENSEMBLES

Structure inference  There are three key assumptions common to all the methods for protein structure determination presented here. Firstly, the forward-model accurately describes the relationship between a conformational state and its contribution to the experimental data. Secondly, the single conformer representation of the Boltzmann distribution is a good approximation. Finally, all experimental observations are uncorrelated and have equal uncertainty. While the two first assumptions have been widely discussed in the literature and consequently also in this thesis, the final one has not.

It is often convenient to assume that experimental data are uncorrelated and have equal variance. We assume this about data used for protein structure inference. This assumption is most apparent if we consider the harmonic error-model eq. 1.32. Here, we chose a diagonal covariance matrix with one variance $\sigma^2$ common to all diagonal elements. This in turn reduces to a single weighing factor of the sum of square distances between the forward-model and the experimental data. This is completely analogous in the square-well and other closely related potentials. Following the Gauss-Markov theorem this will result in the best linear unbiased estimate if and only if the data indeed fulfills these assumptions [131]. It is reasonable to suggest that this is not the case: for instance, sets of NOEs observed between adjacent atomic pairs may be highly correlated. The potential consequence of this includes over-confident wrong structural models.

It appears, however, that such problems have been overcome in practice by using strong conformational priors in subsequent refinement procedures [100]. In spite of this, it might be an interesting future endeavor to investigate if proper handling of the correlation between observations improves the quality of structural models and perhaps alleviates the need for subsequent processing. This may be achieved by replacing the diagonal matrix, $(\sigma^{-2}I)$, from eq. 1.32 with a full covariance matrix.

Unlike the minimization heuristics, inference through sampling of the posterior distribution was reported to be very computationally intensive – it requires a computer clusters of 50 computers to run for a number of days. Since such computational demands are excessive for most experimental labs, the application of ISD has been limited. Consequently, a more efficient implementation may stimulate more experimental labs to adapt this more rigorous methodology. We have proposed a new implementation, see section 2.1.
Restrained simulations In addition to the assumption made above concerning the accuracy of the forward-model similar assumptions regarding the representation and lack of correlations between observations.

It has been suggested that the discrete representation of the conformational distribution is sufficiently accurate to capture anisotropic motion, in particular when only very fast motions are represented in the data [132]. However, a recent paper demonstrated that measurable bias is introduced when discrete ensembles are employed [121]. This may to a large extend be alleviated by using a large number of conformations in the representation. However, it remains an inherent problem that the number of replicas need to be estimated – and most of all there is currently no systematic way to make this decision or account for its uncertainty. While cross-validation may provide a sensible approach to make this assessment it is generally undesirable for reasons mentioned above. This problem may become more apparent with increasingly accurate and dense datasets of experimental data averaged over long time scales, as for instance residual dipolar couplings. Similarly, when intrinsically disordered proteins are studied, the anisotropic nature of the conformational ensemble may become quite significant which may require the number of conformations necessary to become computationally intractable. Consequently, the need for methods that allow restraining these simulations which are independent upon discretizations is desirable.

Cavalli and co-workers provided a nice proof of the validity of the using experimental uncertainties in the cases of multiple replica restraining potential [121]. However, the replica independent maximum entropy formulation shown by Pitera and Codera as well as Cavalli and co-workers does not explicitly handle experimental uncertainty. Consequently, any uncertainty in the data will affect the estimated $\alpha$ or $\lambda$ parameters. In section 2.2 we discuss one possibility of how to overcome this shortcoming.

In multiple replica restrained simulation we assume, as in structure determination, the observations to be uncorrelated. This is not immediately obvious, however, it becomes evident if we consider the continuous case presented by Pitera and Chodera for $N$ datapoints. In that case, a parameter $\alpha_i$ for each representation $f_i(x)$ of a datapoint $i$ is introduced. We get,

$$E_{\text{hybrid–pc2}}(x) = E_{\text{emp}}(x) + \sum_{i=1}^{N} \alpha_i f_i(x)$$

(1.40)

this is equivalent to

$$p(x) \propto \exp(-E_{\text{emp}}(x)) \prod_{i=1}^{N} \exp(-\alpha_i f_i(x))$$

(1.41)
which in statistics corresponds to a generalized linear model of exponential response data also called a log-linear model [133]. The form of the model given in eq. 1.41 assumes the variables $f_i(\cdot)$ to be uncorrelated. As in structure determination, this assumptions will have an effect on the results obtained – it is, however, unclear how severe this will be. Still, the current literature does address such potential shortcomings.

Post hoc ensemble selection This class of methods has had much success, in particular for determining ensembles of intrinsically disordered proteins. However, in addition to the assumptions discussed above they have a number of undesirable properties. First, the resulting explicit ensembles from these methods are not unique and easily complement each other. Second, since the sampled pool is typically obtained from rather crude models it is not always clear whether selected conformations represent any physical realism or simply optimize the fit to the experimental data. These shortcomings in general limit the practical use of explicit ensemble models.

These ensemble weighing or selection techniques may however be useful for averaged data from approximately discrete systems. For instance mixtures of monomeric and oligomeric forms of proteins. The data available for such systems is often rather coarse-grained, eg. small angle X-ray scattering. Consequently, in such cases using atomic-detail molecular simulations may not only be computationally expensive but also excessive.

Throughout this introduction I have mapped out different fields and their methodologies relevant to inference problems in structural biology from NMR data. In this final section I have discussed out some of the current problems and shortcomings in existing methodology – the research papers presented next addresses some of these.
CHAPTER 1. GENERAL INTRODUCTION
Cited Literature


CITED LITERATURE


[34] Uversky, V. N. Biochim Biophys Acta Dec (2012).


53


2.1 Generative probabilistic models increase the scope of inferential structural determination

The appearance of inferential structure determination (ISD) displayed a significant step towards rigorous, automated biomolecular structure determination. It improved the quality of NMR structures, and made structure determination possible in sparse data cases, where conventional methods failed. However, the prohibiting computational demands failed to make ISD the method of choice amongst structural biologists.

This paper describes an improved implementation of ISD drawing on a new sampling scheme and generative probabilistic models of protein backbone and side chain conformational space. The generative probabilistic models provide additional information, which significantly reduces the search space. The sampling scheme allows for efficient traversal of the conformational space.

We demonstrate that the extension provide improved convergence rates along with improved structural precision in very sparse data cases. The improved convergence rates in particular are of interest, as ISD becomes possible on singular desktop computers.

This is a first author paper and I was involved in all aspects of the work carried out. The paper was published in Journal of Magnetic Resonance in 2011.
Communication

Generative probabilistic models extend the scope of inferential structure determination

Simon Olsson a, Wouter Boomsma b, Jes Frellsen a, Sandro Bottaro b, Tim Harder a, Jesper Ferkinghoff-Borg b,*, Thomas Hamelryck a,*

a Bioinformatics Center, University of Copenhagen, Department of Biology, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark
b Biomedical Engineering, DTU Elektro, Technical University of Denmark, Ørsteds Plads, DK-2800 Kgs. Lyngby, Denmark

1. Introduction

Current methods for macromolecular structure determination rely on the seminal idea of hybrid energy minimization introduced by Jack and Levitt [1]. However, the choice of model parameters, such as the weight of the experimental data with respect to a physical force field, is intrinsically problematic in this approach – a fact that was already recognized in the original study [1]. With a growing number of sources of experimental data used in protein structure determination, estimation of weights and other nuisance parameters is becoming increasingly problematic. Current methodology relies on a more or less arbitrary choice of these parameters, using heuristic approaches [2]. While a persistent concern towards the applied heuristics has been evident in the literature [3,4], only few quantitative methods have been described to rigorously determine these nuisance parameters [4,5]. These methods, and the underlying Bayesian approach are referred to as inferential structure determination (ISD).

Bayesian probabilistic inference has previously shown great potential in macromolecular structure determination [2,6]. However, the scope of the approach has been limited due to excessive computational demands. The current study describes a new approach to inferential structure determination which draws on the use of generative probabilistic models. Generative probabilistic models, or GPMs, are probabilistic models that allow sampling. Here, we demonstrate that the use of GPMs greatly increases efficiency, precision and scope of rigorous inferential structure determination. As these GPMs contain information about protein structure, they may supersede physical forcefields – especially in cases where data is very sparse.

2. Methods

In the ISD approach, samples are drawn from a joint posterior distribution over conformational space, X, and model parameter space, n, given experimental data, D, and prior knowledge, I:

$$p(X, n|D, I) = p(D|X, n, I)p(n|I)p(X|I).$$

Consequently, a natural result of posterior sampling is an ensemble of conformers representing the experimental uncertainty. That is, the Bayesian formalism accounts for uncertainty and degeneracy, a feature that is difficult to obtain when using schemes that minimize a hybrid energy consisting of a physical and a data-dependent term [7,8].

In ISD, a physical forcefield $E_{phys}$ enters the Bayesian framework as a conformational prior through a canonical ensemble

$$p(X|I) \propto e^{-\beta E_{phys}},$$

where $\beta = 1/kT$, $k$ is Boltzmann's constant and $T$ is the temperature [2]. The data enters as a likelihood function, $p(D|X, n, I)$; its product with the prior distributions, $p(n|I)p(X|I)$,
results in the posterior distribution, \( p(X, n|D, I) \). When the posterior is defined in this way, Markov Chain Monte Carlo (MCMC) sampling requires evaluation of both likelihood and priors explicitly, in each step. This can potentially lead to substantial computational costs. Conversely, using no, or a uninformative forcefield, leaves a vast conformational space [9]. Here, we use GPMs of local protein structure instead of the Boltzmann distribution of a physical forcefield. Consequently, we demonstrate that the explicit evaluation of the prior can be avoided altogether as the information of the prior enters the posterior distribution through sampling.

Recently, our group has published several GPMs of protein conformational space, describing backbone (TorusDBN) [10,11] and sidechain (Basilisk) [12] dihedral angles. These models only provide structural information on a local sequential scale, ideally complementing the long-range information obtained from NMR nuclear Overhauser enhancements experiments (NOE). As generalizations of the commonly used fragment- [13] and rotamer-libraries [14], and related potentials that involve discretization [15], these GPMs also serve to reduce the complexity of the conformational space. The particular GPMs applied here use continuous angular probability distributions to avoid the intrinsic limitations caused by discretization [16]. Furthermore, since these GPMs are probability distributions, probabilities of arbitrary conformations can be evaluated, which is not generally possible for fragment- and rotamer libraries. Consequently, the full posterior probability can be evaluated explicitly when necessary. Here, we demonstrate that the use of GPMs as conformational proposal distributions can dramatically increase convergence in MCMC sampling of protein conformers from a posterior distribution, in addition to providing an increase in precision.

The GPMs, TorusDBN and Basilisk, enter the ISD approach as \( p(X|I) \times p(b|n)p(\chi|a) \), where \( a \) denotes amino acid sequence, while \( b \) and \( \chi \) denote backbone and sidechain conformations respectively. Thus, during simulation we alternate between moving in backbone and sidechain conformational space, conditioned on amino acid sequence. Following Rieping et al. we assume idealized Engh–Huber bond lengths [17] and parameterize conformations as sets of torsion angles [18]. Variations in the bond angles were allowed to facilitate conformational sampling [19].

We used a generalized ensemble Metropolis–Hastings sampling scheme to draw samples from the posterior distribution. To prioritize search in relevant regions of the conformational space we adopted the 1/\( k \)-ensemble implemented using the generalized multi-histogram equations [20,21]. The 1/\( k \)-ensemble allows sampling independently of temperature, thus avoiding nuisance parameters such as the number of replicas, and their temperature span. It is, however, important to stress that the statistical information provided by this sampling scheme is equivalent to the Replica Exchange Monte Carlo scheme used in the original ISD study [22]. We employ the log-normal formulation of the NOE data to evaluate \( p(D|b, \chi, n, I) \), as this provides the least biased formulation of the likelihood [5].

To assess the performance of TorusDBN and Basilisk as conformational priors, and for comparison to previous results, we created a set of conformers corresponding to the lowest posterior samples, using the very sparse (154 constraints) SH3 FYN domain data [2] and the TRP-Cage data set [28]. As a model baseline, we carried out the same simulations without the models of local protein structure. This simple hard-sphere potential corresponds to the use of a prior distribution reminiscent of that of the original ISD implementation [2].

2.1. Posterior sampling

As described previously, we sample from the joint posterior distribution \( p(X, n|D, I) \) [2]:

\[
p(X, n|D, I) \propto \sigma^{-n-1} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \gamma^2(d_i^2 - \mu^2) \right) p(\gamma|a)p(b|a),
\]

with the log-normal chi-square: \( \gamma^2(d, D) = \sum_i \log^2\left( \frac{\gamma d_i^2}{D_i} \right) \), \( D_i \) are experimental data and \( d_i \) calculated distances [5]. \( \gamma \) and \( \sigma \) are ISPA (isolated spin–pair approximation) equilibrium parameter and experimental uncertainty, respectively. A power \( \alpha = -1 \) was used here as all data were derived distances.

Here, we cannot employ the Gibbs sampling scheme applied in Rieping et al. [2], due to the inherent absence of an explicit temperature in the 1/\( k \) ensemble. This absence of an explicit temperature makes the implementation of the soft-sphere potential employed previously difficult without introduction of additional heuristics, and was therefore avoided [2]. Instead, we here use a Metropolis–Hastings approach, where the involved parameters are updated one at the time. The 1/\( k \) ensemble allows us to sample the conformational- and nuisance-space efficiently.

Low acceptance rates in the nuisance sampling was avoided by introducing a scheme exploiting the information about the current state. For the nuisance parameters, \( n = \{\gamma, \sigma\} \), a log-change is proposed from a log-normal distribution with a standard deviation \( \sigma_n = \frac{1.0}{\max\left(\left\|\log p(X,n|D, I)\right\|_{\text{obs}}\right.1.0)} \).

This expression was derived using standard error propagation and adds a simple regularizer which ensures a maximum standard deviation of 1.0 [23]. As a result, we can draw samples efficiently from the joint posterior distribution without the temperature dependent Gibbs sampling scheme. Using the log-normal distribution in this way we can ensure being in the right domain. We avoid additional bias from the log-normal distribution in the posterior, by dividing out the bias in the Monte Carlo acceptance ratio. For completeness, the analytical expressions of the standard deviations are shown here:

\[
\sigma_\gamma = \frac{1.0}{\max\left(\left\|\frac{\log p(X,n|D, I)}{\gamma_{\text{obs}}}\right\|_{\text{obs}}1.0\right)},
\]

where \( \gamma \) is the number of datapoints, and:

\[
\sigma_\gamma = \frac{1.0}{\max\left(\left\|\frac{\log p(X,n|D, I)}{\gamma_{\text{obs}}}\right\|_{\text{obs}}1.0\right)},
\]

with \( k_i = \ln \frac{\gamma_{\text{obs}}}{\gamma_{\text{obs}}} \) corresponding to the log-ratio between the observed and back-calculated experimental data.

For sampling of the conformational space, a series of MCMC moves for backbone (pivot, local [19] and semi-local [24]) and sidechain conformations were employed. All applied moves fulfill detailed balance, and were chosen with even probability with respect to backbone and sidechain conformational and nuisance space. TorusDBN was extended to account for small deviations from ideal cis/trans-angles, using a normal distribution with mean at the ideal values and a standard deviation of five degrees. In the baseline model, all angles \( b, \chi \) were sampled uniformly in the interval [0,2\pi]. Note that Basilisk was used in a backbone independent fashion for simplicity [12]. Samples were accepted or rejected according to the generalized 1/\( k \) ensemble [20]. Convergence was assessed through inspection of diagnostics provided by Muninn: the multi-histogram implementation of the generalized ensemble (http://www.muninn.sourceforge.net/). It is important to stress that convergence of histograms necessarily reflect convergence of posterior samples, additional sampling allow generation of more refined ensembles.
3. Results

3.1. SH3 FYN

When employing the GPMs, the sampling of the posterior distribution defined by the sparse SH3 FYN data set converges in less than 36 h of computation time on a single standard CPU core. In comparison, the previously published ISD ensemble derived from the same data set took 3 days on a 50 core computer cluster [2]. Even given the increase in average computational power since 2005, this is a substantial increase in the efficiency. We do not observe convergence within the same simulation time when applying the baseline model. This illustrates clearly how the GPMs increase efficiency of posterior sampling.

Performing posterior sampling with the baseline prior, gives rise to two distinct conformational basins (Fig. 1b). There is an excited basin corresponding to the mirror image of the native basin. The local geometry of this basin is highly unfavorable. The second basin corresponds to the correct, native fold, observed in the crystal structure. The latter of the two basins is the only one observed when using the informative GPMs as conformational priors (Fig. 1a). Evidently, the experimental data likelihood in conjunction with the baseline prior only modestly distinguishes between the two folds, resulting in slow convergence due to an excessive conformational multiplicity. The basin with the correct fold is not thoroughly explored within the given time frame, resulting in relatively inaccurate structures among the 20 highest posterior conformer ensemble (Fig. 2b). In contrast, the ensembles obtained within the same simulation time using the TorusDBN and Basilisk priors accurately capture the native state (Fig. 2a). This result illustrates the importance of prior information to resolve degeneracies in sparse experimental data. While avoidance of poor

<table>
<thead>
<tr>
<th>Dihedral prior</th>
<th>VADAR</th>
<th>PROCHECK</th>
</tr>
</thead>
<tbody>
<tr>
<td>(φ, ω) core</td>
<td>68.95 ± 4.19%</td>
<td>88.33 ± 2.85%</td>
</tr>
<tr>
<td>(φ, ω) allowed</td>
<td>27.6 ± 4.12%</td>
<td>9.96 ± 3.17%</td>
</tr>
<tr>
<td>(φ, ω) generous</td>
<td>1.7 ± 1.27%</td>
<td>1.65 ± 1.50%</td>
</tr>
<tr>
<td>ω outside</td>
<td>0.0 ± 0.0%</td>
<td>0.05 ± 0.0%</td>
</tr>
<tr>
<td>ω core</td>
<td>100.0 ± 0.0%</td>
<td>91.0 ± 2.17%</td>
</tr>
<tr>
<td>ω allowed</td>
<td>0.0 ± 0.0%</td>
<td>8.0 ± 2.61%</td>
</tr>
<tr>
<td>ω generous</td>
<td>0.0 ± 0.0%</td>
<td>1.0 ± 1.49%</td>
</tr>
</tbody>
</table>

Packing defects | 11.95 ± 2.85 | 5.95 ± 2.06 |
Free energy fold | -40.7 ± 1.88 | -46.07 ± 2.06 |
Res. 95% buried | 2.25 ± 1.22 | 4.30 ± 1.90 |
Buried charges | 0.15 ± 0.39 | 0.30 ± 0.56 |

G-factor (φ, ω) 1ZBJ GPMs |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G-factor (φ, ω)</td>
<td>1.41</td>
<td>–0.72</td>
</tr>
<tr>
<td>G-factor (φ only)</td>
<td>–1.82</td>
<td>0.25</td>
</tr>
<tr>
<td>G-factor (ω only)</td>
<td>–0.54</td>
<td>0.20</td>
</tr>
<tr>
<td>G-factor (overall)</td>
<td>–1.43</td>
<td>–0.28</td>
</tr>
</tbody>
</table>

Table 1

VADAR and PROCHECK structure quality statistics for the previously published ensemble (PDB: 1ZBJ) [2] and current SH3 FYN (GPMs) ensembles and reference values presented by VADAR (Ref). φ, ω core, allowed, generous and outside denote distinct regions of the Ramachandran plot of decreasing favoredness. ω core denotes the percentage of ω-angles in the most favored region (the three other classes are not shown here). Packing defects, free energy folding, percentage of residues 95% buried and buried charges denotes the number of packing defects, free energy of folding and bury ratios for residues and charges, respectively [25]. Percentile reference values were normalized. PROCHECK G-factors reflect average log-odds of (φ, ω), (χ1, χ2), (χ1) and overall dihedral angle combinations.

Fig. 1. Scatter plots of the RMSD of conformational samples to the crystal structure of SH3 FYN (PDB:1SHF chain A) versus −log(p(X,n,D,l)) (posterior density) for (a) TorusDBN and Basilisk and (b) the baseline prior after 400 million MCMC steps. Samples are from the 1/k ensemble.

Fig. 2. Illustration of 20 of the samples with the highest posterior probability using (a) TorusDBN and Basilisk (RMSD: 1.74 ± 0.17Å) or (b) the baseline prior (RMSD: 3.12 ± 0.24 Å), after 400 million MCMC steps. Conformations are aligned to PDB: 1SHF chain A (shown in a black cartoon representation). Figure prepared using PyMOL (DeLano Scientific LLC).
stereochemistry has been pointed out previously as a feature of the ISD approach [2], degeneracy due to poor local structure has remained unaddressed.

The mean heavy-atom (Cα, C and N) root mean square deviation (RMSD) to the crystal structure from the 20 highest posterior probability structures (see Fig. 2) is comparable to the previously published ISD ensemble (1.84 ± 0.20 Å, PDB: 1ZBJ). However, statistics derived from structure validation server VADAR [25], WHATIF [26] and PROCHECK [27] were vastly improved (see Table 1 and Supplementary material) with respect to both packing quality and local structure. Importantly, clustering of (ϕ,ψ)-angle pairs in less favorable regions of the Ramachandran space is reduced dramatically (see SI). Other structure quality indicators such as number of buried charges remain unchanged. While the improvement in local

![Fig. 3](image1.png)

**Fig. 3.** Scatter plots of RMSD of conformational samples to the previously published NMR structure of TRP-Cage (PDB: 1L2Y) versus −log(p(X,n,D,I)) (posterior density) for (a) TorusDBN and Basilisk and (b) baseline prior after 50 million MCMC steps; (c) baseline prior after 500 million MCMC steps. Samples are from the 1/k ensemble.

![Fig. 4](image2.png)

**Fig. 4.** Illustration of 20 of the samples with the highest posterior probability using (a) TorusDBN and Basilisk (RMSD: 0.63 ± 0.12 Å) or (b) baseline prior (RMSD: 1.41 ± 0.39 Å), after 50 million MCMC samples. Conformations are aligned to PDB: 1L2Y (shown in a black cartoon representation). Figure prepared using PyMOL (DeLano Scientific LLC).
structure is an expected consequence of the information contained in TorusDBN and Basilisk, non-local structure quality parameters such as packing defects hint increase in accuracy. The nuisance parameters $\sigma$, $\gamma$ were estimated to be 0.11 $\pm$ 0.01 and 1.00 $\pm$ 0.01, respectively. These values deviate somewhat from the estimates obtained previously. The discrepancy may be linked to a different conformational prior distribution [2].

3.2. TRP-cage

In addition to increasing efficiency and precision, GPMs can account for the information derived from ambiguous NOE constraints. We demonstrate this point on the TRP-cage data set [28]. Of the reported 169 restraints, 37 involve pseudo atoms, which strictly speaking yields them ambiguous. In these particular calculations, the restraints were therefore not included. The resulting set of unambiguous NOE restraints are insufficiently informative to distinguish native-like structures from conformers with an RMSD of up to 3 Å from the previously published NMR structure. However, when we use the GPMs as structural priors, we obtain an ensemble of high resemblance with the previously published structure.

The simulations of TRP-cage were performed identically to those of SH3 FYN using 50 million MCMC steps. Both simulations complete within a few hours (see Fig. 3a and b). The pattern observed for SH3 FYN emerges again: when using GPMs convergence was reached within the simulation time, whereas convergence was not reached using the baseline model. Extending the simulation time with the baseline model to 500 million MCMC steps results in convergence (Fig. 3c). However, the resulting 20 highest posterior ensemble is of significantly lower quality (RMSD: 1.24 ± 0.39 Å) than the ensemble obtained using the GPMs running for 50 million MCMC steps, Fig. 4a. With these results we again demonstrate how efficiency is gained when employing GPMs in the ISD approach. In addition the results illustrate, how the unambiguous constraints [28] can be complemented by the local information contained in the GPMs.

4. Conclusions

In both examples presented here, the difference in accuracy of the selected ensembles is modest, with mean RMSD differences of at most 1 Å. However, the highest probability (or lowest energy) criterion for selection of conformation for these ensembles may not only underestimate the spread of the ensemble [29, 30], but also ignore severe degeneracies (see Figs. 1 and 3). This points to the importance of using appropriate prior information when analyzing sparse data and suggests extra caution be taken when selecting these ensembles.

This communication describes how generative probabilistic models can be applied to significantly increase efficiency and precision of inferential structure determination. As a natural extension, we propose the development of more specialized GPMs, drawing on additional prior information such as protein family membership or chemical shifts. Such models would presumably resolve degeneracies to an even greater extent, further increasing the scope, efficiency and precision of the inferential structure determination approach.

Acknowledgments

We thank F.M. Poulsen, K. Lindoff-Larsen and J.H. Jensen for critically reading the manuscript. T. Harder is funded by the Danish Council for Strategic Research (NaBIT, 2106-06-009). W. Boomsma is funded by the Danish Council for Independent Research (FNU, 272-08-0315). J. Frellsen and S. Olsson are funded by the Danish Council for Independent Research (FTP, 09-06564). S. Bottaro acknowledge funding from Radiometer, DTU.Elektro.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmr.2011.08.039.

References

2.1. EXPANDED SCOPE OF ISD

Supplementary material for Generative probabilistic models extend the scope of inferential structure determination

Simon Olsson
Bioinformatics Center, University of Copenhagen, Department of Biology, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark

Wouter Boomsma
Biomedical Engineering, DTU Elektro, Technical University of Denmark, Ørsted Plads, DK-2800 Kgs. Lyngby, Denmark

Jes Frellsen
Bioinformatics Center, University of Copenhagen, Department of Biology, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark

Sandro Bottaro
Biomedical Engineering, DTU Elektro, Technical University of Denmark, Ørsted Plads, DK-2800 Kgs. Lyngby, Denmark

Tim Harder
Bioinformatics Center, University of Copenhagen, Department of Biology, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark

Jesper Ferkinghoff-Borg
Biomedical Engineering, DTU Elektro, Technical University of Denmark, Ørsted Plads, DK-2800 Kgs. Lyngby, Denmark

Thomas Hamelryck
Bioinformatics Center, University of Copenhagen, Department of Biology, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark

Email addresses: jfb@elektro.dtu.dk (Jesper Ferkinghoff-Borg), thamelry@binf.ku.dk (Thomas Hamelryck)

Preprint submitted to Journal of Magnetic Resonance August 17, 2011
1. Posterior sampling details

In addition to bond lengths, ring planarities are kept fixed. These are reasonable approximations at room temperature. Furthermore, to avoid steric overlap we impose a hard sphere volume exclusion potential with conservative distance cutoffs at 1.3Å for $H - H$ pairs, 1.65Å for $N - O$ pairs and 1.5Å for any other pair of atoms. No clashing conformers were considered. All calculations were carried out using the Phaistos framework (http://www.phaistos.org) [1]. Support for NOE constraints will be added to the Phaistos package in the near future. The conformers for the final ensembles were selected by reweighing the samples from the $1/k$ ensemble according to the posterior distribution and selecting the 20 most likely conformers.

2. Quantitative Assessment of Structure Quality

During sampling the conformational priors, TorusDBN and Basilisk are generally not evaluated explicitly, e.g. $P(X|I)$ is only used as a proposal distribution of the conformational space, without accounting for sampling bias in the Monte-Carlo acceptance ratio. Thus, the likelihood function $p(D|X, I)$ is the only explicitly evaluated term in each evaluation step of the simulation. Posterior probabilities are determined by multiplying prior probabilities evaluated from TorusDBN and Basilisk with the likelihood of samples obtained from the simulation. All results were based on samples saved with a regular interval of 10000 MCMC steps during simulation, for which the full posterior was used.
Table 1: WHATIF Structure quality statistics for the previously published ensemble (PDB: 1ZBJ) (1ZBJ) [2] and current SH3 FYN (GPMs) ensembles and previously published SH3 FYN crystal structure (pdb: 1SHF chain A) (1SHF). Positive values indicate better than average of a set of high quality structures. [3]

<table>
<thead>
<tr>
<th>WHATIF</th>
<th>1ZBJ</th>
<th>GPMs</th>
<th>1SHF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dihedral prior</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st generation packing quality</td>
<td>-7.43 (bad)</td>
<td>-4.550 (poor)</td>
<td>-0.777</td>
</tr>
<tr>
<td>2nd generation packing quality</td>
<td>-4.984 (bad)</td>
<td>-3.015 (poor)</td>
<td>3.847</td>
</tr>
<tr>
<td>Ramachandran plot</td>
<td>-5.667 (bad)</td>
<td>-2.573</td>
<td>-0.310</td>
</tr>
<tr>
<td>Rotamer normality</td>
<td>-7.668 (bad)</td>
<td>0.405</td>
<td>-1.910</td>
</tr>
<tr>
<td>Backbone conformation</td>
<td>-1.049</td>
<td>-1.462</td>
<td>-1.197</td>
</tr>
<tr>
<td>bond length</td>
<td>0.636 (tight)</td>
<td>0.321 (tight)</td>
<td>1.180</td>
</tr>
<tr>
<td>bond angles</td>
<td>0.983</td>
<td>0.951</td>
<td>1.605</td>
</tr>
<tr>
<td>omega angle restraints</td>
<td>0.000 (tight)</td>
<td>0.899</td>
<td>0.961</td>
</tr>
<tr>
<td>side chain planarity</td>
<td>0.038 (tight)</td>
<td>0.023 (tight)</td>
<td>4.446 (loose)</td>
</tr>
<tr>
<td>Improper dihedral distribution</td>
<td>0.754</td>
<td>0.621</td>
<td>2.696 (loose)</td>
</tr>
<tr>
<td>Inside/outside distribution</td>
<td>1.166 (unusual)</td>
<td>1.089</td>
<td>1.028</td>
</tr>
</tbody>
</table>

3. SH3 FYN Ramachandran plots

4. SH3 FYN WHATCHECK/WHATIF

References

Figure 1: Ramachandran plot of the 20 highest posterior conformers a) obtained using TorusDBN and Basilisk as proposal distributions b) previously published (PDB:1ZBJ) [2]. Figures generated by the VADAR webservice (http://redpoll.pharmacy.ualberta.ca/vadar/)
2.1. EXPANDED SCOPE OF ISD


2.2 Inferential determination of protein structural ensembles from sparse, averaged data.

The determination of ensembles of structures which gives rise to a particular spatially and temporally averaged experimental signal continues to be an important problem. We here approach the problem from a probabilistic perspective, and obtain a hierarchical model which contains previously proposed methods as intuitive limits.

To illustrate the model is tractable, we apply the model to two dataset of simulated nuclear Overhauser enhancements. We find that we are able to accurately reconstruct the probability distributions of the geometrical features reflected in the data.

This is a first author paper and I was involved in all aspects of the work carried out. Submitted for publication.
Inference of protein structure ensembles from sparse, averaged data.

Simon Olsson$^1\ast$, Jes Frellsen$^1$, Wouter Boomsma$^2$, Kanti V. Mardia$^3$ and Thomas Hamelryck$^1\ast$

1 Bioinformatics Centre, Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark
2 Structural Biology and NMR Laboratory, Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark
3 Department of Statistics, School of Mathematics, University of Leeds, Leeds, United Kingdom.

* E-mail: Corresponding solsson@binf.ku.dk, thamelry@binf.ku.dk

Abstract

We present a general principle to infer protein structure ensembles from spatially and temporally averaged data obtained in biophysical experiments. The central idea is to compute the Kullback-Leibler optimal modification of a given prior distribution $\tau(x)$ with respect to the experimental data and its uncertainty. This principle generalizes the successful inferential structure determination method and recently proposed maximum entropy methods. Tractability of the protocol is demonstrated through the analysis of simulated nuclear magnetic resonance spectroscopy data of a small peptide.

Introduction

The rigorous analysis of experimental data probing the structure of biological macromolecules forms the foundation of many biophysical studies [1]. The sources of experimental data include nuclear magnetic resonance spectroscopy spectroscopy (NMR) and small-angle X-ray- and neutron scattering. This article addresses several issues which often make inference of biomolecular structure from such data particularly challenging. First, in these experiments, the time-scale of acquisition typically exceeds that of molecular fluctuations. Second, the samples studied are often near molar concentrations. Third, data is frequently incomplete, or even sparse, and subject to experimental noise. Consequently, data obtained from such techniques yield incomplete, noisy, spatially and temporally averaged information on the Boltzmann ensemble of the observed system. Thus, such data are ideally analyzed through models that take these properties into account. While this fact has long been recognized, the analysis of these types of data has revolved predominantly around structure determination – that is, fitting a single conformation to fulfill all derived geometrical restraints [2]. Such structure determination methods do not adequately handle sparse, noisy and averaged data. Here, we propose an alternative method which addresses these shortcomings.

Typically, structure determination from experimental data proceeds through hybrid energy minimization [3]. In this method, an energy function $E_{\text{exp}}$ that brings in the experimental data is combined with an approximative physical forcefield $E_{\text{phys}}$. The term $E_{\text{exp}}$ is typically a straight-forward combination of a forward- and an error-model. A forward-model relates
a protein conformation to experimental data, whereas an error-model concerns experimental errors. Recently, a Bayesian formulation known as inferential structure determination (ISD) emerged, formulating structure determination in a rigorous probabilistic framework [4]. In ISD, a posterior distribution is constructed by combining a data likelihood with prior distributions on conformational and nuisance parameters. The likelihood and the prior of biomolecular structure correspond to $E_{\text{exp}}$ and $E_{\text{phys}}$, respectively. This Bayesian approach extends the common hybrid energy minimization by solving the important problems of choosing appropriate error-models, treating model-parameters coherently and performing inference through posterior sampling rather than minimization. However, by construction, these approaches assume that conformational variability can be represented through uncorrelated, homoscedastic fluctuations around one average structural representation. Consequently, the conformational heterogeneity present in the posterior distribution reflects the quality and completeness of the experimental data and the prior distributions, but not necessarily any physical fluctuations [5]. Despite this well-known limitation, the approximation tends to yield good results for well-folded proteins when conformational fluctuations are modest.

Early attempts to model ensemble NMR data involved averaging along molecular dynamics trajectories [6, 7]. In these protocols, a memory function specifies an averaging time-span which is used to obtain a time-averaged representation of the experimental data. While this approach displayed initial promise, the short timescales accessible through routine molecular dynamics limit its use [8]. An alternative approach, which involves explaining the data using an average of several conformations, emerged around the same time [9]. This approach has since shown to be much more viable, as we discuss next.

During the past two decades there has been an increasing interest in biomolecules that undergo significant conformational fluctuations, such as natively unfolded and partially unfolded proteins [10]. Consequently, there have been many efforts to overcome the limitations of structure determination procedures with respect to the flexibility of these molecular systems. Prevalently, conformational fluctuations are represented by finite ensembles: the data is explained by a weighed average of $N > 1$ conformations, introduced above. In effect, this corresponds to discretizing the Boltzmann ensemble. Such discrete ensembles are constructed in a multitude of ways, including database-derived explicit ensembles [11, 12], data-optimized explicit ensembles [13–16] and multi-conformer refinement, molecular dynamics and Monte Carlo methods [8,17,18] and maximum entropy methods [19]. Another important approach considers multiple replicas in the hybrid energy of molecular simulations [20]. However, the discretization of the conformational ensemble is inherently problematic as determining the optimal ensemble size $N$, and its associated uncertainty of this, is difficult.

Restraining simulations through the average of multiple replicas is a sensible solution, as it was recently shown that multiple replica restrained simulations constitute the least biased method when the number of replicas goes to infinity in the absence of experimental noise [21,22]. However, a measurable bias is introduced when the number of replicas used is too small [22]. Thus, the development of approaches which are independent of this discretization is highly desirable.

In this work, we approach the problem of modeling sparse, spatially and temporally averaged data through the principles of Bayesian statistics and information theory. Unlike the previous Bayesian efforts [4,16], we explicitly take into account the experimental data as noisy, average
quantities of an underlying heterogenous ensemble in continuous space. We derive a general posterior distribution from first principles which imposes the least necessary bias on our prior knowledge to fulfill the experimental data. Furthermore we illustrate its tractability through applying the method to a sparse, synthetic dataset generated from the small peptide GB1 using the PROFASI forcefield at high temperature [23].

**Results & Discussion**

A hierarchical model of spatially and temporally averaged restraints

Ultimately, our aim is to sample from the conditional probability distribution \( p(x \mid d) \), where \( x \) denotes a protein’s conformation and \( d \) denotes spatially and temporally averaged, experimental data. The variable \( x \) represents a positional microstate in atomic detail. Through a forward model \( f(x) \) we can calculate a coarse-grained representation, \( f \), of a protein conformation \( x \). That is, our forward model is a mapping, \( f : \mathbb{R}^{3N} \to \mathbb{R}^M \), of the \( N \) atoms of \( x \) to an \( M < 3N \)-dimensional coarse grained representation, \( f \). Conceptually, \( f \) may be interpreted as the instantaneous ‘experimental data’ back-calculated from a positional micro-state, \( x \). However, as \( d \) represents an averaged quantity we need to introduce a variable, \( e \), to represent an ensemble average of the simulated experimental data \( f \). Consequently, our full posterior distribution becomes \( p(x, f, e \mid d) \).

We clarify the relation between \( f, e \) and \( d \) using the example we will present later on. In the case of nuclear Overhauser enhancement (NOE) data obtained from an NMR experiment [24], the coarse-grained variable \( f \) is a vector related to pairwise distances between atoms in a protein conformation \( x \). In one case, this is simply a vector of these distances. The variable \( e \) is an average of \( f \) vectors from an ensemble of protein conformations. The experimental NOE data \( d \) can be interpreted as a noisy observation of the vector of averages, \( e \). In general, there is no simple relationship between the vector \( f \) and the averaged vector \( e \), but a simple probabilistic model that relates them can be developed, as we discuss next.

We start by considering the coarse-grained representations of the distribution, \( f, e \) and \( d \), without considering the fine-grained representation, \( x \). Following the Bayesian probability calculus, we formulate a posterior distribution:

\[
p(f, e \mid d) \propto p(d \mid f, e)\pi(f, e) = p(d \mid e)\pi_f(f \mid e)\pi_e(e),
\]

where the first term is the likelihood and the second term is the prior distribution. The equality is due to the redundancy of \( f \) in the evaluation of the likelihood function – \( d \) is a noisy observation of \( e \), which does not involve \( f \). The independence assumptions of the model are shown in the corresponding graphical model in figure 1.

Applying the product rule of probability theory to equation (1), we obtain

\[
p(f, e \mid d) \propto p(d \mid e)\pi_f(f \mid e)\pi_e(e).
\]

\( \pi_f(f \mid e) \) is the prior distribution of the simulated data \( f \) given their averaged value \( e \), and \( \pi_e(e) \) is the prior distribution over the simulated ensemble averaged data \( e \).
Equation (2) is a probabilistic model of the relationship between noisy, ensemble averaged data, and conformational micro-states in a coarse-grained space. However, to obtain a probability distribution \( p(\mathbf{x}, \mathbf{f}, \mathbf{e} | \mathbf{d}) \) which features atomic detail, we need to combine (2) with a fine-grained physical forcefield or a probability distribution, \( \tau(\mathbf{x}) \). This can be done by using the reference ratio method [25]:

\[
p(\mathbf{x}, \mathbf{f}, \mathbf{e} | \mathbf{d}) \propto \frac{p(\mathbf{d} | \mathbf{e})\pi_f(\mathbf{f} | \mathbf{e})\pi_e(\mathbf{e})}{\tau_f(\mathbf{f})}\tau(\mathbf{x}).
\]

\( \tau_f(\mathbf{f}) \) is called the reference distribution, and is the distribution induced in the coarse-grained space by the fine-grained prior, \( \tau(\mathbf{x}) \). The reference distribution arises as a prior on \( \mathbf{x} \) directly implies a prior on \( \mathbf{f} \), due to the deterministic relationship between the parameters through the forward model \( f(\cdot) \). The reference ratio method yields the Kullback-Leibler optimal modification of the fine-grained model \( \tau(\mathbf{x}) \) with respect to the coarse-grained information (for proof, see chapter 4 in [26]). Kullback-Leibler optimality is closely linked to the maximum entropy principle of Jaynes [27]. In essence, our approach can be seen as a maximum entropy solution given the noisy observation of an ensemble average. It should be noted that even if the distribution given by Equation (2) is unimodal, the posterior given by Equation (3) can still be multimodal due to the nature of the conformational prior, \( \tau(\mathbf{x}) \).

**The relationship to other methods**

The model given by Equation (3) may be reduced to the ISD framework [4]

\[
p(\mathbf{x}, \mathbf{f} | \mathbf{d}) \propto p(\mathbf{d} | \mathbf{f})\tau(\mathbf{x})
\]

if we choose the Dirac delta function \( \delta(\mathbf{f} - \mathbf{e}) \) for \( \pi_f(\mathbf{f} | \mathbf{e})\pi_e(\mathbf{e}) \) and assume that \( \tau_f(\mathbf{f}) \) is uniform. Choosing the Dirac delta function corresponds to assuming the Boltzmann distribution is infinitely narrow. Hence, our model can be seen as a generalization of ISD. The choice of the uniform distribution for \( \tau_f(\mathbf{f}) \) corresponds to assuming that \( \tau(\mathbf{x}) \) implies a suitable prior for \( \mathbf{f} \) as well. This may be inappropriate in some cases (see below).

We also observe that Equation (2) may be reduced to the previously proposed maximum entropy restraining methods [21,22]. This is evident if we consider the case where \( p(\mathbf{d} | \mathbf{e}) \) is the normal distribution and \( \pi_f(\mathbf{f} | \mathbf{e})\pi_e(\mathbf{e}) \) is a log-linear model \( \left( \mathcal{G}(\cdot) \right) \) with a linear link-function, \( \ell(\mathbf{A}, \mathbf{b}) = \mathbf{A}\mathbf{b} \). The link function allows us to include the Lagrange multipliers used to relate the coarse-grained variable \( \mathbf{f} \) to the mean value \( \bar{\mathbf{f}} \) [28]. Thus, \( \mathcal{G}(\mathbf{f} | \Lambda, \mathbf{e}) = \exp[\mathbf{c} + \mathbf{f}\ell(\Lambda, \mathbf{e})] \propto \exp(\mathbf{f}\Lambda\mathbf{e}) \).

We have

\[
p(\mathbf{f}, \mathbf{e} | \mathbf{d}) \propto p(\mathbf{d} | \mathbf{e})\pi_f(\mathbf{f} | \mathbf{e})\pi_e(\mathbf{e}) = \mathcal{N}(\mathbf{d} | \mathbf{e}, \sigma)\mathcal{G}(\mathbf{f} | \Lambda, \mathbf{e})
\]

where \( \Lambda \) is a diagonal matrix of Lagrange multipliers. If we now consider the limit where the experimental noise vanishes we obtain,

\[
\lim_{\sigma \rightarrow 0} \mathcal{N}(\mathbf{d} | \mathbf{e}, \sigma)\mathcal{G}(\mathbf{f} | \Lambda, \mathbf{e}) = \mathcal{G}(\mathbf{f} | \Lambda, \mathbf{d}).
\]

In minus log-space Equation 6 is proportional to minus \( \mathbf{f}\Lambda\mathbf{d} \). This corresponds to the empirical term of the previously reported maximum entropy method in absence of experimental uncertainty [21]. We note that this method does not explicitly account for the reference distribution...
Reconstructing a high temperature ensemble from sparse data and a probabilistic prior of local protein structure

To test our model, we use synthetic NOE data, obtained from an ensemble of the GB1 hairpin simulated at 400K in the Profasi forcefield [23]. The restraints are visualized on a random conformation of the GB1 hairpin in Figure 2. Historically, NOEs constitute one of the most important sources of semi-quantitative information in NMR structure determination. Under the isolated spin-pair approximation for rigid molecules, NOEs are related to an interatomic distance $r$ as $\text{NOE} \propto \langle r^{-6} \rangle$ [29]. As an example, we will apply equation (2) to two cases of averaged pairwise distance data – these two cases involve the arithmetic mean $\langle r \rangle$, and the power-averaged mean $\langle r^{-6} \rangle$, respectively. They represent two different averaging processes that are common in biophysical data.

We use the log-normal distribution as an appropriate error-model for pairwise distances derived from NOEs, which is the approach adopted by ISD [30]. The choice is less obvious for the prior $\pi_f(f | e)$ and depends on the type of experimental data. Here, we use the exponential distribution with mean $\beta$, $\mathcal{E}(x | \beta) = e^{-\frac{x}{\beta}} \beta^{-1}$, since it constitutes the least biasing continuous distribution on the positive real axis, when no higher order moments are observed [31]. The prior on $f$ thus becomes: $\pi_f(f | e) = \pi_f(f | e, w) = \prod_i \mathcal{E}(f_i | e_i, w_i)$, where the product runs over all data-points and the scale vector $w$ is a free parameter (discussed below). It follows that

$$p(x, f, e | d) \propto \frac{\mathcal{N}(\ln d | \ln e, \sigma) \pi_f(f | e, w) \pi_e(e)}{\pi_f(f)} \tau_f(x), \quad (7)$$

where $\sigma$ is the experimental error, which is fixed and given, and $\mathcal{N}(\cdot)$ is the normal distribution. As prior on the ensemble average $e$ we choose $\pi_e(e) \propto e^{-1}$. This prior has previously been shown to provide good results for variables confined to the positive real axis [32].

Equation (7) provides an exact solution to the problem of modeling averaged NOE data subject to experimental uncertainty. The only parameter to be estimated is the scale vector $w$, which relates $f$ to $e$ in $\pi_f(f | e, w)$.

In the ideal case, an optimal choice for $w$ results in the desired distribution for $f$ as calculated from the structures $x$. More precisely, it results in a marginal posterior distribution of Equation (7) for $x$, such that, when $e$ is fixed, the expectation of $f$ is equal to $e$. In practice, a satisfactory point estimate for $w$ can be obtained in an iterative manner, using an empirical Bayes approach (see Materials and Methods). The parameter $w$ compensates for the approximate nature the reference distribution $\tau_f(f)$, which is difficult to estimate accurately [25]. The introduction of $w$ provides a simple, yet effective measure to compensate for this.

We use equation (7) to model synthetic pairwise distance restraints in the GB1 hairpin. For $\tau(x)$ we use probabilistic models of the conformational space of the main chain [33] and the side chains [34], as these models recently yielded excellent results when combined with the ISD method [35]. As the prior distribution and the likelihood concern local and nonlocal features of protein structure, respectively, their information content shows little overlap. More informative
priors, for example based on physical energy functions, can be envisaged, but this is beyond the scope of this article.

The prior distribution used in this study concerns protein structure on a local length scale, and thus does not model long range distances accurately. Consequently, as a reaction coordinate, we chose a representative distance $f_0$ between atoms $\text{C}_{\alpha}^{41}$ and $\text{C}_{\alpha}^{51}$ – which are separated farthest in sequence – to illustrate the long-range properties of the eight different ensembles considered here. Histograms of this pair-wise distance in the different ensembles are shown in Figures 3 and 4. This pair-wise distance is highlighted with yellow color in Figure 2.

The conformational prior and PROFASI, which was used to generate the averaged data, result in different distance distributions (Figure 3). However, if we modify the prior using the reference ratio method as described above, we obtain good fits with the PROFASI distribution for both linearly and power-averaged data. The ISD ensemble, which does not take the ensemble nature of the data into account, is overly tightly peaked around the (correct) mean.

A similar pattern is observed for the distribution of the gyration radii $R_g$. The gyration radii are not used in the estimation of the probability distributions, and can thus be used for cross-validation. The average and standard deviation of the gyration radii of the ensemble used to generate the data is $9.71 \pm 1.5\text{Å}$. The ensembles obtained with our method from the power averaged and linearly averaged data resulted in a slightly higher average (10.30 ± 1.8Å and 10.19 ± 1.6Å, respectively), but essentially the correct standard deviation. This is an excellent result, as a perfect fit is not expected due to the sparse and noisy nature of the data. Again, the ISD ensemble provides an overly narrow distribution ($9.88 \pm 0.6\text{Å}$). Finally, sampling from the prior distribution alone results in an average radius of gyration of $11.34 \pm 1.8\text{Å}$, which is considerably too high.

In some cases, the bias introduced by the reference distribution is not critical for obtaining good results. As it constitutes an additional obstacle in terms of estimation and simulation time, we evaluate its significance on the obtained results. In the power averaged case we achieved this by choosing the reference distribution $\tau_f(f)$ and the scale vector, $\mathbf{w}$, in equation (7) to be the uniform distribution and the unit vector, respectively. In the linearly averaged case, the scale vector was kept fixed equal to the 1-vector, as $\tau_f(f)$ was assumed to be uniform in the results presented above. The results are shown in figure 4. In the case of the power averaged data, with $\tau_f(f)$ uniform and $\mathbf{w}$ equal to the 1-vector, severely skews the distribution of the distances (green line in figure 4). When the scale vector $\mathbf{w}$ is estimated, while still assuming $\tau_f(f)$ uniform, the fit improves (blue line), but without resulting in a satisfactory distribution. In the linearly averaged case, we find that a 1-vector for $\mathbf{w}$ provides good fit (red line).

If we again consider the gyration radii as providing complementary views of the ensembles, we find that the power-averaged ensemble with uniform $\tau_f(f)$ and unit scale vector yields a overly extended ensemble, $11.94 \pm 1.63\text{Å}$. The results in the linearly averaged case compare to those with an estimated scale-vector, $10.34 \pm 1.69\text{Å}$, presented above.

To summarize, in the power averaged case, both $\tau_f(f)$ and $\mathbf{w}$ are required for a satisfactory distribution. In the case of the linearly averaged data, our results suggest that $\mathbf{w}$ and $\tau_f(f)$ may be approximated by the 1-vector and uniform distribution, respectively. This is particularly interesting as it may be a general feature of applying other kinds of linearly averaged data. This will make the use of these types of data for restraining easier.
Conclusion

In conclusion, we present a Bayesian principle to infer ensembles of protein structures from noisy experimental data subject to ensemble and time averaging. The principle is successfully applied and constitutes a generalization of ISD and previously proposed maximum entropy restraining approaches. We successfully apply the method to analyze noisy linearly and power averaged data.

Our approach combines a coarse-grained Bayesian model of the data with a fine-grained model of protein conformational space. The combination is accomplished using reference ratio method [25], which corresponds to a maximum entropy solution in the presence of experimental noise. The role of the reference distribution $\tau_f(f)$ is considerable. When we assumed $\tau_f(f)$ to be uniform, we were unable to construct sufficiently accurate distribution of pair-wise distance geometry, in the case of power-averaged data.

The Bayesian model can in principle be applied to denser datasets and to other data such as small angle X-ray- or neutron scattering data, or other data derived from NMR experiments. Also, low-resolution data may be combined with more sophisticated physical prior distributions such as those embodied in force fields. The presented method is thus a general and efficient method to obtain physically sound ensemble models of solution and endogenous states of biomolecules, given appropriate experimental data. Similar models could in principle be used for other inference problems in physics, biology or medicine that involve spatially or temporally averaged data.

Materials and Methods

Synthetic datasets

A synthetic dataset was created for the GB1 hairpin (Protein data bank identifier: 1LE3; sequence variant [Y45W, F52W, V54W]). The data were generated by simulating the protein at 400K with the PROFASI forcefield [23], using Engh-Huber parameters for bond-angles and bond-lengths [36]. The high temperature was used to emulate the effect of a denatured, disordered state. A total of $3.5 \cdot 10^8$ steps were performed using the Metropolis-Hastings algorithm in the PHAISTOS Markov chain Monte Carlo framework (http://www.phaistos.org). We used a Monte Carlo move set previously described [35]. Samples were saved in intervals of 5000 steps. These samples were used to form five non-redundant, averaged $C\alpha - C\alpha$ distance restraints (see Table 1).

To mimic the effect of distance averaging in a dipolar interaction undergoing fast motion compared to the cross-relaxation but slow motion when compared to molecular tumbling, we calculated a power averaged variant of the dataset as $I_i = \langle r_i^{-6} \rangle$, where $r_i$ is an inter-atomic distance and the angular-brackets denote ensemble averaging [37]. We used an experimental uncertainty for the power averaged dataset $\sigma_I$ of the same relative amplitude as for the average restraint set $\sigma_d$, by enforcing the signal-to-noise ratio to be constant. Hence, $\frac{d}{\sigma_d} = \frac{I}{\sigma_I} \Rightarrow \sigma_I = \frac{I}{d} \sigma_d$, as $\sigma_d = 1$, where $I$ and $d$ denote the datapoints corresponding to the largest average distance in the power and linearly averaged datasets, respectively. Noise with standard-deviation $\sigma_I$ was added to the power-averaged data.
Estimation of $p(x, e, f \mid d)$ and scale vector $w$

This section describes the estimation of the reference distribution $\tau_f(f)$ and the vector $w$ needed for the posterior distribution:

$$p(x, f, e \mid d) \propto \frac{N(\ln d \mid \ln e, \sigma) \pi_f(f \mid ew) \pi_e(e)}{\tau_f(f)} \tau(x).$$  \hspace{1cm} (8)

In the case of the power-averaged data, the reference distribution $\tau_f(x)$ was approximated by a product of exponential distributions:

$$\tau_f(f) \propto \prod_i \mathcal{E}(-f_i/\beta_i).$$  \hspace{1cm} (9)

The mean $\beta$ was estimated using a Monte Carlo scheme similar to that used to form the synthetic datasets, but only using the prior $\tau(x)$, consisting of the probabilistic models TorusDBN [33] and Basilisk [34] along with a simple binary term assuring atoms do not overlap [??]. The coarse graining used was the inverse pairwise distance between the C$\alpha$ atoms listed in Table 1. For the linearly-averaged data, $\tau_f(f)$ was approximated by a uniform distribution.

We obtain a point estimate of $w$ following an empirical Bayes approach. We start by initializing all the elements of $w$ to unity. Subsequently, we sample an ensemble according to Equation (8), and update $w$ based on the sampled values of $f$ and $e$. To update $w$ we make use of the moment estimator for the mean of the exponential distribution:

$$w_{i+1} = w_i \frac{\bar{f}}{\bar{e}}.$$

$\bar{f}$ and $\bar{e}$ are posterior expectations of the coarse-grained variable and the ensemble averages using scale vector $w_i$, respectively, and $w_{i+1}$ is the updated scale vector. This procedure is repeated until convergence. Convergence was assumed when fluctuations in $\bar{f}$ were within the experimental uncertainty. Each step in the algorithm runs for $2.5 \cdot 10^6$ MCMC steps, and a final production ensemble is produced using $25 \cdot 10^6$ MCMC steps.

Sampling of $e$

To sample $e$ from the prior $e^{-1}$ we sampled a factor $\Delta$ from a log-normal distribution $\Delta \sim \exp [N(0, \sigma)]$, where $\sigma$ has the same order of magnitude as the experimental uncertainty. A change from $e$ to $e\Delta$ was accepted according to the Metropolis acceptance probability $\alpha$:

$$\alpha(e \rightarrow e\Delta) = \min\left(1, \frac{p(e\Delta)}{p(e)}\right),$$

where $p(\cdot) = \frac{p(e, f, x, d)}{\pi_e(e)}$.

Acknowledgments

We thank Kresten Lindorff-Larsen and Jesper Ferkinghoff-Borg for valuable comments, discussions and suggestions. S.O. and J.F. acknowledge funding from the Danish Council for Independent Research (Technology and Production Sciences, FTP, 274-09-0184)
2.2. INFERENCE OF PROTEIN STRUCTURE ENSEMBLES

References


Figure Legends

Tables
Reference ratio
Fine-grained structural prior
\[ p(x, f, e|d) \propto p(f, e|d) \]
\[ \tau_f(f) \]
\[ \tau(x) \]

Coarse-grained ensemble model

Figure 1. A directed graphical model of the ensemble model (on the left) and its interplay with a fine-grained conformational prior distribution (top right) through the reference ratio method, (bottom right). In the graphical model, the black circles are random variables, and the arrows determine their conditional independencies. The parameter \( \sigma \), marked in grey on the left, is fixed and given, and denotes the experimental error in this particular example. \( \tau_f(f) \) denotes the reference distribution.

Table 1. Synthetic datasets used in this study. First column: \( \alpha \) atoms involved in the pairwise distance. Second and last columns: averaged and power-averaged pairwise distances, respectively.

<table>
<thead>
<tr>
<th>Co-pair</th>
<th>( \langle r_i^{-6} \rangle )</th>
<th>( \langle r_i \rangle )</th>
</tr>
</thead>
<tbody>
<tr>
<td>41–51</td>
<td>( 4.799 \cdot 10^{-6} )</td>
<td>19.40</td>
</tr>
<tr>
<td>42–48</td>
<td>( 1.32 \cdot 10^{-6} )</td>
<td>14.36</td>
</tr>
<tr>
<td>44–46</td>
<td>( 2.07 \cdot 10^{-7} )</td>
<td>6.06</td>
</tr>
<tr>
<td>44–54</td>
<td>( 2.17 \cdot 10^{-7} )</td>
<td>19.39</td>
</tr>
<tr>
<td>53–54</td>
<td>( 13.29 \cdot 10^{-4} )</td>
<td>3.51</td>
</tr>
</tbody>
</table>
Figure 2. A random backbone conformation of the GB1 hairpin obtained from the prior distribution (TorusDBN). The restraints listed in Table 1 are shown as dashed lines. The distance shown in red is used as the reaction coordinate $f_0$ used in Figures 3 and 4. This figure was created using PyMOL (DeLano Scientific LCC).
Figure 3. Histograms, $p(f_0)$, of a representative pairwise distance $f_0$ (between C$\alpha^{41}$-C$\alpha^{51}$, in Å) in the ensembles. The black and blue lines are obtained from the PROFASI and ISD ensembles respectively, while the cyan line represent the prior $\tau(x)$. Finally, the green and red lines respectively represent ensembles obtained from the power-averaged and linearly averaged data.
Figure 4. The influence of $\tau_f(f)$ and $w$ on the ensembles. The figure shows histograms, $p(f_0)$, of a representative pairwise distance $f_0$ (between C$\alpha_{41}$-C$\alpha_{51}$, in Å) in the ensembles obtained without the reference distribution $\tau_f(f)$ or the scale vector $w$. The black line denotes the PROFASI target ensemble; the red and green lines denote the ensembles obtained using the linearly and the power averaged data, respectively. The blue line denotes the case of the power averaged data without $\tau_f(f)$, but with $w$. 
2.3 Construction of protein structure ensembles in continuous space through minimal biasing of a simple physical force field.

The study of concerted motions in proteins is an inherently difficult task. This type of motion occur on timescales which are difficult reach directly in experiment and unrestrained simulations. In this paper we seek use a coarse, yet efficient, forcefield restrained by a modestly sized set of local geometric RDC restraints to study the native ensemble of the third immunoglobulin binding domain of protein G (GB3). We achieve this by using our previously proposed hierarchical restraining strategy.

We find that the restrained ensembles obtained in this procedure agree with complementary experimental data significantly better than the unrestrained ensemble. This cross-validation considers data reflecting both of local and non-local geometric features. In addition we find that previously reported dynamic properties are accurately reproduced. Finally, we find a strong anti-correlated motion which may be attributed to the binding GB3s endogenous partner, Fab.

Collectively, these results suggest that we are able to construct realistic distributions of protein structure using comparably small datasets of local geometric restraints by complementing by simplistic forcefields. This may prove to be a powerful technique to allow for the study of concerted motions using accurate local restraints.

This is a first-author paper and I was involved in all aspects of the work. Unsubmitted manuscript.
Construction of protein structure ensembles in continuous space through minimal biasing of a simple physical force field.

Simon Olsson, * and Thomas Hamelryck *

* Bioinformatics Centre, Department of Biology, Faculty of Science, University of Copenhagen, Denmark

draft manuscript

The structural motion of biological molecules is tightly coupled to their endogenous functions. However, while the study of fast motions has become increasingly feasible in recent years, the study of slower, biologically important, motions remains confined to particularly tractable systems. Here we present a native ensemble of the extensively studied third immunoglobulin binding domain of protein G (GB3). We use an efficient, yet coarse, force field restrained by a modestly sized data-set in a minimally biased fashion to accurately describe both local and non-local conformational dynamics. The resulting ensemble is in agreement with previous studies that rely on much more elaborate datasets. In addition, we find a previously unreported anti-correlated motion between the β-sheet and the α-helix. GB3 appears to undergo a similar anti-correlated motion when it binds its endogenous partner, Fab. These results suggest that we are able to construct realistic conformational ensembles of biomolecules in continuous space very efficiently. In a more general setting, this approach may allow for a dramatic reduction in the computational as well as experimental resources needed to obtain accurate models of conformational dynamics in biomolecules.

Under physiological conditions the motions of proteins span several spatial and temporal orders of magnitude. Fast and, in particular, slow molecular motions are essential to obtain, sustain and regulate biological functions which include signal transduction [1] and catalysis [2]. The study of fast molecular motions may be routinely performed using experimental [3] or computational [4] means. Slower motions are, however, notoriously difficult to investigate [5]. Computationally, such studies require either special purpose super-computers or extensive datasets for unrestrained and restrained simulations respectively [6, 7]. Experimental techniques such as relaxation dispersion allow extraction of thermodynamical, structural (chemical) and kinetic information about the exchange between states [8]. However, these experiments are only sensitive provided that the chemical environment is modulated significantly by this motion [9].

Residual dipolar couplings (RDCs) provide geometric information averaged over many time scales (up to ms) and may be used to study the slow molecular motions of proteins [10]. Such an application of RDCs has already been subject to extensive research [6, 11, 12]. Several reports describe fitting of discrete sets of conformations to large experimental datasets [12], performing restrained simulations and selecting sets of protein conformations post hoc to describe structural variability in folded and intrinsically disordered proteins [13, 14]. Common to these procedures is the assumption that motion reflected in the data is accurately represented by a weighted average of a discrete set of molecular conformations. Potential problems in these methodologies have been pointed out previously [15, 16].

In this work, we studied the native ensemble of the third immunoglobulin binding domain of protein G (GB3). We curated a modestly sized dataset of from RDCs obtained through structurally conservative mutagenesis of charged surface residues that were reported previously [17]. Such mutations are common and thus this dataset represents a case which may be easily reproducible for other systems. The data for the different mutants give complementary views of the N-H and C"=H bonds in the protein backbone of GB3 – something which may be difficult to achieve by solely changing alignment media [18]. In principle, the method is applicable to datasets collected under less alignment conditions if RDCs between other nuclei are included. We used these as restraints in a native simulation using the Profasi forcefield [19]. We utilized our previously described strategy to impose the minimal necessary bias on the forcefield to be in accord with noisy experimental data [16]. We devised a robust algorithm to iteratively refine the bias with guaranteed convergence in finite time. We found that the restrained ensemble accurately reproduces complementary experimental data not used to strain the simulations. Furthermore, we found that the local correlated motions present in the unrestrained ensemble were not significantly distorted by our restraining procedure and were in agreement with what has been previously suggested [6, 20]. Finally, we found strong anti-correlated motion between the α-helix and the β-sheet in our restrained ensemble. This motion is consistent with the conformational difference between GB3 in its free state and when it is bound to its endogenous partner.

These results indicate that efficient, yet coarse, forcefields, when appropriately restrained by modestly sized datasets, may be used to obtain accurate native ensembles. These simulations are computationally inexpensive, even on commodity hardware, dramatically reducing the computational and experimental resources necessary to obtain accurate descriptions of the native fluctuations in biomolecular systems. This may in turn allow for rigorous large-scale studies of dynamics in biomolecules.

Reserved for Publication Footnotes
Results

Kullback-Leibler optimal restraining of force fields. Our aim is to use the computationally efficient forcefield Profasi in a native ensemble simulation restrained by experimental observations, which are subject to experimental noise, as well as spatial and temporal averaging. The experimental dataset we use here is relatively small compared to what has been used previously [6, 20]. Consequently, it will not be sufficiently informative to provide an accurate description of the native ensemble of structures devoid of any prior assumptions. Similarly, the forcefield Profasi only provides a simplified view of the conformational distribution of the protein. Therefore, we seek to restrain Profasi using our data in a minimally biasing way. This may be achieved by considering the restraining problem as a problem in Kullback-Leibler optimization. This corresponds imposing the minimal bias on our prior distribution \( \pi(x) \) (the Boltzmann ensemble of the Profasi forcefield) in order for the average \( f \) of the instantaneous representation \( f \) to match our data within the experimental uncertainty. We have previously described a general way of performing this optimization in our current context, temporally and spatially averaging data subject to experimental uncertainty [16].

We have the posterior distribution,

\[
p(f, e, x \mid d) \propto p(d \mid e, \sigma) \pi(f \mid e) \pi(x \mid \pi(e)) \quad [1]
\]

where \( f \) denotes an instantaneous coarse grained representation ('back-calculated experimental data') of a protein structure \( x; e \) represents an ensemble average of this coarse graining, and \( d \) represents an experimental observation consistent with this particular coarse-graining. Here we use a normal distribution, \( \mathcal{N}(\cdot) \), for the likelihood \( p(d \mid e, \sigma) \), and a log-linear model \( \mathcal{G}(\cdot) \) for the prior \( \pi(f \mid e) \). The likelihood \( \mathcal{N}(d \mid e, \sigma) \) models the probability of the experimental data given its uncertainty \( \sigma \) and some unknown ensemble average. The prior \( \mathcal{G}(f \mid V, e) \) models the distribution of protein conformational

given the ensemble average and the scale matrix \( V \). Finally, we chose an uniform prior for the ensemble average, \( \pi(e) \propto 1 \). We wished to obtain a point estimate of the unknown (latent) scale matrix \( V \) such that a given \( e \) yields a posterior expectation of \( f \) according to eq. 1 which is exactly \( e \). This corresponds imposing the least necessary bias to the prior distribution of protein conformational features to fulfill the experimental data. We here devise an EM algorithm to carry out this estimation (see Materials and Methods). Additionally, we make use of a new method to model to account for the different experimental alignment conditions of the datasets (see Materials and Methods).

Expectation maximization algorithm yields minimally biased native ensembles To investigate the native ensemble of GB3 we use eight sets of previously reported residual dipolar couplings (RDCs) [17] to restrain the forcefield Profasi [19]. The data was acquired in seven different alignment conditions, by using structurally conserved mutants K19AD47K, K19ED40N, K19AT11K, K19EK4A, and two which include a C- and N-terminal His-tag, K19EK4A-C-His6, and K19EK4A-N-His6 along with the wildtype of GB3. Seven of the datasets reported on the N-H\(^{-}\) bond whereas one reported on the C-H\(^{+}\) bond. These conditions have been shown to accurately map the five independent parameters associated with molecular alignment [21]. Consequently, using approximate structural prior information these datasets should very accurately describe the fluctuations of the N-H\(^{-}\) bond and the C-H\(^{+}\) bond, albeit to a lesser extent.

We used an EM algorithm [22] to iteratively refine restrained native ensembles of GB3 (EM\(_i\) for \( i = 0 \ldots 7 \)). We find that agreement with the training data (declining Q-factor) correlates with the increase in expected posterior probability (decline in MMLP) of the model (Fig. 1) and agreement with complementary data not used in the restraining (see Supplementary Material). The final restrained ensemble (EM\(_7\)) is stable near the native state of GB3, with considerable flexibility. In contrast, the unrestrained ensemble (EM\(_0\)) partially unfolds during simulation, see Table 1.

Realistic local and non-local dynamics. While it is promising that the restrained ensemble, EM\(_7\), displays considerable flexibility and is in good agreement with the training data (\( Q = 15\% \)), it does not assure physical realism. We are able to test the quality of EM\(_7\) by cross-validation with the wealth of experimental data available for GB3, providing local and non-local geometrical information. We compared the reproduction of the experimental data to two previously published models of GB3: 2LUM, an ensemble refined using exact nuclear Overhauser enhancements (eNOEs), scalar couplings, RDCs and chemical shifts [20] and 2OED a RDC refined x-ray structure [23].

As shown in Table 1, EM\(_7\) is in excellent agreement with previously reported scalar couplings [24] and previously unpublished cross-correlated dipolar relaxation rates. The unrestrained EM\(_0\) ensemble, in general, is of considerably worse quality. Compared to the previously published models of GB3, 2LUM and 2OED, agreement is comparable. In particular, it is striking that EM\(_7\) shows slightly better quantitative \( J_{HN-\alpha C} \) agreement when compared to the 2LUM ensemble, where these data were used in the refinement process. The cross-correlated dipolar relaxation rates \( R_{HN-\alpha C} \) are particularly sensitive to an accurate representation of the N-H\(^{-}\) bond [25, 26]. We see a dramatic improvement in the correlation with these data when comparing the restrained to the unrestrained ensemble. This suggests that the RDCs used to restrain the Profasi forcefields indeed accurately describe the fluctuations of the N-H\(^{-}\) bonds.

The non-local dynamics were evaluated by cross validation with stereospecifically assigned exact nuclear Overhauser enhancements (eNOEs) and hydrogen bond scalar couplings (HBCs). As with the scalar couplings above, the 2LUM ensemble was refined against the eNOE data. However, here this ensemble generally showed the correlation with these data. This is in the case when both the conformational exchange is assumed to be much faster (eNOE\(_{\alpha C} \)) or slower (eNOE\(_{\alpha C} \)) than the molecular tumbling [27]. The RDC refined structure, 2OED, shows slightly worse agreement while the restrained (EM\(_7\)) and unrestrained (EM\(_0\)) Profasi ensembles are approximately 10\% and 20\% worse when compared to 2LUM. Surprisingly, the agreement with the hydrogen bond scalar couplings is quantitatively best in the restrained Profasi ensemble EM\(_7\), closely followed by 2OED, 2LUM and the unrestrained ensemble. This result, in particular, is interesting because the HBCs report long-range correlated motions [28] averaged on the same timescale as RDCs [29]. Collectively, these results suggest that the local and non-local dynamics in EM\(_7\) are of good quality.

Discussion

Concerted motions in the minimally biased ensemble may suggest a conformational selection binding mech-
2.3. CONTINUOUS ENSEMBLES OF PROTEINS

anism. Approximately 90% of the 56 residues of GB3 are in distinct secondary structure elements: an α-helix is wrapped underneath a mixed anti-parallel/parallel four-stranded β-sheet [30]. Consequently, GB3 is typically considered a relatively rigid protein, undergoing minor, albeit distinct, structural transitions when binding to its biological partner, Fab (see Fig. 3) [31]. Still, fluctuations of varying magnitude and character throughout this fold have been reported in previous studies. These motions range from fast, isotropic bond liberations over local crank-shaft moves by anti-correlated torsion-angles in adjacent residues [32, 33] to long-range correlated motions of the β-sheet across hydrogen bonds [6, 20]. In particular the N-terminal sites of the α-helix and the β2-strand have been reported to undergo the most significant fluctuations on sub-nanosecond time scales. Curiously, these same sites are involved in binding with Fab. In EM7 we see a similar behavior (Fig. 2b), local correlated motions within the secondary structure elements: in particular, strongly correlated motions of the α-helix and strong concerted motions of the β-sheet. While these motions are of relatively subtle amplitude their correlated nature is quite striking.

In addition, we find strongly anti-correlated motions between the α-helix and the β-sheet in our restrained EM7 ensemble, in particular in the N-terminal hairpin. It appears that GB3 has to undergo a similar subtle conformational change going from the free-form, to the bound form in the GB3:Fab-complex (see Fig. 3). We superimposed the EM7 ensemble onto GB3 in the crystal structure of the GB3:Fab complex (PDB code, 1HG), and found some conformations that appear to be capable of forming the necessary hydrogen bonds for binding, see Fig. 4. In addition we calculated distances between the conformers of EM7 and GB3 in its free form (PDB code, 1GD) and bound form (PDB code, 1GC) using the Gauss integral based metric, GIT [34]. The GIT distances shown as a scatter-plot in Fig. 5. In general, the GIT metric appears to be a good measure to separate the two conformations due to the low correlation. This is considerably more difficult when using the RMSD measure in this example (not shown). We find that there are sets of structures which are close two one of the states but not the other, and vice versa. Although these results are preliminary, they may suggest that the bound conformation is in dynamic equilibrium with the free form in solution and that the binding to Fab occurs through conformational selection.

The correlated motions previously described for the 2LUM ensemble are present, yet clearly less pronounced (Fig. 2c). For example, the inter-strand correlations in the C-terminal β-hairpin do not show in the heatmap. In addition, this ensemble does not appear to accurately capture the the anti-correlated motion between the α-helix and the β-sheet.

Propagation of experimental information. Our results showed that the local experimental restraints significantly improve the quality of the resulting ensemble. Importantly, we do not observe severe distortions in pairwise atomic correlations by imposing the local restraints, which supports our theoretical considerations of the restrained ensemble as minimally biased. However, we find that measures of both short- and long-range concerted motions are in better agreement with the restrained ensemble as compared to the unrestrained ensemble. This suggests that local restraints propagate into subtle long range restraints when combined with the prior information embodied in a physical forcefield. This is something which is not immediately expected as RDCs provide angular information but not translational information.

In summary, we use the coarse Profasi forcefield restrained by a small set of experimental RDCs subject to spatial and temporal averaging as well as experimental noise, to generate an ensemble of the protein GB3. Through cross-validation, we find that the ensemble is in good agreement with complementary experimental data. We identify correlated motions similar to what has been previously reported. In addition we find a strongly anti-correlated motion between the two major secondary structure assemblies of the protein, compatible with the conformational change of GB3 when binding to its biological partner, Fab. This suggests that the binding mechanism to Fab may happen through conformational selection of a preformed state experimentally characterized in the free form of GB3.

These results suggest that it may be more generally possible to obtain realistic descriptions of the conformational variability through minimal biasing of native ensembles simulation of simple forcefields. In general, we cannot make assessments about the time-scales we consider in the ensembles, at this stage. Still, the results we obtain here are useful to understand the native fluctuations of biological macromolecules.

Materials and Methods

Posterior distribution The full posterior distribution is given by,
\[ p(f, e, x | d, V_i) \propto N(d | e, \sigma)G(f | e, V_i)\exp(-\beta E_{\text{profasi}}(x)) \]
where \( f \) is calculated from a conformational micro-state \( x \) using the alignment model described below. \( E_{\text{profasi}} \) and \( \beta \) is the Profasi forcefield described below and the inverse temperature, respectively. The experimental data and its uncertainty is represented by \( d \) and \( \sigma \), respectively. \( e \) is a variable which represents the average of the instantaneous representations \( f \). \( G(\cdot) \) is a log-linear model described in further detail below. The parameter \( V_i \) is estimated iteratively using an the EM algorithm described below.

Stochastic model of alignment. Here we introduce a new approach to modeling the five independent tensor components, \( S \), of folded proteins in restrained simulations. We make use of the fact that an ensemble average, \( \bar{e} \), is available during the simulation. One may equate \( \bar{e} \) to the experimental data in an ordinary maximum likelihood tensor estimator akin to what was presented earlier [35]. We have the linear system of equations, \( A\vec{s} = \vec{e} \) where \( A \) is a \( N \times 5 \) matrix of \( N \) bond vector projections of \( \vec{X} \) onto the molecular frame, or alignment tensor, \( S \). The variable \( \vec{s} \) is a \( 5 \times 1 \) vector of the alignment tensor components. Finally, \( \vec{e} \) is some sampled ensemble average of the experimental data. The system of equations may readily be solved for \( \vec{s} \), to obtain the maximum likelihood estimate \( \hat{S} \). From this we may obtain the "back-calculated experimental data", \( e' \), of the conformational micro-state given by the matrix \( A \) as \( \vec{f} = \hat{A}\vec{s} \). This approach has a number of immediate advantages. Firstly, it is independent upon the rotational reference frame. Secondly, it readily accommodates the possibility of anisotropic contributions to the observed experimental data. Finally, this approach does not involve a mechanistic simulation of the alignment tensor which renders it very computationally efficient [36, 37].

Log-linear model The prior \( \pi_{\text{f}}(f | e, V_i) \) is modeled by a log-linear model \( G(f | e, V_i) \) with a linear link function, \( \ell(A, b) = Ab \) [38]. The link function allows us to introduce the matrix \( V \) to ensure an appropriate prior in the space of \( f \). We have,
\[ G(f | e, V_i) = \exp(c + f\ell(V_i, e)) = \frac{\exp(V_i e)}{Z} \]
where we have chosen \( c \), to ensure that \( G(f | V_i, e) \) is normalized via \( Z \).

Profasi forcefield The Prof forcefield is parameterized in the torsional space of protein, eg. the \( \phi \) and \( \psi \) torsion angles of the backbone and \( \chi \)-torsions of the sidechain. All-atom specifications were achieved by assuming bond-angles and bond-lengths to be fixed [19]. The mathematical expression for the interaction potential consists of four terms
\[ E_{\text{profasi}}(x) = E_{\text{loc}}(x) + E_{\text{ev}}(x) + E_{\text{hb}}(x) + E_{\text{sc}}(x) \]
where, $E_{loc}(x)$, describes local interactions, that is, interactions between atoms separated through only a few covalent bonds. The remaining three terms account for non-local interactions, such as excluded-volume effects ($E_{ev}(x)$) and hydrogen-bonds ($E_{hb}(x)$). Finally, charge-charge and hydrophobic interactions between side chains is contained in the term $E_{hc}(x)$.

**EM Algorithm** The aim of the expectation maximization algorithm is to estimate the scale matrix, $V^*_E$, such that the optimal posterior probability distribution (eq. 1) is obtained. This is achieved by minimizing the expected minus log-posterior (MLPL) [22].

Initiate $V_0$ to some initial guess, here we use the zero matrix. This corresponds to an unstrained simulation where the marginal posterior probability distribution of $X$ and $F$ coincides with the prior distribution, $\pi_X(x)$ – here the Boltzmann distribution of the Profosi forcefield at 300K.

- **E-step** Samples $S = \{f, e, x\}$ were obtained from the posterior distribution $p(f, e, x \mid V, V_i)$ by Markov chain Monte Carlo. Here, we use the Metropolis-Hastings algorithm [39]. The posterior expectations of, $e$ and $f$ are estimated from as the sample means.

- **M-step** This step yields a new scale matrix $V_{i+1}$, through importance sampling [40]. With the $N$ samples $S = \{f, e, x\}$ from the posterior with $V_i$, and all other things being equal the expectations $\bar{e}$ and $\bar{f}$ for a given $V_{i+1}$ may be approximated by:

$$\bar{e}(\Delta) \approx \sum_{i \in S} e_{i} \exp(f_{i} \Delta e_{i})$$

and

$$\bar{f}(\Delta) \approx \sum_{i \in S} f_{i} \exp(f_{i} \Delta e_{i})$$

where $\Delta = V_{i+1} - V_i$. Here, the sample mean of $e$ is used as it is insensitive to changes in $V$.

We may now obtain $\Delta$ as,

$$\Delta = \arg \min_{\Delta} ||\bar{e}(\Delta) - \bar{f}(\Delta)||^2.$$ 

The minimization is carried out using a simple stochastic gradient descent heuristic:

- Propose new $V' = O(0, I)$, i.e. $\bar{e}(\Delta) = \bar{f}(\Delta)$.
- Evaluate new expectations according to equations above. Accept or reject new $V$ according to the Metropolis criterion [41].

The iterative procedure was repeated until convergence was assessed. The EM algorithm ran for 8 steps in total. Each E-step ran for 400000 MCMC steps and the M step is run for 15000 minimization steps. The MCMC simulation was carried out in the PHAISTOS framework [42] starting from pdb-file 20ED, and covers the space of $\{x, e, f\}$. Snapshots of $\{x, e, f\}$ were stored every 5000 step and used in the following M step. The sampling of the all-atom conformational micro states, $X$, involves local and non-local Monte Carlo moves, all of which fulfilled detailed balance [43]. A full EM step takes about 12hrs on a quad-core desktop computer.

**Final ensembles** The final ensembles $EM_i$, with $i = \{0, 1, 2, \ldots, 7\}$ were constructed by uniformly sub-sampling 5000 conformations of the 8000 snapshots saved during the E step to ease subsequent analyses. No significant differences were found between different subsampled ensembles.

**Cross correlated relaxation and hydrogen bond scalar couplings** For a macromolecule undergoing slow, isotropic tumbling the cross-correlated dipolar relaxation depends on the angle $\Theta$ between the between the two inter-atomic bonds, $X-Y$ and $U-V$, as:

$$R_{X-Y;U-V} = \frac{\gamma_X \gamma_Y}{\gamma_X + \gamma_Y} \frac{\mu_X}{\mu_Y} \frac{V_{X-Y}}{\Omega_{U-V}} \frac{\Omega_{U-V}}{\Omega_{X-Y}}$$

where $\gamma_X$ is the gyromagnetic ratio of nuclei $X$, $\mu$ is Planck's constant divided by $2\pi$, $\Omega$ is the permeability of free space, $\tau$ is the correlation time of the molecule and $\Omega_{X-Y}$ is the inter-atomic bond length between nuclei $X$ and $Y$. For $\tau$, we used experimentally determined values at $\Delta T = 4.1$ns for $R_{HN \cdot HN+1}$ and $4.0$ns for $R_{H20 \cdot H20+1}$, respectively.

The through hydrogen bond scalar couplings, $h_3 J_{NC'}$, between $^{14}$N-$^{13}$C nuclei were calculated using the empirical equations provided by Barfield [28]:

$$h_3 J_{NC'} = (-0.31 \cos^2 \Theta + [0.62 \cos^2 \Theta + 0.92 \cos \rho + 0.14]) \sin^2 \Theta$$

$$+ \frac{3}{2} (2R_{HO' \cdot H20'} - R_{HO' \cdot H20}) + 0.01$$

where $\tau_{HO' \cdot H20'}$ and $\rho$ are the $H \cdot O$ hydrogen bond length, $HO'$-angle and the $H-O'-C'$-N dihedral angle, respectively. The empirical constant $h_3^0$ is fixed at 1.760A.

**Training dataset details** A total of 413 previously published experimental RDCs were used to restrain the simulations in the EM algorithm. Uncertainties were uniformly assumed to be $1 \text{Hz}$.

**Acknowledgments.** We thank B. Vögel for kindly providing most of the experimental data on GB3 and LD Antov for editorial assistance. S.O. acknowledges funding from the Danish Council for Independent Research (FTP: 09-06654).

17. Yao L, Vögeli B, Ying J, Bax A (2008) Nmr determination of amide n-h equilibria, the H-O-C-D-H angle, and the H-O-C-D-H dihedral angle, respectively. The empirical constant $h_3^0$ is fixed at 1.760A.
Fig. 1. Plot of $-\log(p(\mathbf{f}, \mathbf{e}, \mathbf{x} | \mathbf{d}, \mathbf{V}_i))$ (MMLP) and the quality factor of the training set (Q-factor) versus iteration count, $i$, of the expectation maximization algorithm.

Fig. 2. Heatmaps of pairwise correlations, $\rho$, of fluctuations of $C^\alpha$ atomic positions along the backbone of GB3. (a) unrestrained Profasi (or EM$_0$), (b) restrained Profasi (or EM$_7$) and (c) the 2LUM ensemble. Correlation matrices were calculated using THESEUS [44].

Table 1. Reproduction of experimental data in various structural models of GB3$^{**}$.

<table>
<thead>
<tr>
<th></th>
<th>EM$_0$ $^*$</th>
<th>EM$_7$ $^\dagger$</th>
<th>2LUM$^\ddagger$</th>
<th>2OED$^\S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e$NOE$^-$6 ($\rho$)</td>
<td>0.6477</td>
<td>0.7247</td>
<td>0.8457</td>
<td>0.7870</td>
</tr>
<tr>
<td>$e$NOE$^-$3 ($\rho$)</td>
<td>0.6855</td>
<td>0.7795</td>
<td>0.8904</td>
<td>0.8675</td>
</tr>
<tr>
<td>$R_{HN_{i-1}HN_{i+1}}$ ($\rho$)</td>
<td>0.6291</td>
<td>0.9253</td>
<td>0.8966</td>
<td>0.8945</td>
</tr>
<tr>
<td>$R_{HN_{i-1}H_{\alpha_{i+1}}}$ ($\rho$)</td>
<td>0.6781</td>
<td>0.6665</td>
<td>0.6392</td>
<td>0.6438</td>
</tr>
<tr>
<td>$R_{H_{\alpha_{i}}H_{\alpha_{i+1}}}$ ($\rho$)</td>
<td>0.7623</td>
<td>0.7197</td>
<td>0.6896</td>
<td>0.7243</td>
</tr>
<tr>
<td>$h_3$ $J_{NC'}$ (rmsd, s$^{-1}$)</td>
<td>0.1538</td>
<td>0.1281</td>
<td>0.1421</td>
<td>0.1349</td>
</tr>
<tr>
<td>$3^J_{HN-H\alpha}$ (rmsd, s$^{-1}$)</td>
<td>0.9613</td>
<td>0.5548</td>
<td>0.3589</td>
<td>0.4455</td>
</tr>
<tr>
<td>$3^J_{HN-C\beta}$ (rmsd, s$^{-1}$)</td>
<td>0.5285</td>
<td>0.2747</td>
<td>0.3045</td>
<td>0.2759</td>
</tr>
<tr>
<td>$3^J_{HN-C'}$ (rmsd, s$^{-1}$)</td>
<td>0.6761</td>
<td>0.3776</td>
<td>0.3021</td>
<td>0.2724</td>
</tr>
<tr>
<td>$\langle$RMSD$\rangle$ (Å)</td>
<td>2.27</td>
<td>1.33</td>
<td>0.72</td>
<td>-</td>
</tr>
</tbody>
</table>

$^*$ Ensemble from unrestrained Profasi force field at 300K
$^\dagger$ Ensemble from RDC restrained Profasi force field at 300K
$^\ddagger$ refined against $e$NOE$^-$6, $3^J_{HN-H\alpha}$, $3^J_{HN-C\beta}$ and $3^J_{HN-C'}$
$^\S$ RDC refined x-ray crystal structure.
$^\S$[45]
$^\S$Average pair-wise $C^\alpha$ root-mean square deviation.
$^{**}$Pearsons correlation coefficient is denoted by $\rho$
Fig. 3. Superpositioning of the free (teal) and Fab bound (red) forms of GB3 [31].

Fig. 4. The restrained Profasi ensemble EM7 (green) superimposed onto GB3 in its binding site in Fab (blue) [31]. Figure created using PyMOL (DeLano Scientific LCC)
Fig. 5. Scatter plot of a Gauss integral-based distance between the conformations in EM7 and the bound (1IGC) and un-bound (1IGD) conformations [34].
Supplementary material for: Construction of protein structure ensembles in continuous space through minimal biasing of a simple physical force field.

Simon Olsson, * and Thomas Hamelryck *

* Bioinformatics Centre, Department of Biology, Faculty of Science, University of Copenhagen, Denmark

draft manuscript

Supplement

Reserved for Publication Footnotes
Table 1. Cross validation as a function of Expectation Maximization step †.

<table>
<thead>
<tr>
<th></th>
<th>EM₁</th>
<th>EM₂</th>
<th>EM₃</th>
<th>EM₄</th>
<th>EM₅</th>
<th>EM₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>eNOE − eNOE − 6 (ρ)</td>
<td>0.7093</td>
<td>0.7116</td>
<td>0.7079</td>
<td>0.7262</td>
<td>0.7232</td>
<td>0.7276</td>
</tr>
<tr>
<td>RHNᵢ−HNᵢ₊₁ (ρ)</td>
<td>0.8983</td>
<td>0.9086</td>
<td>0.9179</td>
<td>0.9051</td>
<td>0.9241</td>
<td>0.9263</td>
</tr>
<tr>
<td>RNᵢ−Hαᵢ (ρ)</td>
<td>0.6596</td>
<td>0.6677</td>
<td>0.6734</td>
<td>0.6684</td>
<td>0.6759</td>
<td>0.6653</td>
</tr>
<tr>
<td>RHαᵢ−Hαᵢ₊₁ (ρ)</td>
<td>0.7150</td>
<td>0.7212</td>
<td>0.7163</td>
<td>0.7283</td>
<td>0.7221</td>
<td>0.7169</td>
</tr>
<tr>
<td>h₃J_{NC'} (rmsd, s⁻¹)</td>
<td>0.1393</td>
<td>0.1315</td>
<td>0.1329</td>
<td>0.1309</td>
<td>0.1290</td>
<td>0.1237</td>
</tr>
<tr>
<td>³J_{HN−Hα} (rmsd, s⁻¹)</td>
<td>0.6632</td>
<td>0.6242</td>
<td>0.6225</td>
<td>0.5854</td>
<td>0.5226</td>
<td>0.5309</td>
</tr>
<tr>
<td>J_{HN−Cα} (rmsd, s⁻¹)</td>
<td>0.3494</td>
<td>0.3275</td>
<td>0.3124</td>
<td>0.3103</td>
<td>0.3021</td>
<td>0.2758</td>
</tr>
<tr>
<td>³J_{HN−C'} (rmsd, s⁻¹)</td>
<td>0.4652</td>
<td>0.4462</td>
<td>0.4247</td>
<td>0.4194</td>
<td>0.3655</td>
<td>0.3626</td>
</tr>
<tr>
<td>⟨RMSD⟩ (Å) *</td>
<td>1.39</td>
<td>1.50</td>
<td>1.46</td>
<td>1.31</td>
<td>1.30</td>
<td>1.42</td>
</tr>
</tbody>
</table>

*Average pair-wise Cα root-mean square deviation.
†Pearsons correlation coefficient is denoted by ρ
Chapter 3

Closing

3.1 Conclusion and Outlook

The work I have presented in this thesis addresses a central problem in structural biology: How do we construct sensible biomolecular models from incomplete, noisy and averaged data? – I have approached this question with a probabilistic mindset and have found some exciting new methods to achieve this. I will present a summary of the key findings below.

In the first paper, we presented a new implementation of a previously reported probabilistic method for protein structure determination, namely inferential structure determination. We made two conceptual changes to the original approach. Firstly, we employed more detailed structural prior knowledge embodied in the knowledge-based potentials TorusDBN and Basilisk. Secondly, we employed a different sampling scheme to efficiently traverse the conformational space. We found that the sampling from the posterior distribution converged rapidly compared to what had been reported for the previous approach. Also, the precision of the obtained structures improved. This was in particular the case for the local structure, due to the prior distributions employed. The key result of this study was that inferential structure determination had become viable on desktop computers, whereas computer clusters were needed before.

While the inferential structure determination (ISD) approach formally solves the problem of protein structure determination in a rigorous probabilistic framework, it inherently assumes that only a single conformation is reflected in the experimental data. In the second paper, we aimed to generalize the method to the inference of distributions of protein structures, rather than single structures. From first principles we found a hierarchical Bayesian network to be an accurate representation of spatially and temporally aver-
aged data subject to experimental noise. Importantly, we found ISD constitutes an intuitive limit of the model where the ensemble becomes infinitely narrow. Similarly, we found a recently proposed Maximum Entropy method for restraining molecular simulation to also constitute a special case of our model. We found that we were able to reconstruct a high-temperature ensemble of a small hairpin using simulated data subject to both linear and power averaging. We found the reference distribution to introduce significant bias in the case of power averaged data – this could however be alleviated by using the reference ratio method. While this study showed great promise on the proof-of-concept level, it remained unclear whether this method would be tractable in more complex cases.

In the third paper we applied the hierarchical Bayesian model presented in the second paper in the context of restraining native ensemble simulations with a physical forcefield. Here, the number of experimental restraints was more practically relevant, as was the size of the studied system. We found that the restrained ensemble reproduced complementary experimental evidence significantly better than the unrestrained simulation. In addition we found that the fluctuations present in the distribution structures of the restrained ensemble was in agreement with what had previously been proposed. Finally, we found a strong concerted motion which had not previously been reported for the free form of the system. These positive results suggests that the hierarchical model of averaged data is tractable to more realistic cases.

Discussions of conformational ensembles appear to take up an increasing amount of space on the pages of scientific journals, and rightfully so. However, the inference of single structures from experimental data remains an important tool, when conformational fluctuations are modest. Still, structure determination from NMR data has been largely limited by the need for expert assessments with regard to both assignment of resonances to atoms as well as adjustment of empirical parameters in the structure calculation itself. The ISD approach showed immediate promise to alleviate problems in the latter. In principle it could automatically assess both the quality of the data and make a balanced weighing with respect to prior information – this, in a rigorous probabilistic framework. Unfortunately, excessive computational demands prohibited most experimental labs from adapting this method. Our new implementation will hopefully attract a broader audience to apply this approach in a more routine fashion – in particular, in cases where the approach stands out compared to other methods, e.g. when the available experimental evidence is scarce. Future endeavors in this field could include the usage of other more universally accessible data, such as chemical
shifts. This has already been reported to be a fruitful direction, outside the context of ISD $^1$ $^2$. We are currently working on a full reimplementation of the presented results into the PHAISTOS framework (see section 4.3) to be made publicly available. In this connection we are already working on including chemical shifts into the structure calculation. Furthermore it would be interesting to experiment with different methods to account for potential correlations in experimental observables. Both of these endeavors may improve the quality and scope of probabilistic structure determination even further.

Over the past decades there has been steady progress in structural biology from considering primarily the singular, rigid structures to understanding the structure and its motions – in particular, within the past decade or so, with the emergence of intrinsically disordered proteins as a major class of functional biological molecules. Consequently, the focus on better understanding disorder, flexibility and dynamics in biological molecules and their relation to function has been invigorated $^3$. From my discussion of methods to characterize the flexibility of proteins using averaged experimental data it appears that using restrained simulations will be the most sensible endeavor. These provide a framework to bias approximative simulations such that they agree with experimental data, and may be iteratively refined as more experimental evidence becomes available. More specifically, the restraining procedure I presented in the second paper, and other similar approaches $^4$ $^5$, show great promise as to establish realistic descriptions of the distribution of biomolecular structure – see for example the preliminary results in my third paper. Going forward it would be very interesting to apply these methods to obtain distributions of describing the conformational variability in more

---

$^1$Andrea Cavalli, Xavier Salvatella, Christopher M. Dobson, and Michele Vendruscolo. *Protein structure determination from NMR chemical shifts* (2007) PNAS 104(23) p.9615

$^2$Oliver F. Lange, Paolo Rossi, Nikolaos G. Sgourakis, Yifan Song, Hsiau-Wei Lee, James M. Aramnic, Asli Ertekin, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione and David Baker. *Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples* (2012) PNAS 109(27) p. 10873

$^3$Kaare Teilum, Johan G. Olsen and Birthe B. Kragelund. *Protein stability, flexibility and function* (2011) Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1814(8) p.969


$^5$Andrea Cavalli, Carlo Camilloni and Michele Vendruscolo. *Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle* (2013) JCP 138(9) p. 094112
dynamic systems such as molten globule or intrinsically disordered proteins. This is something I have already started working on.
3.2 Acknowledgments

Firstly, I’d like to thank my friends and family for their continued support and encouragements during the work carried out over the past three years. Secondly, I’d like to thank my supervisor Thomas Hamelryck for inviting me to a PhD in his group and supporting me throughout the process. Third, the rest of the staff and alumni of the Bioinformatics center, in particular Wouter Boomsma, Jes Frellsen, Jan Valentín, Lubomir Antonov, Kristoffer Eneé, Mikael Borg, Christian Andreetta, Tim Harder, Kasper Stovgaard and Martin Paluszewski are thanked. Also to the former DTU group: Jesper Ferkinghoff-Borg, Sandro Bottaro and Pengfei Tian and SBiN-lab members Magnus Kjærgaard, Robert Dagil, Kaare Teilum, Kresten Lindorff-Larsen and in particular the late Flemming M. Poulsen, are thanked for their time and encouraging scientific discussions and great collaborations. Also Beat Vögeli at the ETH in Zürich, Switzerland is thanked for interesting discussion about NMR and modeling of ensemble data.

Finally, I’d like to thank Douglas L. Theobald and his group a welcoming and stimulating stay at Brandies in July of 2012.

The work carried out in this thesis was funded by a grant from the Danish Council for Independent Research (FTP, 09-066546).
Appendix A: Technical note on adaptive kernel proposals

Given a non-linear function $f(x)$ with $x = \{x_i\}_{i=1}^{N}$ the its expected change $\partial f$ under a given uncertainty $\sigma(x)$ is given by the propagation of uncertainty,

$$\partial f = \sqrt{\sum_{i=1}^{n} \left[ \frac{\partial f}{\partial x_i} \sigma(x) \right]^2}. \quad (4.1)$$

When only one parameter, $x_i$, is updated at the time we can without any loss of generality simplify this to:

$$\partial f_i = \|\frac{\partial f}{\partial x_i}\| \sigma(x), \quad (4.2)$$

where $\partial f_i$ corresponds to the expected change in $f$ under the variance $\sigma^2$ in $i$. We can now chose a desired change in $f_i$, as $\partial f_i = \epsilon$ and isolate an expression of the standard deviation $\sigma$. We obtain,

$$\sigma(x) = \frac{\epsilon}{\|\frac{\partial f}{\partial x_i}\|}. \quad (4.3)$$

Thus, equation 4.3 yields an optimal choice of standard deviation of our proposal distribution of a variable $x_i$ given a desired change in $f_i$, that is, energy, and the partial derivative of $f$ with respect to that variable. Special care must be taken as to assure the change in the proposal distribution is accounted for in the Metropolis-Hastings criterion.
4.2 Appendix B: TYPHON

This paper describes the combination of three advancements made in the
groups of Thomas Hamelryck and Jesper Ferkinghoff-Borg (TorusDBN, Basilisk
and CRISP) with a network of Gaussian restraints, to yield a probabilistic
equivalent of the popular flexibility predictor tConcoord. In spite of its
simplicity, the model exerts good agreement with experimental data while
keeping the local structure realistic.

This is a co-author article. It was published in Structure in 2012. I was
involved in analysis of the results and wrote large parts of the manuscript.
An Efficient Null Model for Conformational Fluctuations in Proteins

Tim Harder, 1 Mikael Borg, 1 Sandro Bottaro, 2 Wouter Boomsma, 2,3 Simon Olsson, 1 Jesper Ferkinghoff-Borg, 2 and Thomas Hamelryck 1,*
1The Bioinformatics Section, Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark
2DTU Elektro, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
3Departments of Astronomy and Theoretical Physics, Lund University, SE-223 62 Lund, Sweden
*Correspondence: thamelry@binfo.ku.dk
DOI 10.1016/j.str.2012.03.020

SUMMARY

Protein dynamics play a crucial role in function, catalytic activity, and pathogenesis. Consequently, there is great interest in computational methods that probe the conformational fluctuations of a protein. However, molecular dynamics simulations are computationally costly and therefore are often limited to comparatively short timescales. TYPHON is a probabilistic method to explore the conformational space of proteins under the guidance of a sophisticated probabilistic model of local structure and a given set of restraints that represent nonlocal interactions, such as hydrogen bonds or disulfide bridges. The choice of the restraints themselves is heuristic, but the resulting probabilistic model is well-defined and rigorous. Conceptually, TYPHON constitutes a null model of conformational fluctuations under a given set of restraints. We demonstrate that TYPHON can provide information on conformational fluctuations that is in correspondence with experimental measurements. TYPHON provides a flexible, yet computationally efficient, method to explore possible conformational fluctuations in proteins.

INTRODUCTION

Over the past few decades it has become increasingly accepted that proteins are dynamic molecules. Although many proteins adapt unique and specific folds, their inherent flexibility is often essential to the protein’s function. However, flexibility can also lead to pathogenesis through misfolding, possibly leading to the formation of aggregates and fibrils (Dobson, 2003; Teilmann et al., 2009a).

Computer simulations have emerged as important tools to study the dynamics of proteins, complementing the data obtained from biophysical experiments. A variety of methods are available, ranging from detailed all-atom molecular dynamics (MD) simulations (McCammon et al., 1977; Kaprillus and McCammon, 2002; Hess et al., 2008) to coarse-grained and approximate methods, such as normal mode analysis (NMA; Levitt et al., 1983), elastic networks (Zheng et al., 2007), tCONCOORD (de Groot et al., 1997; Seeliger and De Groot, 2009), and FRODA (Jacobs et al., 2001; Wells et al., 2005). All methods come with a trade-off between the level of detail and the computational cost for obtaining useful information.

The concept behind MD simulations is to approximate the physical forces acting on a protein and to calculate the motion of particles in the system by applying Newton’s laws of motion (McCammon et al., 1977; Kaprillus and McCammon, 2002; Hess et al., 2008). Because the calculation of these physical forces is computationally expensive, MD simulations are usually limited to short timescales—typically in the range of hundreds of nanoseconds. The high level of detail in MD simulations makes general physical conclusions viable (van Gunsteren et al., 1996; Brooks et al., 2009). However, the timescales routinely accessible through MD simulations rarely cover the full dynamic range of proteins. Coarse-grained MD simulations sacrifice certain atomic details to gain a computational advantage, thus allowing longer simulation times or simulations of larger systems. Merging multiple atoms into so-called beads or pseudoatoms is a common approach to reduce the number of particles in the system (Marrink et al., 2007). Another solution to overcome the computational cost of MD simulations is to use faster computer hardware. Shaw and colleagues were able to achieve a millisecond simulation using custom built special-purpose hardware (Klepeis et al., 2009; Shaw et al., 2010).

Many faster heuristic alternatives to MD have been developed. The idea behind the elastic network (EN) models is that the dynamics of folded, native proteins are rather limited compared to unfolded dynamics and are overall governed by the interresidue contact topology (Bahar and Rader, 2005). Over the past years, the computationally efficient EN models have replaced the original harmonic potentials in many NMA approaches (Bahar and Rader, 2005; Yang et al., 2009). In EN models, the protein’s atoms are viewed as point masses that are interconnected by springs. Often, only the backbone Cα atoms are included. Subsequently, a number of conformations are sampled and a principal component analysis is performed on the generated ensemble, yielding the normal modes (Levitt et al., 1983). However, ensembles sampled from EN models can be also used in different scenarios (Zheng et al., 2007); vice versa, normal modes can be also calculated from ensembles generated in MD simulations (Hess et al., 2008).

Other heuristic approaches that include atomic detail have gained popularity over the past years. FRODA (Jacobs et al., 2001; Wells et al., 2005) identifies rigid substructures in the
protein structure to reduce the degrees of freedom for the subsequent simulation. Another widely used heuristic tool is tCONCOORD (de Groot et al., 1997; Seeliger et al., 2007; Seeliger and De Groot, 2009), which has been successfully applied in different contexts (Zachariae et al., 2008; Seeliger and de Groot, 2010). Here, the input structure is analyzed to create a network of constraints. Subsequently, tCONCOORD randomly perturbs the atom coordinates within a box around their initial positions in the native structure. Then, a Monte Carlo procedure changes the perturbed atomic positions until they again satisfy the constraints. In this procedure, the atomic positions are subject to changes sampled from a uniform distribution. Consequently, all the information is encoded in the constraint network; in the absence of constraints, there is no information on how to arrange the atoms.

Here, we present TYPHON, which adopts a probabilistic approach to exploring conformational fluctuations in proteins. TYPHON is based on two recent innovations: TorusDBN (Boomsma et al., 2008) and BASILISK (Harder et al., 2010). TorusDBN and BASILISK are probabilistic models of the conformational space of a protein’s main chain and its amino acid side chains, respectively. Both models are formulated as dynamic Bayesian networks (DBNs) and make use of directional statistics (Mardia and Jupp, 2000)—the statistics of angles and directions—to represent protein structure in a natural, continuous space (Hamelryck et al., 2006; Boomsma et al., 2008; Harder et al., 2010). Together, TorusDBN and BASILISK constitute a probabilistic model of protein structure in atomic detail. This model is generative; plausible protein conformations can be efficiently sampled. Furthermore, TYPHON incorporates CRISP (Bottaro et al., 2012), an efficient method for applying local modifications to the protein’s conformation.

The application of these probabilistic models in TYPHON ensures that the structure remains protein-like on a local length scale throughout the conformational sampling. The long-range structure is maintained by imposing different types of distance-based restraints, which are heuristic representations of nonlocal interactions, such as hydrogen bonds. TYPHON uses Gaussian distributions to implement the restraints, resulting in a valid probabilistic description of the restraint network and the local structure of proteins. This well-justified probabilistic formulation differs from previous ad hoc approaches. TYPHON explores the conformational space accessible to a protein, within the limits imposed by the restraint network. In the absence of a restraint network, sampling is solely guided by the probabilistic models and results in an ensemble of extended conformations with realistic local structure, conceptually reminiscent of an “unfolded state.”

In short, TYPHON can be considered a null model of conformational fluctuations, given a set of probabilistic restraints. We again stress that our method is well justified, given a chosen set of restraints; the biological relevance of the obtained conformations will necessarily depend on the relevance of the heuristic restraints. However, TYPHON provides default restraints, which typically deliver good results for common applications, as discussed below.

In the following, we compare results obtained from TYPHON with experimental measures describing the native ensemble of folded proteins, including B-factors, nuclear magnetic resonance (NMR) order parameters, and residual dipolar couplings (RDCs). The different measures allow us to investigate how well TYPHON captures the flexibility of a folded protein. We then demonstrate how local unfolding caused by the loss of metal ions is correctly modeled by TYPHON. Finally, we show how fluctuations of local structure can be investigated under the control of the probabilistic models, which is an additional attractive and innovative aspect of our approach.

RESULTS

Overview of TYPHON

TYPHON samples protein structures from a joint probability distribution that includes local and nonlocal interactions (described in more detail in the Experimental Procedures). TYPHON incorporates several sophisticated probabilistic models to maintain the local structure and uses simple Gaussian restraints to maintain relevant nonlocal interactions. Although the choice of these nonlocal restraints is heuristic, the resulting joint probabilistic model is well defined and rigorous. In other words, if a suitable restraint network can be chosen for the problem of interest, TYPHON will typically deliver good results, obtained from a well-defined probability distribution.

By default, TYPHON automatically detects the hydrogen bond network. The geometry of the individual hydrogen bonds is restrained using a simple model based on four distances modeled by Gaussian probability distributions. Disulfide bridges are, by default, treated in a similar way. By default, TYPHON also restraints all distances between Cα atoms that are five or more residues apart in the amino acid chain and within six Å of each other. The latter restraints aim to capture general interactions that stabilize the protein, such as the hydrophobic effect.

The user can manipulate and verify the restraint network. For example, it is possible to disregard all hydrogen bonds involving side chains or to add or remove restraints between arbitrary atom pairs. In this manuscript, we use different restraint networks to answer different questions. These networks range from involving Cα atoms (see Experimental B-Factors) over hydrogen bonds (see Generating a Native Ensemble) to a small number of disulfide bridges (see Local Structure under the Control of Probabilistic Models).

TYPHON is obviously limited with respect to modeling the formation and dissolution of nonlocal interactions themselves, as the restraint network is fixed throughout the sampling procedure. However, the secondary structure can be, to some extent, put under the control of the probabilistic models (see Local Structure under the Control of Probabilistic Models), allowing for formation and dissolution of certain hydrogen bonds, notably, in helices.

Experimental B-Factors

The Protein Data Bank (PDB: Berman et al., 2000) currently contains over 77,000 solved structures; the majority of them are determined by X-ray crystallography. Experimental B-factors associated with the atoms of a crystal structure often give a first indication of the conformational fluctuations within a protein. The B-factor reflects both the thermal vibrations of single atoms and small structural differences between molecules in the crystal. The latter contribution is of interest for inferring protein flexibility.
In this test, we analyze whether TYPHON is able to reproduce the flexibility that is indicated by the B-factors of a protein.

TYPHON makes it possible to sample an ensemble of structures that is close to the native structure. We illustrate this with the crystal structure of the 317-residue-long protein Candida antarctica Lipase B (CalB; PDB: 1tca; Uppenberg et al., 1995). CalB is an enzyme with industrial applications that adopts an a/b-fold. A short helix, consisting of residues 139 to 147, is suspected to act as a flexible lid that is important for catalysis, making it a prime subject of dynamics studies (Skjøt et al., 2009). For comparison, we translated the experimental B-factors of the crystal structure into root-mean-square fluctuations (rmsf) using the following relation (Kuzmanic and Zagrovic, 2010):

$$\text{RMSF}_{i} = \frac{3B_i^2}{2\pi^2}$$

where $B_i$ is the B-factor for the $i$-th residue.

TYPHON used the crystal structure as sole input, from which Gaussian distance restraints were derived (see the Experimental Procedures). The sampling ran for 50 million iterations. Figure 1 shows RMS fluctuation calculated from the experimental B-factors for the crystal structure and from 1,000 sampled conformations chosen with regular intervals. The overall flexibility along the sequence is well captured. The lid region clearly displays a higher level of flexibility in correspondence with its dynamic nature (Skjøt et al., 2009). The good agreement with the experimental measure is also reflected in the Pearson correlation coefficient, which is equal to 0.71.

**Generating a Native Ensemble**

Advances in nuclear magnetic resonance (NMR) spectroscopy over the past decades made more detailed studies of dynamics in proteins possible. The $S^2$ order parameter is a measure arising from NMR experiments describing the amplitude of motion of an N-H vector (Lipari and Szabo, 1982). A backbone segment that is unrestricted in its movement, usually in a region of high flexibility, will have a low $S^2$ value. For segments in more constrained or rigid regions of the protein, the $S^2$ value will be higher. Analyzing $S^2$ order parameters provides a more direct view on the dynamics of a protein compared to the B-factors. In this test, we analyze whether TYPHON is able to capture the fast dynamics of a protein as implied by the $S^2$ order parameters.

Ubiquitin is a well-studied protein in terms of its dynamics; its relatively small size of 76 amino acids allows for both extensive MD simulations as well as NMR studies. Ubiquitin consists of a five stranded, twisted, and antiparallel $\beta$ sheet with an a-helix lying across. A number of recent publications discuss the molecular recognition mechanisms using ubiquitin as a model system (Lange et al., 2008; Wlodarski and Zagrovic, 2009; Long and Brüschweiler, 2011).

TYPHON sampling started from a single crystal structure of ubiquitin (PDB: 1ubi; Ramage et al., 1994), with 46 automatically detected hydrogen bonds as restraints, and ran for 50 million iterations. A total of 1,000 structures were sampled in regular intervals. We also generated an ensemble of 1,000 structures using tCONCOORD, starting from the same ubiquitin crystal structure and using default settings. For further comparison, we also included the order parameters calculated from an MD simulation of ubiquitin (Maragakis et al., 2008).

Figure 2 shows the $S^2$ order parameters calculated from the TYPHON ensemble following Best and Vendruscolo (2004) and the order parameters obtained from an experiment by Tjandra et al. (1995). The figure further shows order parameters calculated from a tCONCOORD ensemble obtained with default parameters and from an MD simulation (Maragakis et al., 2008). Overall, the $S^2$ parameters calculated from the TYPHON ensemble following Best and Vendruscolo (2004) and the order parameters obtained from an experiment by Tjandra et al. (1995) agree well.
ensemble are in good agreement with the experimental measurements; the correlation coefficient for the two curves is 0.73. The most rigid region is located in the well-ordered $\alpha$-helix between residues 23 and 33. This region is indeed rigid in the TYPHON ensemble as well though overly so compared to the experimental results (Tjandra et al., 1995). The terminal regions are the most flexible (see Figure 2). Recently, it was found that the increased flexibility in the C-terminus and in loop I between the I1 and I2 strands is of importance for the molecular recognition mechanism of ubiquitin (Lange et al., 2008; Wlodarski and Zagrovic, 2009). The ensemble generated by TYPHON accurately reflects the conformational fluctuations in these regions of interest.

The order parameters calculated from the MD simulation match the experimental values less well; the correlation coefficient is 0.52. Although the MD ensemble accurately reflects the flexibilities in loop I, it does not well reproduce the fluctuations in the C-terminus. The $S^2$ order parameters calculated from the tCONCOORD ensemble match the general trend of the experimental curve. The correlation coefficient is 0.53, which is also lower than for TYPHON. The generated ensemble appears to overemphasize the flexibility in certain loops, including the functionally important loop I—around residues 7 to 10. In addition, loop V—around residues 63 to 65—shows considerable discrepancy. Leaving out the flexible C-terminal region, following Lindorff-Larsen et al. (2005), results in correlation coefficients equal to 0.50, 0.55, and 0.28 for the MD, TYPHON, and tCONCOORD ensembles, respectively. In conclusion, TYPHON matches the experimentally determined order parameters, indicating that the fast dynamics—as described by the Lipari-Szabo $S^2$ parameters—are well captured in the generated ensemble.

Residual dipolar couplings (RDCs) probe the bond vector geometry relative to an external magnetic field. Data acquisition in a nematic phase solvent or in the presence of a paramagnetic center can make measurement of RDCs in the solution state possible (Tjandra et al., 1997; Banci et al., 2004). RDCs are anisotropic quantities and thus average out when molecules undergo isotropic rotational diffusion.

For ubiquitin, Cornilescu et al. (1998) obtained six sets of backbone RDCs in a nematic phase solvent based on phospholipid bicelles. The experimental data was obtained from the Biological Magnetic Resonance Data Bank (BMRB entry: 6457; Ulrich et al., 2008). We used the same TYPHON and tCONCOORD ensembles as in the previous section. Ensemble averages were calculated from these ensembles using the procedure described by Showalter and Brüschweiler (2007; Lindorff-Larsen et al., 2005).

Figure 3 shows experimentally determined $C_{\alpha} - CO$ RDCs in comparison with RDCs that are calculated from a TYPHON ensemble. Figure S1 (available online) additionally shows correlation plots for all RDCs. In general, there is a good correlation between the values obtained from the TYPHON and the experimental data (see Table 1). The agreement with experiment for the TYPHON ensemble is comparable to the tCONCOORD ensemble and the crystal structure (1UBI). However, Q-factors for the TYPHON ensemble (0.37) are larger than for the tCONCOORD ensemble (0.28) and the crystal structure (0.23), suggesting better qualitative agreement of the tCONCOORD ensemble (Lipsitz and Tjandra, 2004).

Although the reproduction of experimental data, such as residual dipolar couplings and order parameters, serves as a sanity check, it is difficult to make a quantitative assessment of the physical timescales sampled (Showalter and Brüschweiler, 2007). However, collectively, the results suggest that TYPHON samples broader ensembles in some regions of ubiquitin as compared to tCONCOORD. Regions that appear overstabilized may be attributed to the employed restraints, suggesting that TYPHON ensembles can be improved by input of expert knowledge. In view of the excellent structural quality of the generated decoys (compare section Quality of the Sampled Structures), these observations support the interpretation of TYPHON as a suitable “null model” of conformational fluctuations in proteins for a given set of restraints; given the nonlocal

![Figure 3. C_{\alpha} - CO RDC Values for Ubiquitin](image)

The figure shows a comparison between experimentally determined $C_{\alpha} - CO$ RDCs (green line) and RDCs calculated from a TYPHON ensemble using the procedure described in Showalter and Brüschweiler (2007), (blue line), where the RDCs are plotted on the y axis against the residue index on the x axis. See also Figure S1.

### Table 1. Statistics for the RDC Values Obtained from the TYPHON and tCONCOORD Ensembles of Ubiquitin

<table>
<thead>
<tr>
<th></th>
<th>N - NH</th>
<th>CO - NH</th>
<th>$C_{\alpha} - H_{\alpha}$</th>
<th>N - CO</th>
<th>$C_{\alpha} - CO$</th>
<th>$C_{\alpha} - C_{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation coefficient average RDC TYPHON$^a$</td>
<td>0.91</td>
<td>0.90</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.90</td>
</tr>
<tr>
<td>Correlation coefficient average RDC tCONCOORD$^a$</td>
<td>0.96</td>
<td>0.91</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>Correlation coefficient Crystal structure (1UBI)$^b$</td>
<td>0.98</td>
<td>0.96</td>
<td>0.93</td>
<td>0.99</td>
<td>0.99</td>
<td>0.97</td>
</tr>
</tbody>
</table>

$^a$Correlation coefficients of the TYPHON and tCONCOORD ensembles with the experimental data, respectively.

$^b$Correlation between the crystal structure 1UBI and for all six RDC types.
restraints, the probabilistic models of local structure ensure a thorough exploration of the remaining conformational space.

**Functional Dynamics of an Enzyme**

Ribonuclease (RNase) A is a pancreatic protein that cleaves single-stranded RNA; its structural dynamics are essential for its enzymatic function (Doucet et al., 2009; Formoso et al., 2010). The protein has 124 residues and adopts a β/α fold that consists of two domains flanking a catalytic site. In this experiment, we analyze whether TYPHON can reproduce the functional dynamics of RNase A. In addition, we compare the TYPHON ensemble to results obtained from NMA.

We initialized TYPHON sampling from the RNase A crystal structure (PDB: 7RSA; Wlodawer et al., 1988) and used the automatically detected hydrogen bond network with default settings, resulting in 76 hydrogen bonds and four disulfide bridges. The sampling was run for 100 million iterations, from which 1,000 structures were retained.

As a measure of the structural flexibility of RNase A, we analyzed 132 experimentally determined structures with a maximum of one point mutation (for a complete list see Table S1). We superimposed the experimental structures using iterative root-mean-square deviation (rmsd) minimization to the average structure and calculated the rmsf of the C\_α atoms. We call this set the high-sequence similarity PDB ensemble (Best et al., 2006).

In addition, we compare our result to the dynamics of the enzyme according to the elastic network model (ENM), a coarse-grained model of protein dynamics that has been used to analyze collective motions, residue fluctuations, and conformational changes (Tirion, 1996; Hinsen, 1998; Bahar and Rader, 2005; Ma, 2005; Kimber et al., 2010). In the ENM, the protein structure is approximated as a network of coupled harmonic oscillators between all C\_α atoms closer than a specified cutoff radius. The collective motions of the system can be then calculated using NMA. The ENM analysis was performed with the elNémo server and default parameters, using an 8 Å cutoff distance to identify elastic interactions (Suhre and Sanejouand, 2004). The server reports the rmsf calculated from the scaled Eigen vectors of the first hundred modes.

The fluctuations found within the PDB and the TYPHON ensembles (Figure 4) are in good agreement; the correlation coefficient is 0.72. The overall flexibility pattern along the amino acid chain indicates increased mobility in the same regions. The amplitude of the fluctuations is, however, significantly larger for the TYPHON ensemble, indicating that a large volume of the conformational space is sampled. This again confirms the interpretation of TYPHON as a suitable null model of conformational fluctuations for a given set of restraints. Notably, loop I—consisting of residues 14 to 25—has a high degree of flexibility (Figure 4). The dynamics of this loop are especially important for the catalytic activity of the enzyme (Doucet et al., 2009; Formoso et al., 2010). TYPHON sampling started from other crystal structures of RNase A in the PDB yielded similar results (PDB codes 3LXO [Doucet et al., 2010] and 2G8O [Leonidas et al., 2006]). In contrast, although having only a slightly lower correlation coefficient to the PDB ensemble (0.67), the result from the ENM analysis does not show an elevated flexibility in this loop.

The dynamics of loop I is a requirement for the functional dynamics of RNase A; RNase A has been shown to function through a concerted motion between an open form that can bind substrate and a closed form, where catalysis occurs (Watt et al., 2007). To investigate how the TYPHON ensemble relates to these motions, we performed a principle component analysis on the TYPHON samples and isolated the main modes. The first mode, which contains the most important variations of the ensemble, indeed shows an opening and closing of the catalytic cleft, lending further evidence that the TYPHON ensemble can be used to explore enzyme dynamics. A video of the motion is available online (see Movie S1).

**Induced Change in Flexibility**

Large-scale motions or major changes in flexibility in proteins are often induced by binding or releasing ligands. These ligands can be as complex as multiantom substrates, inhibitors, or drugs or as simple as single metal ions. In this test we use TYPHON to simulate partial unfolding upon loss of metal ions. This application illustrates how the probabilistic models “step in” to provide information in the absence of restraints.

Cu/Zn superoxide dismutase (SOD1) is a ubiquitous protein in the cytoplasm that is associated with the neurodegenerative disease amyotrophic lateral sclerosis (ALS). ALS results in paralysis and respiratory failure within one to five years from onset (Pasinelli and Brown, 2006). The oligomerization of SOD1 is associated with a gain in toxic function. Experimental evidence suggests that a loss of the two metal ions induces structural changes to the monomeric form of SOD1 and subsequently leads to pathogenic aggregation (Teilum et al., 2009b). However, the exact pathway is still unknown. We used the PDB:2v0a crystal structure as starting point for our experiments (Strange et al., 2007). SOD1 consists of a β barrel with long loops connecting the antiparallel strands. It contains a disulfide bridge and has
two associated metal ions: a copper ion that is coordinated by four histidines (residues 46, 48, 63, and 120) and a zinc ion that is coordinated by three histidines and an aspartate (residues 46, 48, 63, and 120). The spike in flexibility around residue 57 can be attributed to the reduced disulfide bridge, which, in the native structure, covalently binds this surface loop. The increased flexibility in other parts of the protein is likely due to the loss of the metal ions. An interesting observation is also the increased flexibility in loop II around residue 25, which is not in direct contact with any of the mutated sites. We speculate that the overall increased mobility in the long loop IV and VI also influenced the flexibility in this region.

The results closely resemble those of Ding and Dokholyan (2008), which were obtained from discrete MD simulations. A TYPHON experiment requires about 20 hours, which would allow scanning of larger sets of clinically known mutations (Andersen et al., 2003). We point out that the increased mobility in loop II was not observed in the MD study of Ding and Dokholyan (2008), which illustrates that TYPHON can deliver results that suggest starting points for new hypotheses or follow-up studies. It should be noted that TYPHON only includes the steric component of the ion loss; changes in electrostatics or solvent accessibility are not directly accounted for. Nonetheless, in this case, modeling the effect of the metal ions as simple Gaussian restraints accurately reproduces the results obtained from much more sophisticated simulations and leads to potentially interesting and new observations.

**Local Structure under the Control of Probabilistic Models**

The Gaussian restraints obviously do not allow for formation or dissolution of nonlocal interactions; the restraint network is rigorously fixed during the sampling procedure. However, certain nonlocal interactions, such as hydrogen bonds in helices, can be put under the control of the probabilistic models instead. In practice, this means that certain conformational fluctuations of the protein backbone on a local length scale could be investigated. In this application, we explore and illustrate this approach with a small helical protein and investigate helical mobility and $\alpha/3_{10}$-helix transitions.

The Mature T Cell Proliferation Gene 1 (MTCP1) is a known oncogene that is linked to certain types of leukemia (Barthe et al., 2002). The structure of the human p110$\gamma$-MTCP1 protein has been solved by NMR and consists of three helices.
A stable α hairpin connecting helix I and II is covalently held together by two disulfide bridges between residues 7, 38, and 17, respectively. A third less restricted and stable helix (helix III) is also connected to helix II with a third disulfide bridge between residues 39 and 50 (Barthe et al., 1997). MD simulations indicate that helix III is fairly flexible with respect to the α hairpin (Barthe et al., 2002).

We first investigate to what extent the helices move with respect to each other. We therefore started from the first model of a p8MTCP1 NMR ensemble (PDB: 2hp8; Barthe et al., 1997). The experiment ran for 100 million iterations with the three disulfide bridges as only restraints. However, we also imposed the secondary structure of the native structure according to DSSP (Kabsch and Sander, 1983) through TorusDBN (Boomsma et al., 2008). This is a more flexible and “soft” way to restrain the sampling, as the helical regions are allowed to bend or, to a certain extent, form and dissolve hydrogen bonds under the influence of the probabilistic model.

Despite the absence of restraints, besides those involving the three disulfide bridges, all helices remain stable throughout the sampling. Figure 6A shows five representative structures from the ensemble. Helix I and helix II are tightly fixed by the interhelical disulfide bridges, which only allow limited movements. Helix III is only tethered by a single disulfide bond in the beginning of the helix, which results in higher flexibility. As indicated in Figure 6A, helix III slightly tilts away from the other two helices, a behavior that also has been observed in MD simulations (Barthe et al., 2002).

Figure 6C shows the secondary structure content over the course of the first experiment. The consistent red bars show that all three helices remain fully helical throughout the sampling. In the beginning of helix III, we observe transitions between α- and 310-helix, which is again in agreement with the results of a MD simulations (Barthe et al., 2002).

In the second experiment, we investigate the stability of the helices themselves. We again included restraints concerning the three native disulfide bonds. However, this time we did not provide any secondary structure information to TorusDBN. In other words, this means that TorusDBN still enforces protein-like conformations but does not require them to be helical.

Again helix I and helix II remain stable throughout the sampling as indicated by the consistent red bars in Figure 6D. This is not surprising because both helices are covalently connected near their respective start and end. The entire protein structure is, however, significantly more flexible, expressed by the movement of the helices with respect to each other (compare Figure 6B). In contrast to helices I and II, helix III quickly unfolds up to residue 50, where it is covalently attached to helix II via a disulfide bridge.

In addition to the unfolding helix III, we observe significant differences compared to the first experiment in the loop regions. In particular, for loop II, which connects helix II and III and stretches from residue 39 to 47, we observe a transition to an α-helix. The terminal 18 residues of helix III readily unfold (see Figure 6D), which points to a difference in stability between the first two and the third helix.

This experiment strikingly demonstrates the possibilities of probabilistic models. In the first experiment, which includes the disulfide bridges and secondary structure information, we observed specific movements of the helices with respect to each other and transitions from an α- to a 310-helix in the beginning of helix III. Both observations concur with the results obtained from MD simulations (Barthe et al., 2002). In the second experiment, which includes the disulfide bridges but not the secondary structure information, we obtained some information on the relative stability of the helices themselves. Helices I and II remain stable, whereas helix III readily unfolds. Again, this difference in stability is in accordance with MD simulations (Barthe et al., 2002).

**Quality of the Sampled Structures**

To evaluate the quality of the structures, we analyzed 50 random structures from an RNase A ensemble, generated as described previously, using PROCHECK (Laskowski et al., 1993). For comparison, we generated 50 tCONCOORD (Seeliger et al., 2007) samples for the same protein (starting from PDB: 7rsa). The
detailed PROCHECK reports can be accessed as Documents S2 and S3.

The Ramachandran map divides the main chain’s conformational space, as parameterized by the $\phi$ and $\psi$ angles, in different regions, some sterically more favorable than others (Ramachandran et al., 1963). Well-refined protein structures are expected to have 90% or more of the backbone dihedral angles in the most favorable regions. The PROCHECK analysis indicates that the TYPHON samples are of good quality; over 88% of all angles are in the most favored regions. In contrast, the tCONCOORD samples have less than 70% of the backbone angles in these favored regions (Table 2).

PROCHECK’s G factor is a measure of how well the analyzed structures match the observed distributions of bond lengths, bond angles, and dihedral angles in crystal structures and is expected to be $\approx 0.5$ or higher for well-refined structures. Also in this respect, TYPHON samples have a higher quality than do tCONCOORD samples; the G-factor is $-0.13$ versus $-0.69$. The G factor takes the side-chain quality into account; in this respect, TYPHON undoubtedly benefits from the detailed side-chain modeling in BASILISK (Harder et al., 2010).

Additionally, we performed a WHATIF (Vriend, 1990) packing analysis of the TYPHON and tCONCOORD ensembles of RNase A. The structures generated by TYPHON have an average packing environment score of $-1.495$. Those generated by tCONCOORD have an average score of $-1.944$. As well-refined structures have a score around $-0.5$, both methods might be improved in this respect.

### Computational Efficiency

TYPHON is computationally efficient. The ubiquitin experiments used in this study were performed on a regular desktop computer (Intel Core i7, 2.8GHz) and ran for around 10 hr on a single CPU core. The human p8MTCP1 protein experiments comprising 100 million iterations ran for about 15 hr. Naturally, the runtime increases as the number of restraints in the network grows, though extensive caching in the calculations minimizes this effect to an extent. With increasing protein size, more iterations will be necessary to achieve a comparable level of convergence. Although a parallelization of a single run onto multiple cores is not possible in the current implementation, it is possible to perform several TYPHON experiments in parallel to obtain better statistics.

## DISCUSSION

In this paper we present TYPHON, a probabilistic approach to explore conformational fluctuations in proteins. TYPHON incorporates detailed probabilistic models of the conformational space of a protein’s main chain and its amino acid side chains (Boomsma et al., 2008; Harder et al., 2010) and an efficient local backbone resampling algorithm (Bottaro et al., 2012). During sampling by TYPHON, the conformational space is restricted by a set of restraints imposed on the structure. These restraints typically concern nonlocal interactions, such as hydrogen bonds, disulfide bridges, or interactions with metal ions. The protein structure on a local length scale, including main chain and side chains, is controlled by the probabilistic models.

In this study, we show that TYPHON is able to generate structural ensembles that closely resemble native ensembles described by experimental measures. This includes fluctuations as measured by S2 order parameters, as well as measured by RDC values. The RNase A study shows not only that TYPHON captures the functional dynamics in the correct regions but also that a principal component analysis of the results is feasible to identify large-scale motions. The analysis of the superoxide dismutase results shows that TYPHON can be used to model effects due to the gain or loss of a ligand, including partial unfolding.

Its computational efficiency makes TYPHON a promising tool for larger screening efforts, for example, of known mutations with clinical relevance. Another interesting application lies in generating suitable candidate structure for docking experiments, allowing for some degree of flexibility in the binding pocket (Henzler and Rarey, 2010). The high quality of the generated structures indicates that no irrelevant parts of the conformational space are explored. On the other hand, TYPHON thoroughly samples the relevant conformational space.

The results of the human p8MTCP1 protein experiments demonstrate another strength of our approach. With only a minimal set of restraints defined for the system, the effect of the probabilistic models becomes obvious. They control the local structure and maintain the overall secondary structure, while still allowing for significant conformational fluctuations. It should be noted that it is also possible to run TYPHON without explicitly defining the secondary structure, leading to significantly broader sampling.

In the current implementation, TYPHON keeps the constraint network fixed during the sampling. As a next step, it would be advantageous to allow more flexibility in the constraint network, such as the dissolution or formation of arbitrary hydrogen bonds as the sampling progresses. However, this will require the development of a suitable probabilistic model of nonlocal interactions in proteins and its seamless combination with the probabilistic models of local structure. Fortunately, important theoretical progress was recently made in this respect (Hamelryck et al., 2010). Another interesting addition would be to include information from experimental data (Olsson et al., 2011).

### Availability

TYPHON is available as part of the Phaistos package and can be obtained freely from SourceForge under the GNU public license.

<table>
<thead>
<tr>
<th>Residues in Regions</th>
<th>tCONCOORD</th>
<th>TYPHON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most favored (%)</td>
<td>69.8</td>
<td>88.1</td>
</tr>
<tr>
<td>Additionally allowed (%)</td>
<td>26.3</td>
<td>10.8</td>
</tr>
<tr>
<td>Generously allowed (%)</td>
<td>3.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Disallowed (%)</td>
<td>0.7</td>
<td>0.5</td>
</tr>
<tr>
<td>$\phi/\psi$ G factor</td>
<td>$-0.93$</td>
<td>$-0.24$</td>
</tr>
<tr>
<td>$\chi_1$ G factor</td>
<td>$-0.26$</td>
<td>0.10</td>
</tr>
<tr>
<td>Overall G factor</td>
<td>$-0.69$</td>
<td>$-0.13$</td>
</tr>
</tbody>
</table>

The table lists the results of a PROCHECK analysis of a set of TYPHON and tCONCOORD samples. Well-refined structures usually have 90% or more of all residues in the most favored regions. The G factor is a log odds score; higher numerical values denote higher quality. See also Documents S2 and S3.
http://sourceforge.net/projects/phaistos/). Currently the Phaistos package is limited to single chain proteins. However, support for multiple chains will be added in the next release.

EXPERIMENTAL PROCEDURES

Overview
The TYPHON network calculation starts from a full atom protein structure, including all of the hydrogen atoms. A restraint network is either loaded from an input file or created in accordance with the protocol described in the following section. In the course of the sampling, the dihedral angles in both the main chain and the side chains are modified under the control of TorusDBN (Boomsma et al., 2008) and BASILISK (Harder et al., 2010), respectively. An efficient local moves method makes subtle movements of the protein backbone possible (Bottaro et al., 2012) and also affects the bond angles in the backbone (see Protein Backbone Move).

Restraint Network Calculation
TYPHON currently supports three classes of restraints involving hydrogen bonds, disulfide bridges, and distance restraints between arbitrary atoms. In the absence of any user input, the program suggests a network using default parameters, which are described in the following paragraphs. This default network is mainly based on biologically relevant restraints, such as hydrogen bonds and disulfide bridges. To stabilize parts of the protein that are naturally stabilized by effects that are not modeled explicitly, TYPHON also connects residues that are far apart in the amino acid sequence but close in space. The user can edit the generated network by adding, removing, or modifying restraints between arbitrary atoms.

We evaluate all potential hydrogen bonds using the DSSP hydrogen bond energy (Kabsch and Sander, 1983). Following Kabsch and colleagues, we discard all candidates with a DSSP energy higher than −0.5 kcal/mol. If an atom has multiple potential hydrogen bonding partners, only the one with the lowest energy is retained. Following the general idea of the DSSP hydrogen bond energy, the hydrogen bond geometry is modeled using four distances. For backbone-backbone hydrogen bonds, these respective distances are explained in more detail in Figure S3. For hydrogen bonds involving side chains, the corresponding standard hydrogen bond acceptors and donors are used; asparagine, aspartate, glutamine, and glutamate can act as hydrogen bond acceptors; arginine, asparagine, glutamine, histidine, lysine, serine, threonine, tryptophan, and tyrosine can act as hydrogen bond donors.

Disulfide bridges are required to have a S→S distance of 3 Å or less. Similar to hydrogen bonds, the geometry of the disulfide bond is also modeled by four distances consisting of the S→Cj, Cj→Cj, Sj→Cj, and Cj→Cj distances.

The last class of restraints that are detected by default connects residues that are far apart in the amino acid sequence but close together in space. These restraints stabilize parts of the protein that are naturally stabilized by effects not accounted for explicitly in TYPHON, such as hydrophobic interactions. Residue pairs that are five or more residues apart in the sequence but within six Å (Cα–Cα distance) are modeled by a Gaussian probability distribution on the distance between the two Cα atoms. The distance in the input structure is used as mean μ. The variance σ2 is set proportional to the square of the distance:

\[
\sigma^2 = \frac{(\mu^2 - \mu)}{\delta}
\]

This value was chosen by trial-and-error and produces reasonable results. It allows for more flexibility with increasing distance.

The automatically detected restraints will not always yield the best results, especially when modeling large-scale movements. To keep the framework flexible and utilize the expert knowledge of the researcher, TYPHON allows modifying the restraints and adding distance restraints between arbitrary atom pairs in the structure. In that way, the researcher may additionally stabilize certain parts of the structure or allow more flexibility in other parts. It is also possible to remove automatically detected restraints, for example, when a certain hydrogen bond is known to be weak.

To further simplify this process, TYPHON can generate a PyMOL (Schrödinger, 2010) script that visualizes the restraints. This makes it possible to quickly detect regions that need manual, expert interaction. Figure 7 shows different restraint networks visualized using the generated PyMOL script.

Unstable Hydrogen Bonds
Hydrogen bonds that are in direct contact with solvent molecules are known to be significantly less stable than those that are well shielded. Fernandez and colleagues (Fernández and Berry, 2002; Fernández et al., 2002; Fernández and Scott, 2003; Fernández, 2010) proposed the concept of dehydrons, insufficiently shielded hydrogen bonds that are more likely to break. They showed that the number of the carbonaceous groups, CHn, in a shell around the hydrogen bond is a good estimate of water accessibility. tCONCOORD incorporates this convenient measure to judge the stability of a hydrogen bond (Seeliger et al., 2007). We extended their approach, which was only formulated for backbone hydrogen bonds, to apply to hydrogen bonds involving side chains as well. We therefore moved the centers of the two spheres composing the dehydration shell to the donor nitrogen and the acceptor carbon atoms (Fernández and Berry, 2002; Fernández et al., 2002; Fernández and Scott, 2003; Fernández, 2010). We recalibrated the measure using counts of carbonaceous groups derived from a set of high-resolution crystal structures previously used as training data for BASILISK (Harder et al., 2010). Following Fernandez and colleagues (Fernández and Berry, 2002; Fernández et al., 2002; Fernández and Scott, 2003; Fernández, 2010), we defined the threshold between weak and strong hydrogen bonds as the 4% percentile of the counts. This resulted in thresholds equal to 14, 9, and 7 for backbone-backbone, backbone-side-chain, and side-chain-side-chain hydrogen bonds, respectively. All weak hydrogen bonds are removed from the restraint network by default.

Protein Backbone Move
TYPHON sampling is usually started from the native state of a protein, that is, from a densely packed, compact structure. To capture the subtle movements and flexibilities in compact proteins, it is important to propose local updates of the backbone conformation. A local move only affects a limited part of the protein backbone—such as a stretch of five residues—whereas the rest of the protein remains unchanged.

Figure 7. Restraint Network
Depicted are three different calculated networks for ubiquitin (PDB 1ubi).
(A) A network that includes all hydrogen bond types (red: backbone hydrogen bonds; purple: backbone-side-chain hydrogen bonds; and yellow: side-chain-side-chain hydrogen bonds), as well as Cα contacts (green).
(B) A network that includes only the backbone hydrogen bonds.
(C) A network that only includes Cα contacts. The cutoff was 7 Å. The minimum sequence separation between the residues in the chain was two.
See also Figure S3.
In TYPHON, we use a recently developed type of local move called CRISP (Bottaro et al., 2012). Similar to other methods (Go and Scheraga, 1970; Dodd et al., 1993; Hoffmann and Knapp, 1996; Ulmschneider and Jorgensen, 2003), a local move consists of a concerted rotation of the bond and dihedral angles of the backbone atoms of neighboring residues. Each move involves four elementary steps:

1. Choose a random stretch in the protein chain.
2. Precord: Propose a set of bond and dihedral angle variations in the first \( N \) – 6 degrees of freedom.
3. Postrotation: Calculate the six remaining degrees of freedom such that the loop closes.
4. Calculate the bias introduced by performing such a nonrandom modification of the chain. The bias calculation is important when the method is used in a Markov chain Monte Carlo sampling scheme to ensure detailed balance.

This geometrical problem is tackled in an original manner. We derived an analytical solution for the postrotational step, thus avoiding the tedious numerical solution of a system of six equations for the six unknown degrees of freedom. The analytical solution is used to derive an efficient strategy to draw tentative updates of the chain. This scheme makes it possible to continuously control the angular variations of all degrees of freedom involved. The CRISP method thus improves on previous concerted-rotation methods in which, to satisfy all geometrical restraints, tentative updates of the chain are often radically different from the original structure or introduce a suboptimal local structure.

### Protein Side-Chain Move

To propose a new side-chain conformation, we use our previously developed probabilistic model of side-chain conformational space, BASILISK (Harder et al., 2010). BASILISK is a dynamic Bayesian network that makes it possible to sample side-chain conformations for all relevant amino acids in continuous space. By default, TYPHON resamples a single, randomly picked residue at a time, proposing an entirely new set of \( X \) angles for the side chain. Both the bond length and the bond angles remain unchanged. To have a roughly equal number of accepted changes affecting side chains and backbone, TYPHON on average resamples five side chains for every backbone move because a local move affects five backbone residues.

### Sampling Strategy and Scoring Functions

For sampling, we use a classic Markov chain Monte Carlo (MCMC) approach. According to the Metropolis–Hastings (Metropolis et al., 1953; Bishop, 2006) sampling scheme, a proposed \( X' \) structure is accepted with the following likelihood:

\[
P_{\text{acc}}(X \rightarrow X') = \min \left( 1, \frac{P(X'|X)Q(X \rightarrow X')}{P(X|X')Q(X' \rightarrow X)} \right),
\]

where \( P_{\text{acc}} \) is the probability of accepting to move from structure \( X \) to structure \( X' \); \( P(X) \) and \( P(X') \) are the probabilities of \( X \) and \( X' \), respectively; \( Q(X \rightarrow X') \) and \( Q(X' \rightarrow X) \) are the probabilities of proposing to move from \( X \) to \( X' \) and from \( X' \) to \( X \), respectively. \( P(X) \) is defined as

\[
P(X) = P_a(R)|P_t(T)|P_b(B)\Delta(X),
\]

where \( A \) is the amino acid sequence; \( P_a(R) \) is the probability density of the restraint network \( R \), consisting of the product of the probability densities of the individual Gaussian restraints; \( P_t(T|A) \) is the density of the backbone angles \( T \) according to TorsusDBN; \( P_b(B|A) \) is the probability density of the side-chain angles \( B \) according to BASILISK; and \( \Delta(X) \) is a clash term that is either one or zero. This simple clash function is introduced to avoid close contacts between atoms. We reject every structure with one or more atom pairs below a specific distance cutoff. The exact cutoff distance depends on the atoms involved: 1.5 Å for a hydrogen atom and any other atom; 1.8 Å for \( S_i \) atoms, to allow disulfide bridges; and 2.3 Å for any other atom pair.

The proposal distributions consist of resampling of side-chain conformations using BASILISK (Harder et al., 2010) or local moves using CRISP (Bottaro et al., 2012). To facilitate smooth local perturbations of the backbone chain, CRISP allows for small variations of the backbone bond angles. Each angle is modeled by an atom specific Gaussian distribution with parameters chosen in accordance with the bond-angle term of the OPLS-AA force field (Jorgensen et al., 1996; Kaminski et al., 2001).

### SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures, two tables, one movie, and two documents and can be found with this article online at doi:10.1016/j.str.2012.03.020.

### ACKNOWLEDGMENTS

The authors thank Francesco Carbone for help with the p8MTCP1 study. We thank Kaare Teilum (University of Copenhagen, Copenhagen, Denmark), Kresten Lindoff-Larsen (University of Copenhagen), Thomas Poulsen (Novozymes, Bagsvaerd, Denmark), and Leonardo De Maria (Novozymes) for valuable comments and suggestions. T.H. and M.B. are funded by the Danish Council for Strategic Research (NABIT, 2106-06-0009). W.B. and S.O. are funded by the Danish Council for Independent Research (FNU, 272-08-0315, and FTP, 274-09-0184, respectively). S.B. acknowledges funding from RADIOMETER, DTU Elektro.

Received: June 27, 2011
Revised: March 8, 2012
Accepted: March 12, 2012
Published online: May 10, 2012

### REFERENCES


Schrödinger, L. (2010). The PyMOL molecular graphics system (version 1 1.3r1), Schrödinger, LLC.


4.3 Appendix C: Phaistos

This paper presents the framework which has been in development in Thomas Hamelrycks group to conduct Markov chain Monte Carlo simulations of proteins. The special features of PHAISTOS include the native incorporation of the probabilistic models of local protein structure TorusDBN and Basilisk, along with efficient sampling through generalized ensembles (Muninn). This is a co-author article. The paper is accepted for publication in Journal of Computational Chemistry. I was involved in extensive testing of the framework and have committed some core functionality.
Abstract

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 the OPLS-AA/L force-field with the GB/SA 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.sourceforge.net. The software is implemented in C++, and has been tested on Linux and OSX platforms.

Keywords: Markov chain Monte Carlo simulation, protein structure, probabilistic models, local moves, conformational sampling

*to whom correspondence should be addressed
αDepartment of Biology, University of Copenhagen, 2200 Copenhagen, Denmark
βDepartment of Astronomy & Theoretical Physics, University of Lund, SE-223 62 Lund, Sweden
γCenter for Bioinformatics, University of Hamburg, 20146 Hamburg, Germany
δScuola Internazionale Superiore di Studi Avanzati, 34136 Trieste, Italy
εDepartment of Biomedical Engineering, DTU Elektro, DTU, 2800 Kgs. Lyngby, Denmark
ζNiels Bohr Institute, University of Copenhagen, 2100 Copenhagen, Denmark
ηUni Computing, Uni Research AS, Norway
θDepartment of Chemistry, University of Copenhagen, 2100 Copenhagen, Denmark
ιBILS, Science for Life Laboratory, Box 1031, 171 21 Solna, Sweden
We present a new software package for simulation and inference of protein structure. The PHAISTOS framework contains a range of novel sampling techniques and probabilistic models, constituting a versatile toolkit for efficient simulations of protein structure. The package provides tools for a variety of tasks, including reversible folding simulations and probabilistic inference of protein structure from experimental data. The source code is released under an open source license and full documentation is available online.
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. 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. In addition, the statistical nature of MCMC methods make them particularly well-suited for Bayesian inference of protein structure from experimental data.

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 paper, we present a software framework designed with this goal in mind. The PHAISTOS framework contains implementations of recently developed tools that increase the 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 multi-canonical ensemble).

The framework contains many of the established moves from the literature, including various pivot moves (\texttt{move-pivot-uniform}, \texttt{move-pivot-local}), the crankshaft/backrub local move (\texttt{move-crankshaft})\textsuperscript{6,7}, the CRA local move (\texttt{move-cra})\textsuperscript{8}, and the semi-local biased Gaussian step (BGS) (\texttt{move-semilocal})\textsuperscript{9}. Side-chain conformational sampling can be done either from Gaussian distributions given by rotamer libraries (\texttt{move-sidechain-rotamer})\textsuperscript{10} or through Gaussians centered around the current side-chain conformation (\texttt{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\textsuperscript{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 C\textalpha{} trace of a protein\textsuperscript{13}, COMPAS models a reduced single-particle representation of amino acid side-chains, while TORUSDBN and BASILISK, respectively, model backbone and side-chain structure in atomic detail\textsuperscript{14,15}. All models can be applied both as proposal distributions in the form of Monte Carlo moves (\texttt{move-backbone-dbn}, \texttt{move-sidechain-basilisk}, \texttt{move-sidechain-compas}) and as probabilistic components of an energy function (\texttt{energy-backbone-dbn}, \texttt{energy-basilisk}, \texttt{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 Energy section below.

### 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 (pre-rotation), while other are modified deterministically to bring the chain back to a closed state (post-rotation). From the work of Gō and Scheraga\(^16\), it is known that in general, six degrees of freedom are required for the post-rotation step. CRISP distinguishes itself from previous methods by merging these two steps, modifying the stochastic pre-rotation step so that it takes the resulting post-rotation 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 \(n-6\) pre-rotation degrees of freedom \(\bar{\chi}_\text{pre}\)

\[
P(\delta \bar{\chi}_\text{pre}) \propto \exp \left( -\frac{1}{2} \delta \bar{\chi}_\text{pre}^T C_{n-6} + S^T C_6 S \delta \bar{\chi}_\text{pre} \right).
\]

Here, \(C\) is an inverse diagonal covariance matrix specifying the desired fluctuations for the individual angular degrees of freedom, \(S\) is a linear transformation mapping the pre-rotational degrees of freedom to the corresponding post-rotational values, and \(\lambda\) 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 (sidechain-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 semi-local 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 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 forcefields were selected to provide support for a broad range of simulation tasks. The efficiency of the PROFASI forcefield makes it possible readily to conduct reversible folding simulations of peptides and small proteins\(^17\). The OPLS forcefield 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 non-bonded force-field terms in both forcefields have been implemented using the chaintree data structure\textsuperscript{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\textsuperscript{17}

\[ E = E_{ev} + E_{hb} + E_{sc} + E_{loc} \]  

(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 neighbouring peptide units along the chain, with a correction term for improved consistency with the Ramachandran plot, and a side-chain torsion potential. Since 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\AA. While 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\textsuperscript{17}, while still being fast enough for many-body aggregation simulations\textsuperscript{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, while the torsion term has the form\textsuperscript{18}

\[ E_{torsion} = \sum_i \sum_{j=1}^{3} w_j (1 + (-1)^{j+1} \cos(j \theta_i)) \]  

(3)

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

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

(4)

where \( r_{ij} \) is the distance between atoms \( i \) and \( j \), \( q_i \) and \( q_j \) are the corresponding partial charges, \( \varepsilon_0 \) is the vacuum permittivity and \( \sigma_{ij} \) and \( \epsilon_{ij} \) are calculated using the combination...
rules $\sigma_{ij} = \sqrt{\sigma_i \sigma_j}$ and $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$. 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 neighbours $(1,2)$ 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. 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 should be combined with a suitable solvent model to reproduce physiological conditions. In order to model the effect of the solvent on hydrophobic interactions and electrostatics, we use the OPLS forcefield in combination with the Generalized Born Surface Area (GB/SA) implicit solvent model.

Many implicit solvent models express the solvation free energy $G_{\text{solv}}$ as a sum of non-polar and electrostatic contributions

$$G_{\text{solv}} = G_{\text{npol}} + G_{\text{pol}}$$

(5)

Here, $G_{\text{npol}}$ is the free energy of solvating the molecule with all the partial charges set to zero, and $G_{\text{pol}}$ is the reversible work required to increase the charges from zero to their full values. In GB/SA, the non-polar contribution $G_{\text{npol}}$ 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

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

(6)

where $\varepsilon$ is the dielectric constant of the solvent and $q_i$ is the partial charge of atom $i$. $f_{\text{GB}}(r_{ij}) = \sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp \left( -r_{ij}^2/4\alpha_i \alpha_j \right)}$ is a function of the distance $r_{ij}$ and of the so-called Born radii $\alpha$, 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.

**Incorporating Probabilistic Models in the Energy Function**

When using moves that are based on probabilistic models such as TORUSDBN or 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. In order 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

$$LL(\bar{X}, \bar{A}) = \ln \sum_{\bar{H}} P(\bar{X}, \bar{A}, \bar{H})$$

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

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. 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 which can be done efficiently using dynamic programming.

**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). 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, or more sophisticated evaluations of the environment of each residue.

**Monte Carlo**

Although PHAISTOS can be used for Monte Carlo minimization, the primary focus of the framework is Markov chain Monte Carlo 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 $x'$

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

\[
\frac{P_\alpha(x \to x')}{P_\alpha(x' \to x)} = \frac{\pi(x')P_s(x' \to x)}{\pi(x)P_s(x \to x')} \tag{9}
\]

Most of the moves are symmetric, in the sense that \( P_s(x' \to x)/P_s(x \to x') = 1 \). However, for moves such as the local and semi-local 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

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

This is the default simulation method used in PHAISTOS ([monte-carlo-metropolis-hastings](https://muninn.sourceforge.net/)). 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 Metropolis-Hastings 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

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

for a given weight function \( w(x) \). The typical choice is the multi-canonical 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/](http://muninn.sourceforge.net/)). It employs the generalized multi-histogram 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\textsuperscript{20}.

Finally, through a modular build-system, developers can readily write their own modules utilizing the library. Modules are separate code entities that are auto-detected 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\textsuperscript{32}, demonstrating several of the features described above.

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

```
$ ./phaistos --aa-file beta3s.aa \
    --energy profasi-cached backbone-dbn[weight:-1] \
    --move backbone-dbn sidechain-uniform semilocal-dbn-eh \
    --threads 8 --temperature 283 \
```

```
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 8 parallel threads and set the temperature to 283K. MUNINN is chosen to be the Monte Carlo engine, using a $\beta$ (inverse temperature) range of $[0.6; 1.1]$. These $\beta$ factors are unit-less, specified relative to the inverse temperature, and thus correspond to a temperature range of $[(1.1 \cdot 1/283K)^{-1}; (0.6 \cdot 1/283K)^{-1}] = [257K; 472K]$. 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 ($<[E]_{257K}; [E]_{472K}>$).

We ran PHAISTOS with the settings above on an eight-core 3.4GHz Intel Xeon processor for one week. Figure 3 gives an overview of the results. The free energy plot in Figure 3(a) shows the distribution of energy versus RMSD extracted directly from the samples dumped during the simulation. Since the samples were generated using a generalized ensemble 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 3(b).

To find representative structures in the ensemble, we use the PLEIADES clustering module\textsuperscript{33}, 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 283K, the protein is marginally stable in the PROFASI force-field, with native-like populations at 2Å and 4Å 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%\textsuperscript{32}. This result is compatible with a previously published simulation of the same protein using an unbiased simulation technique\textsuperscript{17}. 

---

**--monte-carlo muninn [min-beta:0.6,max-beta:1.1] \**
**--observable backbone-dbn rmsd [reference-pdb-file:beta3s.pdb] \**
**--observable xtc-trajectory**
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\textsuperscript{34}. 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 or 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 contains low resolution information on the overall shape of a protein, which can be useful for determining the relative domain positions and orientations in multi-domain 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\textsuperscript{35,36}.

NMR experimental data can provide high resolution structural information which 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\textsuperscript{37,38}. 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\textsuperscript{39}.

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\textsuperscript{40} 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\textsuperscript{41}, which provides dramatic computational speedups\textsuperscript{33}.

Native ensembles

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

The TYPHON module\textsuperscript{42} rapidly explores near-native ensembles by using the CRISP move in combination with a user-defined set of non-local restraints. Local structure is under the control of probabilistic models of the backbone (TORUSDBN) and side chains (BASILISK), while non-local 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 paper.

Of the available Monte Carlo software packages, the ROSETTA package\textsuperscript{11} is perhaps the most widely used, and has an impressive track record for protein structure prediction and design\textsuperscript{43}. 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\textsuperscript{44}, in the sense that both implement the BGS move\textsuperscript{9} and the PROFASI energy function\textsuperscript{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 forcefield 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\textsuperscript{1} and the Monte Carlo package in CHARMM\textsuperscript{2}, which both provide functionality for conducting Markov chain Monte Carlo 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 molecular dynamics 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.

ACKNOWLEDGMENTS

This work was supported by the Danish Council for Independent Research [FNU272-08-0315 to W.B., FTP274-06-0380 to K.S, FTP09-066546 to S.O., J.V., FTP274-08-0124 to K.E.J.], the Danish Council for Strategic Research [NABIIT2106-06-0009 to J.F., Ti.H., C.A., M.B.], the Novo Nordisk STAR Program [A.S.C.], a Hallas-Møller stipend from the Novo Nordisk Foundation [K.L.L.] and Radiometer (DTU) [S.B.].
References


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, while each X node is a bivariate distribution over ($\phi$, $\psi$) 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 ($\phi$, $\psi$) 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 neighbouring $H$ 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 ($\phi$, $\psi$) 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.

Figure 2: A UML-diagram of the major classes in the PHAISTOS library (black diamond: composition, white diamond: aggregation, arrow: inheritance). A MonteCarlo simulation object contains a MoveCollection object, which consists of a selection of moves, and an Energy object, which is comprised of 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.

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 multi-canonical (flat histogram) ensemble. b) Free energy plot as a function of energy and RMSD, reweighted to the canonical ensemble at 283K. c-f) Representative cluster medoids found with reweighted clustering using the PLEIADES module, compared to the native structure (shown in black). Figures created using Pymol. g) RMSD vs time of one of the eight threads in the simulation.
Figure 1
4.3. APPENDIX C: PHAISTOS

Figure 2
Figure 3