On the application of Bayesian statistics to protein structure calculation from nuclear magnetic resonance data

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Martin Mechelke

aus Homberg (Efze)

Tübingen
2014

Tag der mündlichen Qualifikation:	14.10.2014
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Dr. Michael Habeck

Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Tübingen, 17. Oktober 2014

Martin Mechelke

Abstract

In the present work, we use concepts of Bayesian statistics to infer the three-dimensional structures of proteins from experimental data. We thus build upon the method of inferential structure determination (ISD) as introduced by Rieping et al. (2005). In line with their probabilistic approach, we factor the probability of a three-dimensional protein structure given the experimental data, into a prior distribution that captures the protein-likeness of a structure and the likelihood that describes how likely the experimental data were generated from a given three-dimensional structure. In this Bayesian framework, we attempt to develop structure calculation from NMR experiments into a highly accurate, objective and parameter-free process.

We start by focusing on integrating new types of data, as ISD currently does not entail a mechanism to incorporate chemical shifts in the calculation process. To alleviate this shortcoming, we propose a hidden Markov Model that captures the relationship between protein structures and chemical shifts. Based on our probabilistic model, we are able to predict the secondary structure and dihedral angles of a protein from chemical shifts.

Another means to high quality structures involves improving the potential functions that form the core of ISD's prior distributions. Although potential functions are designed to approximate physical forces, there are still parameters, such as force constants and temperatures, that are set on an ad hoc basis and can bias the structure calculation. As an alternative, we propose an algorithm based on Bayesian model comparison to determine these parameters from the data. Further, we demonstrate that optimal data-dependent parameters lead to improved accuracy and quality of the final structure, especially with sparse and noisy data. These findings dismiss the notion of a single universal parameter and advocate the estimation of free parameters based on experimental data instead.

Third, we focus on the estimation of new potential functions to include even more prior information in the structure calculation process. Currently, only a few methods allow the estimation of potential functions from a database of known structures. Our method provides a sound mathematical solution of this problem, which is also known as the inverse problem of statistical mechanics.We demonstrate the effectiveness of our approach on the examples of simple fluids and a coarse-grained protein model.

Zusammenfassung

Im Rahmen dieser Arbeit stellen wir neue Ansätze, basierend auf der Bayes'schen Statistik, zur Interpretation von experimentellen Daten in der NMR-Spektroskopie vor. Dabei bauen wir auf den Ergebnissen von Rieping et al. (2005) auf, die das Prinzip der inferentiellen Strukturbestimmung (ISD) eingeführt haben. Ihr probabilistischer Ansatz beruht auf der Faktorisierung der A-posteriori-Verteilung in die A-priori-Verteilung, welche die Proteinähnlichkeit einer möglichen Struktur bewertet, und die Likelihood-Funktion, welche die Übereinstimmung mit den experimentellen Daten beschreibt. Ziel dieser Arbeit ist es, die Qualität, aber auch die Vergleichbarkeit der Strukturberechnung in der NMR-Spektroskopie zu verbessern.

Zuerst beschäftigen wir uns mit der Integration neuer experimenteller Datentypen in die Strukturrechnung. Dazu schlagen wir ein Hidden-Markov-Modell vor, das beruhend auf der chemischen Verschiebung die Dihedralwinkel und Sekundärstruktur vorhersagt.

Eine Alternative zur Integration zusätzlicher experimenteller Information ist die Verbesserung der A-priori-Verteilung. In ISD beruht die A-priori-Verteilung auf einer Potentialfunktion, welche die frei Energie approximiert. Dennoch gibt es freie Parameter in Potentialfunktionen, wie die Temperatur oder doe Kraftkonstante, die festgelegt werden müssen. Wir benutzen Bayes'sche Hypothesentests, um die freien Parameter objektiv und beruhend auf den experimentellen Daten zu bestimmen. Die Anwendung der Bayes'schen Hypothesentests ermöglicht es uns, verschiedene Potentialfunktionen zu kombinieren, um aus verrauschten und unvollständigen Daten noch exakte Strukturen zu bestimmen. Weiterhin zeigen unsere Studien, dass für statistische Potentiale keine allgemeingültige Kraftkonstante existiert und diese anhand der experimentellen Daten bestimmt werden sollte.

Im dritten Teil dieser Arbeit führen wir eine Methode ein, um neue Kraftfelder aus Strukturdatenbanken zu erlernen und damit die A-priori-Verteilung noch weiter zu verbessern. Dieses nichtlineare Problem ist auch als inverses Problems der statistischen Mechanik bekannt, das wir durch eine Generalisierung des Konzepts der *Configurational Temperature* lösen. Wir benutzen unsere Methode, um die Potentialfunktionen von vereinfachten Moleküldynamik Kraftfeldern zu rekonstruieren.

Acknowledgements

First and foremost, I would like to thank my supervisor, Michael Habeck, for the many thoughtful and inspiring conversations, and for giving me the freedom to pursue many interesting projects. My gratitude goes out to Andrei Lupas for his support and for allowing me to work in his lab. I also want to thank Oliver Kohlbacher for supervising this thesis at the Eberhard Karls University, Tübingen and for the helpful discussions. Thanks go to my colleagues Vikram Alva, Klaus Kopec, Simeon Carstens, Ivan Kalev and the rest of department 1 for the many fascinating conversations, coffee breaks and table soccer matches. I am grateful to Karin, Michael, Silvester and Vikram for their comments on the manuscript.

Special thanks go to my parents who always encouraged me and had no doubt that I would reach my goals. I thank all my friends for reminding me that there is a real world outside of science. Finally, my greatest gratitude goes to Ines for her constant compassion and support.

Contents

1

Introduction

The aim of structural biology is to analyse the structure and function of biological macromolecules. This discipline looks back on more than 50 years of history. Throughout these years, X-ray crystallography has remained the method of choice for the elucidation of protein structures. Over time, nuclear magnetic resonance spectroscopy (NMR) and cryo-electron microscopy (cryo-EM) were added to the repertoire of experiments that provide structural insights. What has changed are the means to analyse and interpret the experimental data. While Kendrew and Perutz, the first to solve protein structures, used manually adjusted wire-frame models to explain the diffraction pattern, most of the analysis and interpretation today are performed by computer algorithms that require little to no human interaction. The solution of a new structure has become a routine task and we are able to elucidate many biological reactions in atomic detail. As the number of known structures keeps growing at an increasing pace, more and more questions on the molecular principles of many biochemical pathways can be answered.

One of the main contributor and, at the same time, beneficiary of structural biology is the pharmaceutical industry. Knowledge of the atomic detail of a receptor or enzyme is of great help in the search for specific and strong binders. A prominent example is the HIV protease, an enzyme that is represented by several hundred structures in public databases, most of which are complexes with potential drugs and inhibitors. There are probably hundreds more in the databases of pharmaceutical companies. Overall, this makes the virus protein one of the best-studied structures. This immense structural knowledge allowed researchers to develop highly effective inhibitors that are able to significantly slow the progression of the

virus. Despite such success stories, however, structural biology offers many open challenges that can be addressed computationally.

One of these challenges is the application of NMR spectroscopy to large biological macromolecules. NMR spectroscopy is a powerful technique that provides insights into the functionally important motions and transient interactions of protein structures. An important limitation of this technique is molecule size; typically NMR is restricted to small proteins with less then 50 kDa. But, eukaryotic cells depend on the function of many large complexes, some of which are composed of more than a hundred different proteins. Special labelling techniques are used to gather structural and functional information on proteins of sizes up to 1 MDa (Sprangers et al., 2007). The downside of these labelling techniques is that they result in a sparse set of distance restraints that can prove difficult to solve by conventional methods.

New approaches are required to derive meaningful ensembles from sparse data. One solution are hybrid approaches that integrate different structural information at different levels of resolution. An alternative are methods that complement sparse structural information with precise potential functions. In this work, we pursue both routes using Bayesian statistics.

Predicting structural information from chemical shifts

The integration of chemical shift data provides additional information that can guide structure calculation. Chemical shifts have become increasingly important in structure calculation, as the chemical shift of an atom depends on important structural factors, like backbone conformation, secondary structure and the position of aromatic rings (Wishart and Case, 2001; Wüthrich, 1986). Recently, progress was made by Cavalli et al. (2007); Shen et al. (2008); Thompson et al. (2012), who use this information to determine the structure of proteins accurately.

Encouraged by their success, we focus on extracting structural data from chemical shifts. While the problem has received a lot of previous attention (Mielke and Krishnan, 2009; Cornilescu et al., 1999; Shen et al., 2009), almost all previous algorithms are ad hoc. Too often, scientists only look at the output behaviour of a complex algorithm, treating it like a ”black box”, but what is inside the box is also very important. To alleviate this issue, we present a principled, clean, and transparent algorithm based on hidden Markov models (HMMs) for solving this extensively

studied problem. The predicted secondary structure can be incorporated as distance restraints into the structure calculation process.

Our second approach is concerned with the prediction of dihedral angles from chemical shifts. The predicted dihedral angles can also serve as restraints in structure calculation. We propose an HMM, based on a discrete representation of protein structure by Boomsma et al. (2008), to infer the dihedral angles from chemical shifts.

Objective priors for structure calculation

The Bayesian methodology introduced by Rieping et al. (2005) provides an elegant and powerful approach to structure calculation. Instead of relying on energy minimization, Rieping et al. generate samples from a probability distribution, which is the product of the prior distribution that captures the protein-likeness of a structure, and the likelihood, which captures the goodness-of-fit to the experimental data. Within this framework it is possible to elegantly determine additional model-parameter, that previously were choose empirically (Habeck et al., 2006). However, there are still parameters that that cannot be estimated from the data, yet need to be adjusted like the temperature of the potential function. Bayesian inference stipulates determining such parameters based on the model evidence. This is challenging because the model evidence cannot be calculated analytically, not even for small proteins. So far, all methods to approximate the evidence rely on assumptions regarding the functional form of either the likelihood or the prior (MacKay, 1992; Li, 2009; Pryce and Bruce, 1995; Zhou et al., 1997), which makes them unsuitable in the context of structure calculation. We introduce a replica-exchange Monte Carlo (REMC) scheme that allows us to estimate the model evidence through use of multiple histogram reweighting for a large number of problems in Bayesian data analysis.

Combination of prior information

In the case of sparse and noisy data, we need to rely on the potential function to guide structure calculation towards the native conformation and resolve the inherent ambiguity of the experimental data. There have been two different approaches

to the construction of potentials: physics-based potentials, that are inspired by physical laws, and knowledge-based potentials, that are derived from databases. A fusion of data-driven and physics-based approaches, which are often complementary, should increase the accuracy of potential functions (Pande, 2011). We propose a method based on our earlier results on Bayesian model comparison, to combine different potentials in the context of structure determination. The resulting potential functions, optimized for a particular protein and experimental data set, lead to more accurate structures. Our results argue against the existence of a transferable combination of physics-based potentials and knowledge-based potentials.

Estimating new potential functions

The apparent incompatibility of physics-based and knowledge-based potentials motivated us to investigate new techniques to estimate potential functions, whereby physical interactions are used as prior knowledge but the model is then estimated from a database of known structures. The simplest strategy is to collect database statistics for certain geometric descriptors, such as distances and dihedral angles, and compute a statistical potential by inversion of the histogram (Sippl, 1995). But the resulting energy functions are, at best, potentials of mean force and generally differ from the potential energy function as they do not take the multiplicity into account and tend to neglect correlations (Ben-Naim, 1997; Shortle, 2003). Rather than compute statistical potentials, we aim to extract force field parameters directly from configurations that are drawn from the canonical ensemble, which amounts to solving the inverse problem of statistical mechanics. We propose an extension of the configurational temperature that allows us to infer a parameterized approximation of the potential function. It turns out that our approach is a generalization of Score Matching Hyvärinen (2005) and the generalized Yvon-Born-Green (gYBG) equations (Mullinax and Noid, 2009, 2010).

1.1 Structure of the Thesis

This thesis is divided into seven chapters. Following this introduction, the biological and mathematical background are introduced in Chapter 2. In Chapter 3, we intro-

duce a probabilistic model to calculate secondary structure elements and dihedral angles from chemical shifts with the aim of using this information for structure calculation. In order to incorporate the new potential functions into the structure calculation process, we need to estimate the influence of the potential function in the light of the experimental data. A general answer to this question is provided in Chapter 4, where we introduce a method to weight the prior information in Bayesian data analysis. Subsequently, in Chapter 5, we use the results from the previous chapter to find an optimal combination of different potentials for structure calculation. In Chapter 6, an extension to the configurational temperature is presented, which allows us derive new force fields from ensembles of structures. Chapter 7 provides a general conclusion to the work presented here.

2

Background

2.1 Protein structure

Proteins are fundamental to all living organisms. They are simple linear polymers composed of 20 different α amino acids. Each gene of an organism encodes the amino acid sequence of at least one protein. A gene is first transcribed into RNA and then translated at the ribosome to produce a protein, where it folds itself into a three-dimensional structure. The folding process is driven by the physiochemical properties of the individual amino acids and thus the structure of a protein is uniquely defined by its sequence. Over billions of years, driven by selection pressure, the sequence of amino acids evolved, to adopt the distinct three-dimensional conformation that we observe today and to carry out a vital function in the organism. These functions range from digesting nutrients and providing scaffolding for the cell to transferring signals and killing pathogens.

Each amino acid has the same basic composition: a central carbon atom, called C_α, to which a hydrogen atom, a carboxyl group ($COOH$) and an amino group (NH_2) are attached. They differ only in the fourth substituent of the central carbon. The 20 different substituents that make up the naturally occurring amino acids are diverse and range from single hydrogen atoms to complex aromatic ring systems. During synthesis of the protein at the ribosome, a peptide bond between the carboxyl group of one amino acid and the amino group of the next one is formed. This reaction, which is called condensation, covalently links the two amino acids and thereby releases a water molecule. The reaction creates electron delocaliza-

tion along the newly formed peptide bond, rendering it planar; thus the amino acid occurs only as cis or trans isomer. As a result, the angles of rotation around the $N - C_\alpha$ and $C_\alpha - C'$ bond, called the ϕ and ψ angles, are the only degrees of freedom of the protein backbone. As such, the conformation of the backbone can be completely described by all ϕ and ψ angles. Typically, the ϕ and ψ angles are visualized by the Ramachandran plot (Ramachandran et al., 1963).

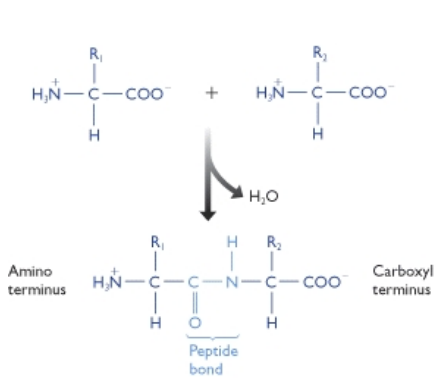
The structure of proteins is organized into primary, secondary, tertiary and quaternary structure. Each level of this hierarchy adds more information. The lowest level, the primary structure, refers to the sequence of amino acids along the polypeptide chain. The primary structure includes no three-dimensional information, but allows us to deduce the evolutionary relation between proteins. Two proteins of common decent are called homologous to each other and will often show a high similarity in their sequence and structure.

The concatenation of local repetitive elements along the protein chain stabilized by backbone hydrogen bonds is called the secondary structure. The most common secondary structure elements are the α helix and the β strand. The protein backbone of an α helix forms a right-handed spiral, where the CO group of each residue is hydrogen-bonded to the backbone NH group of the fourth residue along the chain. Thus, all residues of the α helix, except the first and the last, are connected by two hydrogen bonds. One turn of an α helix comprises 3.6 residues and is about 5.4 Å in height. An average helix in globular proteins comprises about three turns or 10 residues. The α helix is associated with ϕ, ψ angles pairs of around $-50° - -60°$ each.

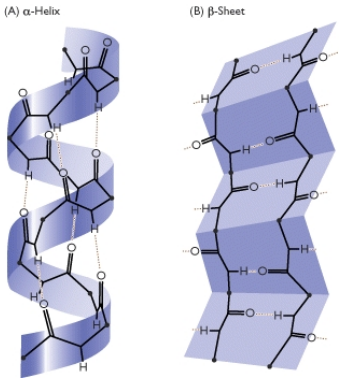
The polypeptide chain in a β strand adopts an extended conformation, which is stabilized by though-space backbone hydrogen bonds with another β strand. Two or more parallel or anti-parallel aligned β strands form a β sheet. Figure 2.1b shows an anti-parallel β sheet. In terms of geometry, parallel and anti-parallel β sheets have a characteristic hydrogen-bond pattern. The ϕ,ψ angles for a β strand are less well-defined than those of α helices and the angles potentially occupy the upper left part of the Ramachandran plot.

Most proteins are made up of either β strands or α helices joined loosely by structured loop regions of various lengths and shapes. Loop regions, in general, do not form hydrogen bonds and comprise charged and polar residues, as they are often exposed to the solvent. Besides these common secondary elements, there exist rare variants like the $3 - 10$ helix, the π-helix and collagen.

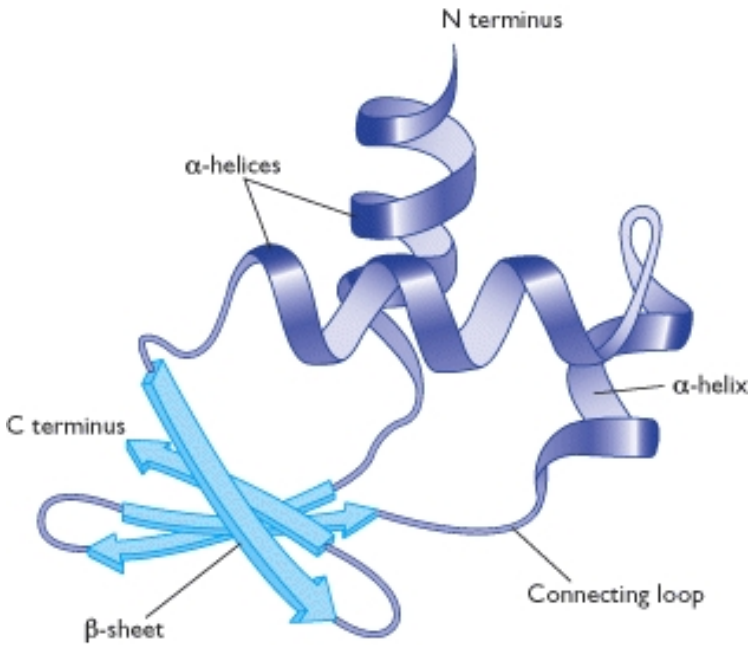
The tertiary structure refers to the atomic coordinates of the folded structure of



(a) Polypeptide synthesis



(b) Secondary structure



(c) Tertiary structure

Figure 2.1. Protein structure synthesis and organization. Panel 2.1a shows the condensation reaction that forms the peptide bond. The upper right Panel 2.1b shows an α helix and the arrangement and hydrogen bond pattern of an anti-parallel β sheet. The lower Panel 2.1c shows a cartoon representation of the tertiary structure and highlights the secondary structure elements. Figure taken from Brown (2002)

a protein. Tertiary structures are often composed of one or multiple autonomously folding substructures, which are called domains. Each domain fulfils a specific function within the tertiary structure. Despite the wide range of known domains, it is possible to classify them according to taxonomic levels (e.g. subfamily, family, superfamily).

A large number of proteins are unable to function on their own and form larger assemblies stabilised by non-bonded interactions. The quaternary structure refers to this assembly. It is not uncommon that the assembly consists of identical subunits. The largest multi-protein complexes like the nuclear pore complex can have a mass of up to ≈ 125 MDa.

Upon synthesis, proteins fold themselves into the final, functional, three-dimensional conformation. Hydrophobic side chains are buried in the core of the protein while the polar/charged side chains are left accessible to the solvent. Secondary structure elements are formed through hydrogen bonds and are further stabilised by disulphide bridges, charged interactions and hydrophobic interactions. These stabilising interactions are opposed by entropic contributions. In the end, the sum of stabilising effects outweighs the destabilising ones and proteins adopt their three-dimensional structures. The folding process itself is viewed as the path of the folding protein chain through an energy landscape, which funnels the forming protein towards the native state. Within this energy landscape, multiple folding pathways and metastable intermediates exist on the route to the native state.

2.2 Nuclear magnetic resonance

Nuclear magnetic resonance (NMR) can be used to determine the three-dimensional structure of proteins Wüthrich (1986). In addition, NMR can be used to investigate time-dependent biochemical phenomena like reaction kinetics and molecular dynamics (Cavanagh et al., 1996). The effects underlying NMR spectroscopy depend on the nuclear spin and were first reported by Bloch et al. (1946) and Purcell et al. (1946). NMR soon developed into a standard technique for the research on small molecules, and Bloch and Purcell were awarded the nobel price in physics in 1952 *"for their discovery of new methods for nuclear magnetic precision measurements and discoveries in connection therewith"*.

The correlations between protein structures and NMR spectra were realized quickly. Yet *de novo* structure calculations seemed out of reach until Ernst et al.

(1990) introduced the use of the Fourier transformation and pulsed frequency radiation. This advancement was awarded the Nobel Prize in Chemistry in 1991. These discoveries paved the way for multidimensional NMR spectroscopy, which led to the solution of the first protein structure by NMR spectroscopy by Williamson et al. (1985). This feat also led to Kurt Wüthrich being awarded the Nobel Prize in Chemistry in 2002.

Nuclear resonance

The theory of NMR revolves around a quantum mechanical property of the atomic nucleus called spin (Sakurai et al., 1995; Levitt, 2001). The spin $\boldsymbol{s}$ is the quantum mechanical analogue of the angular momentum and it is intrinsic to all elementary particles. It describes the magnetic field surrounding the nucleus. Due to its quantum mechanical nature, the spin is quantized (i.e. it may only be a discrete value) and can be characterized by the spin quantum number I, which is a multiple of $\frac{1}{2}$. The spin quantum number is connected to the norm of the spin $\|\boldsymbol{s}\|$ by

$$\|\boldsymbol{s}\| = \sqrt{I\,(I+1)}\hbar \tag{2.1}$$

, where $\hbar$ is the reduced Planck constant. It is defined as the Planck constant h divided by 2π. Since NMR relies on the existence of the magnetic field, only nuclei with a spin quantum number greater than or equal to $\frac{1}{2}$ are NMR active.

The magnetic moment of a nuclei $\boldsymbol{\mu}$ is given by

$$\boldsymbol{\mu} = \gamma \boldsymbol{s}\,, \tag{2.2}$$

where γ denotes gyromagnetic ratio specific to each nucleus. Given an arbitrary direction z, defined by an external magnetic field, the spin z-projection is given by

$$s_z = m_z\hbar \qquad \text{with} \quad m_z \in \{-I, -I+1, \ldots, I-1, I\}\,. \tag{2.3}$$

Thus, there are $2I+1$ different values of s_z. Nuclei with spins greater than $\frac{1}{2}$ are typically not considered in NMR as the greater number of spin states complicates the later analysis of spectra. The most important nuclei with spin $\frac{1}{2}$ is that of hydrogen $^1\mathrm{H}$, as it has high natural abundance. Many other common nuclei, such as $^{12}\mathrm{C}$

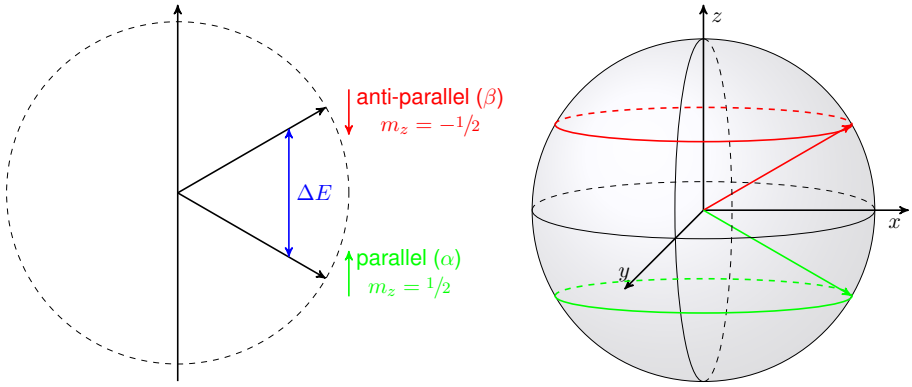


Figure 2.2. Energy level diagram reflecting the alignment of a 1/2 spin nuclei in an external magnetic field.

and $^{14}\mathrm{N}$, are NMR inactive. The isotopes of these elements, $^{13}\mathrm{C}$ and $^{15}\mathrm{N}$, are used to make them visible to NMR.

In the absence of an external magnetic field, the energy levels of each orientation m_z are equal. If a magnetic field $\boldsymbol{B_0}$ is applied to a probe, each orientation of the spin has a different energy. It depends on the direction of the spin relative to the external field, the strength of the external magnetic field, and the strength of the intrinsic magnetic moment γ of the nucleus. The energy is given by

$$E = -\mu_z \|\boldsymbol{B_0}\| = -m\gamma\hbar\|\boldsymbol{B_0}\|. \tag{2.4}$$

It should be noted that the energy is quantitized as well. Orientations along the magnetic field have lower energy and, following Boltzmann's law, are more highly populated. There exist two distinct energy levels for a nucleus with spin $\frac{1}{2}$ (see Figure 2.2), α with $m = 1/2$, and β with $m = -1/2$. In the α state, the spin is oriented parallel to the external field. The spin in the β state is oriented anti-parallel to the external field, thereby resulting in a higher energy. The energy difference ΔE between the two levels again depends on the strength of the magnetic field and is given by:

$$\Delta E = \gamma\hbar\|\boldsymbol{B_0}\|. \tag{2.5}$$

If an electromagnetic wave is applied to the probe and the energy of that wave is exactly ΔE, spins on level α are able to change to state β and a strong absorption occurs, which can be measured. The frequency at which the absorption occurs is called the resonance frequency of the nucleus. The resonance frequency of a pro-

ton is in the range of several hundred MHz for commonly used NMR spectrometer. As described by Equation 2.3, only the z component of the angular momentum is quantitized, while the vector is free to rotate as shown in Figure 2.2. The resulting motion, which is called precession, is similar to that of a gyroscope. The frequency of the motion is named after Joseph Larmor. The Larmor frequency f_{Larmor} is identical to the resonance frequency of the nucleus.

Chemical shift

How does the nuclear resonance give us any information about the probe? Why is it that not all hydrogen atoms in a protein have the same absorption frequency? To answer all these questions, we need to have a closer look at the magnetic field. The magnetic field is not homogenous; it differs locally. Every nucleus induces a weak magnetic field. Therefore, the magnetic field at a particular nucleus depends on its surrounding and only rarely will two nuclei experience the same field and have the same resonance frequency. The deviation of a resonance frequency from that of a reference nucleus is called the chemical shift. The chemical shift δ is usually measured in parts per million [ppm] and does not depend on the spectrometer used. It is defined as

$$\delta = \frac{\nu - \nu_0}{\nu_0} \tag{2.6}$$

where ν is the measured resonance frequency and ν_0 is the resonance frequency of a reference. For example, tetramethylsilane (TMS) is used as a reference for hydrogen atoms and $^{13}\mathrm{C}$. The exact computation of chemical shifts from first principles is only possible for small molecules. In the case of larger system, the complexity of the interactions is prohibitive. Nevertheless, it is possible to arrive at a good empirical approximation for proteins.

Fourier Transform NMR

We still need to answer the question of how NMR can be used to elucidate a protein structure. The goal of NMR is to obtain the intensity of the resonance as a function of the frequency. In the early experiments, a frequency sweep was used to collect a

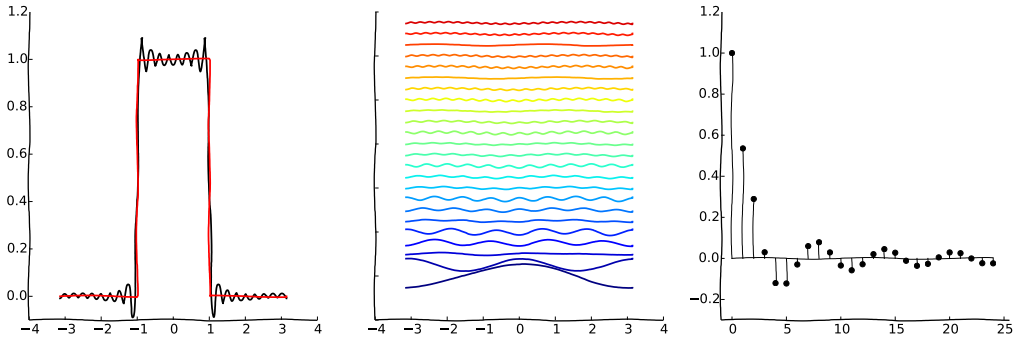


Figure 2.3. The discrete Fourier transform of a square wave of frequency 2π truncated after 20 coefficients. The square wave can be seen as the combination of different wavelengths. The image on the left depicts the Fourier transform of a square wave up to the 20th Fourier coefficient. In the he middle image we show 20 individual waves that comprises the truncated Fourier transform. The image on the right shows the corresponding Fourier coefficients.

spectrum, which means that each possible frequency was probed individually. The downside of this procedure is that it suffers from a poor signal-to-noise ratio and long measurement times.

Today, almost all NMR spectrometers use pulsed Fourier-transformed NMR, which does not suffer from the shortcomings of the frequency sweep. In this technique, a radio frequency square pulse is used to excite the probe. This pulse contains contributions of all nearby wavelengths as depicted in Figure 2.3, and is able to excite all resonance frequencies at once. As a result, the net magnetization of the probe is no longer aligned with the external field and starts tumbling. The tumbling magnetic field induces a periodic current in a pick-up coil that is measured over time. The overall tumbling is caused by the precession of the individual spins. The resulting signal contains contributions from all resonance frequencies and is called a free induction decay (FID). By applying a Fourier transform to the FID signal, a frequency-domain NMR spectrum can be obtained.

Protein NMR

A protein of typical size will easily contain several thousand protons. Although it is theoretically possible to measure a unique signal for each of the shifts, in practice, one will encounter many overlapping signals. The difference in the resonance

frequency between two protons cannot be resolved by a spectrometer. Multidimensional NMR techniques that result in two- or more-dimensional spectra are used to ameliorate this problem (Ernst et al., 1990). The peaks along the main diagonal of the spectra correspond to the one-dimensional experiment. The off-diagonal peaks can reveal interactions through-space or through-sequence. Depending on the applied radio pulses and relaxation times between the pulses, different interaction types can be revealed. Correlation spectroscopy (COSY), a common two-dimensional experiment, gives off-diagonal peaks for nuclei which are covalently linked via one or two atoms. The Nuclear Overhauser effect spectroscopy (NOESY) reveals through space contacts between nuclei (Neuhaus and Williamson, 2000). The information collected by NOESY is paramount to revealing the three-dimensional structure of a protein. As there is no simple relation between the recorded peaks and the sequence of the polypeptide chain, these experiments do not disclose a congruence of peaks and nuclei in a protein. Sequential assignment is used (Cavanagh et al., 1996) to discover these relationships. The assignment process for all the Nuclear Overhauser effect (NOE) peaks is the most tedious and time-consuming step in NMR structure determination and might take several months, if not years, in the worst case.

There is an upper limit to the size of protein structures solvable by NMR. In part, this is due to an increase in overlapping peaks and to a faster loss of magnetization in larger macromolecules. As a consequence of the loss of magnetization, there is less time to measure the signal, which prohibits many multidimensional spectroscopy experiments. Ultimately, this lead to broad peaks in a crowded spectra with many overlapping peaks, making assignment impossible. A remedy is to use selective labelling, for example it is possible to make only the methyl groups of cystine residues NMR active while ignoring all other nuclei. But this introduces a new problem: using conventional methods, it is no longer possible to solve these structures based on such sparse data sets.

Finally, at the end of the assignment process, it is possible to arrive at a list of pairwise distance constraints from the NOE peaks. These serve as the most important information for structure calculation. Other common types of experimental data for structure calculation are J-couplings and residual dipolar couplings (RDC). J-couplings, which is also known as indirect dipole dipole couplings, are used to define angular restraints along the main chain rotational degrees of freedom. Using RDCs, it is possible to derive restraints on the orientation of bonds relative to the external magnetic field. The careful combination of sufficient restraints allows the computation of a protein structure with atomic resolution.

2.3 Bayesian Inference

Few theorems in statistics are as controversial as the one named after the 18th-century Reverend Thomas Bayes (Bayes, 1763). Bayes was the first to describe a special case of what is today known as Bayes' theorem in his posthumously published assay "An Essay towards Solving a Problem in the Doctrine of Chances". It was actually Pierre Simon Laplace (Laplace, 1774), who introduced the theorem in today's general form,

$$p(\boldsymbol{x} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{x})\, p(\boldsymbol{x})}{p(\mathcal{D})} \tag{2.7}$$

where $\boldsymbol{x}$ denotes the variable of interest, $\mathcal{D}$ the data we have collected on $\boldsymbol{x}$, and p a probability measure. The controversy surrounding this theorem lies not in the formula itself, as it is derived from the very rules of probabilistic calculus, but in the interpretation of probability as a "degree of belief" instead of the frequency of occurrence of $\boldsymbol{x}$. Based on this interpretation, Bayes' theorem allows us to update the initial degree of belief in $\boldsymbol{x}$, also called the prior $p(\boldsymbol{x})$ to arrive at the probability of $\boldsymbol{x}$ after we have observed $\mathcal{D}$, which is called the posterior. The likelihood $p(\mathcal{D} \mid \boldsymbol{x})$ is a function of the parameter $\boldsymbol{x}$ and expresses how probable the data are if we assume $\boldsymbol{x}$ to be true. Within this framework we can easily update our assessment on $\boldsymbol{x}$ as new information arrives. It allows us to quantify the confidence in the inference through the posterior distribution.

Let us illustrate the Bayesian probability through a simple example inspired by Sivia and Skilling (2006). Suppose we face the task of estimating the probability θ that a coin will come up heads if tossed. In this case, the likelihood is the binomial distribution, which gives us the probability of observing h heads out of n trials $p(h|\theta, n) = \binom{n}{h}\theta^h\,(1-\theta)^{n-h}$. We are less restricted in the choice of prior. Any proper and improper probability distribution on the interval $[0., 1.]$ that encodes our assumptions about that coin is possible. If we assume a uniform prior, then the posterior will be proportional to the likelihood. A uniform prior encodes our initial ignorance about the coin: all probabilities are equal. But does a uniform prior really reflect our everyday experience with coins? Would it not be more sensible to assume that even heavily biased coins are never below $\theta = 0.3$ or above $\theta = 0.7$? On the other hand, we might be more cautious and assume that this is a loaded coin from the start. Such a prior would locate most of the probability mass around $\theta = 0.0$ and $\theta = 1.0$. Figure 2.4 shows the effect of these prior distributions and

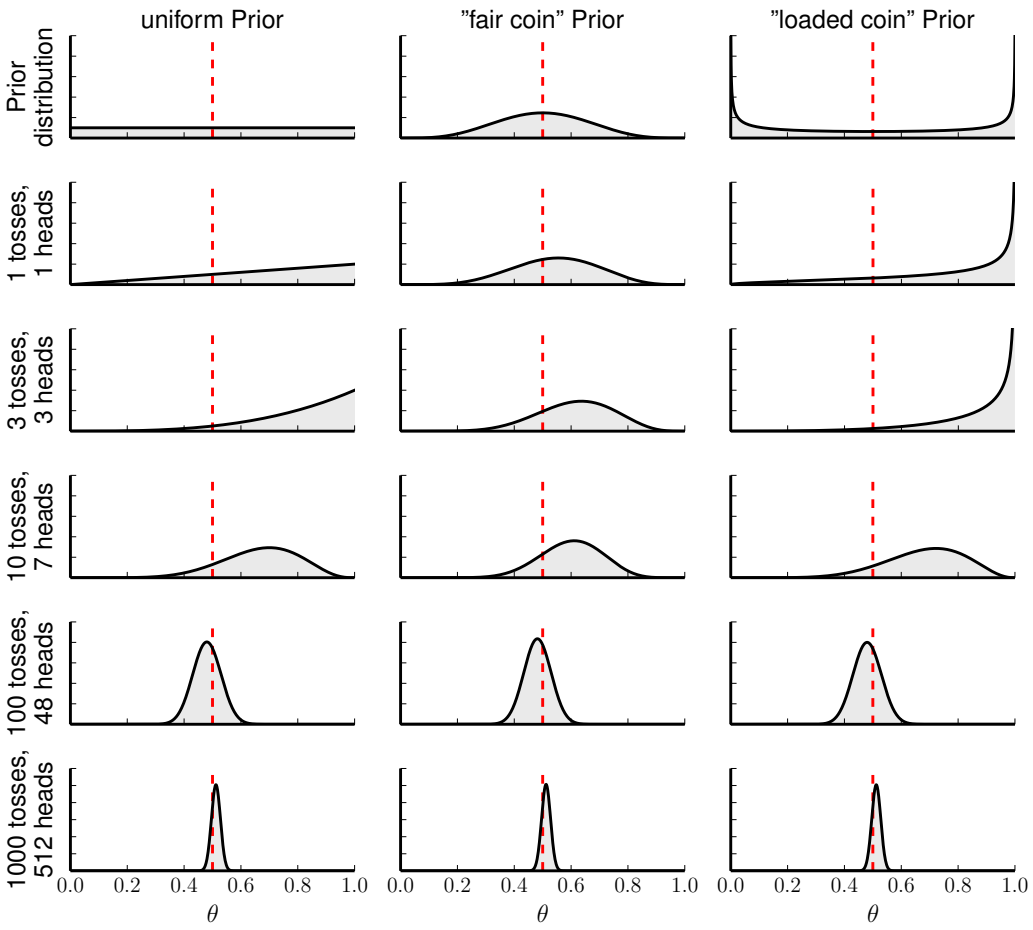


Figure 2.4. A coin toss example. This example demonstrates Bayes' Theorem at work. Each column details the inference process with a different prior distribution. The broken line shows the position of the true value of θ. The leftmost column uses a uniform prior. The prior used in the middle column assume that a reasonably fair coin is used. The rightmost column uses a prior that assumes a loaded coin; thus it emphasizes the extreme probabilities of 0.0 and 1.0. The effect of the prior is best observed for only a few rolls. The non-uniform prior keeps us from drawing excessively rash conclusions after seeing heads thrice, while the other posterior distributions are already fairly convinced that we are dealing with a loaded die. After 500 trials, the influence of the prior distribution is diminishing against the overwhelming evidence of the data and the posteriors are centred close to the true value.

how the posterior distribution evolves as more and more data is collected. When the volume of data is small, the posteriors provide quite different estimates of θ. As we observe more outcomes, the posteriors become more sharply peaked and converge towards the same conclusion.

Today, the Bayesian view of probability is well accepted within the statistics community. For a long time over the course of history the field of Bayesian statistics was perceived as being absurd (Gelman and Robert, 2011). The ideas of the early Bayesians like Laplace were abandoned by mainstream statistics in favour of Fisher's new randomization methods, sampling theory and tests of significance. It was Jeffreys (1939), Cox (1946), Good (1950) and De Finetti (1970) who again advocated for the use of the Bayesian interpretation of probability for scientific deductions. Furthermore, Cox showed that the Bayesian probability theory can be seen as an extension of Aristotelian logic to the realm of events under uncertainty. Their publications sparked a heated debate over the usefulness of the Bayesian interpretation of probability. Today, these quarrels between Bayesians and Frequentists have been settled and ideas from both statistical disciplines are used side by side.

The most important factor in the renaissance of Bayesian methods was the discovery of Markov chain Monte Carlo methods (Metropolis et al., 1957). These methods, together with the advent of powerful computers, made it possible to solve problems that resisted a Bayesian treatment because they were too complex to be tackled using only pencil and paper. Recently, probabilistic models and machine learning methods based on Bayesian principles were adopted in the field of structural bioinformatics and phylogeny, and led to the solution of challenging problems (Hamelryck, 2009a).

Bayesian methods are not restricted to drawing inferences on x. It is possible to compare several competing hypotheses on the basis of the evidence in favour of each one after having observed $\mathcal{D}$ Jeffreys (1939); Kass and Raftery (1995). Suppose we wish to compare which prior in our coin toss example reflects the observed data better. In this case, each choice of prior implies a different hypothesis $\mathcal{H}_1 \dots \mathcal{H}_n$. Recalling the Bayes' principles, we will to evaluate the posterior after having observed some data $\mathcal{D}$. Thus the posterior $p(\mathcal{H}_i|\mathcal{D})$ is given by

$$p(\mathcal{H}_i|\mathcal{D}) = \frac{p(\mathcal{H}_i)p(\mathcal{D}|\mathcal{H}_i)}{p(\mathcal{D})} \tag{2.8}$$

The prior $p(\mathcal{H}_i)$ allows us to express a preference for a hypothesis. This allows accounting for the notion that extraordinary claims require extraordinary proof. Thus, a critical statistician might assign a low prior probability to a hypothesis that seems unusual or unexpected. For example, a hypothesis claiming life on Mars would be assigned a much lower prior probability than would its opposite. But, it is usually assumed, that all hypotheses share the same prior probability. Moreover, $p(\mathcal{D})$ can be ignored as we compare all the hypotheses on the same set of data; thus,

$p(\mathcal{D})$ is a constant. This leaves us with only one term, which is usually called the model evidence $p(\mathcal{D}|\mathcal{H}_i)$. As an intuition, the evidence can be seen as the overlap between the prior and the posterior of the model for the data. More formally, this can be expressed as an integral over all parameters of the model underlying the hypothesis.

$$p(\mathcal{D}|\mathcal{H}_i) = \int d\boldsymbol{x}\, p(\mathcal{D} \,|\, \boldsymbol{x},\, \mathcal{H}_i)\, p(\boldsymbol{x},\, \mathcal{H}_i) \tag{2.9}$$

An alternative way of describing the evidence can be seen as the probability of generating the observed data from the model $p(\mathcal{D} \,|\, \boldsymbol{x},\, \mathcal{H}_i)$ whose parameters were drawn from the prior.

The evidence also includes a trade-off between model complexity and goodness-of-fit. An example for this intrinsic property of Bayesian inference is explained in Figure 2.5. Hence, it will select the simplest hypothesis that fits the facts. As the probability mass is limited, a hypothesis with high evidence will have its mass concentrated around the observed data. A more complex model will be able to explain a more varied data set, but at the expense of spreading its mass over a larger area of data set space. In contrast, for a very simple model the probability mass will have a sharp peak, but this mass will have little overlap with the observed model. These concepts have been used successfully in the field of phylogenetics (Huelsenbeck et al., 2001), where different evolutionary examples need to be compared.

2.4 Inferential Structure Calculation

Magnetic resonance data is usually not sufficient to determine a macromolecular structure by itself. The final structure depends on a number of subjective choices, which make the assessment of the quality of the structure difficult. This is also mirrored in the algorithmic structure calculation process, where typically a hybrid energy $E_{\text{hybrid}} = E_{\text{phys}} + w_{\text{data}} E_{\text{data}}$ is minimized. Here, E_{phys} describes how physically plausible a structure is, E_{data} captures the goodness of fit of the data, and w_{data} is the weight of the data chosen ad hoc. However, the formulation of structure calculation as minimization is an ill-posed inverse problem, as there is no guarantee of a unique answer to this question. Moreover, within this framework it is unclear how to estimate auxiliary parameters like w_{data}. Typically, these parameters are set

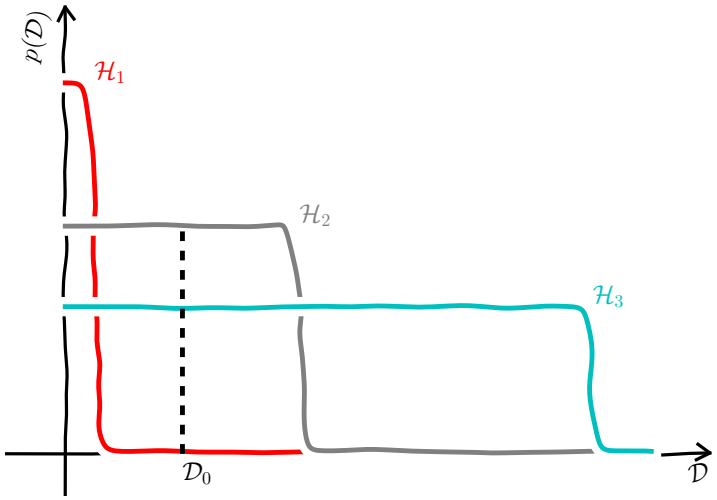


Figure 2.5. Schematic illustration of model evidence. This illustration shows how Bayesian model comparison includes a trade-off between model complexity and goodness of fit. For each hypothesis, we plot the model evidence against the space of all observable datasets. Bayesian reasoning stipulates to choose the most probable hypothesis, given the observed data D_0. But as the posterior $p(H_i|\mathcal{D})$ is a probability measure, it must assign a mass of one to the entire probability space. Hypothesis 3 is very flexible and can support most data sets. But its probability mass is limited; hence the mass associated with the observed data D_0 is small. In contrast, Hypothesis 1 is very specialized and can only explain a small subset of all observations. This allows H_1 to allocate a lot of probability mass to relatively few data sets. The actual observation D_0 is only weakly supported by H_3 and H_1. Ergo we select Hypothesis 2, which is the middle ground between H_1 and H_3, as the most probable one.

beforehand, thereby adding a possible source of bias, or need to be estimated via cross-validation.

To alleviate these problems, Rieping et al. (2005) developed a new approach called inferential structure determination ISD. The novelty of their approach is that they treat structure determination as a Bayesian inference problem. As discussed earlier, the use of probability theory allows reasoning in the presence of uncertainty. In the framework of ISD probabilities depend only on the observed data $\mathcal{D}$ and all relevant prior information $\mathcal{H}$. Hence, the probability of a conformation $\boldsymbol{x}$ is given by the posterior probability $p(\boldsymbol{x}|\mathcal{D},\mathcal{H})$. Using Bayes' theorem, the posterior is factored into

$$p(\boldsymbol{x}|\mathcal{D},\mathcal{H}) \propto p(\boldsymbol{x}\,|\,\mathcal{H})\; p(\mathcal{D}\,|\,\boldsymbol{x},\mathcal{H}). \tag{2.10}$$

The components of the posterior bear some similarity to the hybrid energy.

The prior $p(\boldsymbol{x}\,|\,\mathcal{H})$ describes how well a conformation $\boldsymbol{x}$ matches our expectations of protein structures. More formally it is defined as a Boltzmann distribution, which is determined by a physical force field E_{phys} and the inverse temperature of the system w_{phys}

$$p(\boldsymbol{x}\,|\,\mathcal{H}) \;=\; \frac{1}{Z(w_{\mathrm{phys}})} \exp\left[-w_{\mathrm{phys}} E_{\mathrm{phys}}(\boldsymbol{x})\right], \tag{2.11}$$

where $Z(w_{\mathrm{phys}})$ is the partition function. As $Z(w_{\mathrm{phys}}) = \int d\boldsymbol{x}\, \exp\left[-w_{\mathrm{phys}} E_{\mathrm{phys}}(\boldsymbol{x})\right]$ only depends on the temperature and not on $\boldsymbol{x}$, it is typically not computed.

The likelihood $p(\mathcal{D}\,|\,\boldsymbol{x},\mathcal{H})$ captures how likely it is to observe the data for a given conformation. The calculation of the likelihood involves a forward model that calculates idealized observations from the conformation. Furthermore, an error model is employed to assess the deviations between prediction and observation.

The ideas of ISD are best explained using a simple example. Let us assume a structure with just a single angular degree of freedom φ (see Figure 2.6). Let us further assume that we have measured the three-bond scalar coupling constants (J-coupling) for this angle. The theory put forward by Karplus (1963) allows us to relate the J-coupling to the torsion angle φ through:

$$J(\varphi) = A\cos^2(\varphi + D) + B\cos(\varphi + D) + C \tag{2.12}$$

To complete the forward model, we assume for the sake of simplicity that A,B,C and D are known. The resulting function is not injective, as shown in Figure 2.6a.

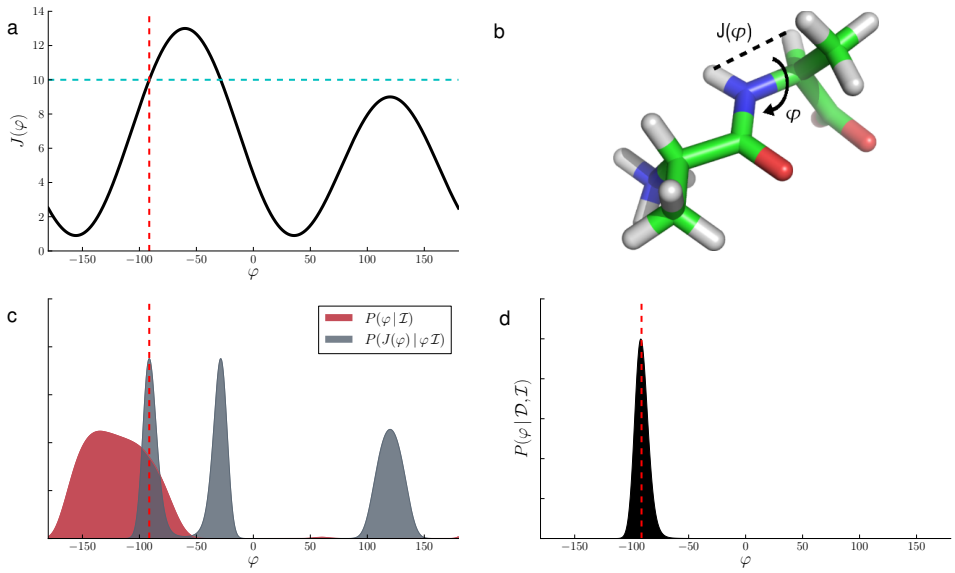


Figure 2.6. In Figure 2.6a, we show the Karplus curve for our example measurement as given in Equation 2.12. The measured values $J(\varphi)$ is indicated by a dashed cyan line. The correct angle is indicated by a red dashed line ($\varphi = -91.36°$). The structure of our toy example is shown in 2.6b. 2.6c shows the likelihood resulting from a scalar coupling measurement in grey and the combined prior in red. 2.6d shows the posterior probability (i.e. the product of the likelihood and the prior) in black.

If we assume a measurement of $J(\varphi) = 10.0$Hz generated by the true angle $\varphi = -91.36°$, there are two exact solutions: $\varphi = -91.36°$ and $\varphi = -28.64°$. If we also take a Gaussian distributed error of the measurement into account, we can compute the likelihood function depicted in Figure 2.6c. The likelihood function features three modes: two at the locations of the exact solutions and a third at lower probability for which J could not be exactly reproduced. In the absence of additional information, we cannot resolve the ambiguity. This information is encoded in a physical force field that is the base for the prior distribution shown in Figure 2.6c. In the prior distribution, the angles leading to steric clashes are now masked out. Hence, in the posterior, the product of prior and likelihood, the only regions to remain are those where both distributions show significant probability mass (see Figure 2.6d). Moreover, the resulting posterior has only one mode at the true solution.

On paper, it seems relatively straightforward to write down the posterior distribution. But in practice, exploring the posterior distribution can easily become the real problem. For the one-dimensional example, it is still possible to enumerate all values of the posterior, however in the case of more than a few variables enumeration is infeasible.

ISD employs an elaborate sample scheme Habeck et al. (2005a) to draw inference from the posterior. Sampling offers several advantages over minimization, as it gives estimates on the uncertainty of the inference results. Furthermore, many quantities of interest are defined as high-dimensional integrals and sampling, if done correctly, can provide a good approximation of these integrals. Still, sampling is no silver bullet. Many algorithms are quickly trapped in local modes and will only explore a very small part of solution space. To overcome this problem, ISD uses Replica Exchange Monte Carlo (REMC) sampling Swendsen and Wang (1986). The concept of REMC borrows from statistical physics and expands simulated annealing Kirkpatrick et al. (1983). Multiple copies of the system, which are called replicas, are sampled in parallel. These systems do not interact and are simulated at different temperatures of a temperature ladder. Higher temperatures flatten the probability distribution from which samples are generated. States are exchanged between neighbouring replicas according to the Metropolis criterion Metropolis et al. (1957). This mechanism prevents states from being trapped in local modes, as they are always able to diffuse to higher temperature.

ISD uses a variation of this method, that involves two temperature-like parameters controlling the shape of the prior and likelihood independently. While the likelihood is scaled by a temperature, a Tsallis transform Tsallis (1988) is used for

the prior. The Tsallis ensemble is controlled by a parameter $q > 1$ and its energy is given by

$$E_q(\boldsymbol{x}) = \frac{q}{\beta\,(q-1)}\,\log\Big\{1 + \beta\,(q-1)\,(E(\boldsymbol{x}) - E_{\min})\Big\} + E_{\min} \tag{2.13}$$

where $E_{\min} \leq E(\boldsymbol{x})$ must hold for all configurations $\boldsymbol{x}$. For $q = 1$ it holds $E_{q=1}(\boldsymbol{x}) = E(\boldsymbol{x})$. The transform becomes smoother for $q > 1$, which enhances sampling and facilitates the crossing of energy barriers.

For local sampling within a replica hybrid Monte Carlo (HMC) sampling Duane et al. (1987) is used. This is an adaptation of the well known Metropolis–Hastings algorithm, where new states are not generated by random sampling but by a short MD simulation. So far, ISD has been successfully employed in a wide range of structure determination projects. Besides structure calculations from NMR data, applications include the combination of of NMR and X-ray data(Bayrhuber et al., 2008; Honndorf et al., 2012), and the solution of structures from solid state NMR (Shahid et al., 2012).

3

Predicting secondary structure from chemical shifts

First, we try to extract structural information from chemical shifts. Assigned chemical shifts are available relatively easily and early on in the NMR structure elucidation process. Hence, any information gained at this stage will benefit all further steps of structure elucidation. In the best case, the structural information is sufficient to determine the structure themselves.

In the following, we introduce and evaluate the methods that we have developed. Parts of this chapter have been published in Mechelke and Habeck (2013b).

3.1 Introduction

In Chapter 2.2 we introduced protein chemical shifts and explained that they depend not only on the nucleus itself, but also its immediate surrounding. So, why not use this subtle information about the local chemical environment of nuclear spins to draw first inferences about the protein structure in question? One of the problems is that the relationship between the three-dimensional structure and the local shift information is very complex. Despite the obvious problems, there has been growing interest in the recent years to access this information and utilize it for biomolecular structure determination (Case, 1998; Wishart and Case, 2001). The recent progress by Cavalli et al. (2007); Shen et al. (2008) in combining chemical

shifts with protein structure prediction programs shows that it is possible to obtain structures in atomic detail from chemical shift information alone.

The prediction of secondary structures is a simpler, but still a worthwhile task. The connection between secondary structure and chemical shifts has been known for a long time (Markley et al., 1967; Williamson and Madison, 1990; Pastore and Saudek, 1990; Wang and Donald, 2004). Secondary structure elements are defined by a characteristic hydrogen-bonding pattern and geometry (see Chapter 2.1). These regular patterns have a notable influence on the chemical shifts that can be detected. This correlation is used at all stages of NMR analysis including chemical shift assignment and structure calculation. The direct application is complicated as it is theoretically involved and computationally demanding to describe the connection between shifts and secondary structure on a fundamental physical level. Therefore, empirical computational methods have been proposed that aim to quantify this correlation and exploit it to predict secondary structure from mainly backbone chemical shifts (for a recent review see Mielke et al.(Mielke and Krishnan, 2009)).

Over the past years, the use of chemical shifts to predict secondary structure elements has been the subject of many studies. The first technique to employ chemical shifts to predict secondary structure was the chemical shift index (CSI) (Wishart and Sykes, 1994). CSI uses the deviation of the secondary chemical shift, a normalized transformation of the chemical shift, as an indicator for secondary structure. CSI is easy to implement and performs well despite its simplicity. Wang and Jardetzky (2002) introduced a probabilistic approach, probability-based secondary structure identification (PSSI). This method relies on univariate Gaussian distributions to approximate the distribution of chemical shifts for a given amino acid and secondary structure. The posterior probability for a particular secondary structure assignment is computed as the combination of the probabilities of several nuclei. The predictions are smoothed to arrive at more 'protein-like' predictions using a five-residue window. PLATON Labudde et al. (2003) uses a homology-based approach to predict secondary structure and compares the observed chemical shift patterns with those of a reference database. The retrieved reference patterns are then used to calculate probabilities for the secondary structure classes.

PsiCSI (Hung and Samudrala, 2003) uses the combination of an extended version of CSI and PsiPred (Jones, 1999), a homology-based secondary structure prediction algorithm. The final prediction of PsiCSI is generated by a neural network that uses the output of CSI and PsiPred as input. PECAN (Eghbalnia et al., 2005) combines sequence information and residue-specific statistical potentials to

yield energetic secondary structure scores. The approach by (Wang et al., 2007a), which is called 2DCSi, clusters all possible pairs of chemical shifts and secondary structure elements from a database and applies a nearest-neighbour classifier to derive the predictions. The scores of all possible chemical shift pairs are combined into a final prediction. Two very successful programs are TALOS (Cornilescu et al., 1999) and TALOS+ (Shen et al., 2009). These use a homology-based search to predict primarily dihedral angles from chemical shifts but also provide secondary structure predictions. The method DANGLE (Cheung et al., 2010) is a probabilistic homology search that also provides secondary structure predictions. Among the above-mentioned methods, PsiCSI and DANGLE seem to be the most accurate methods (Cheung et al., 2010) with prediction accuracies of more than 80%. Most of the recent methods use neural networks or complex statistical learning methods, which involve many parameters that are difficult to interpret and that obscure the relationship between secondary structure and chemical shift. Too often biologists and biophysicists consider only the output behavior of a complex algorithm, treating it like a "black box". But what is inside the box is equally important. In this chapter, we are interested in simple and transparent probabilistic models that capture the connection between secondary structure and protein chemical shifts. Our approach is based on hidden Markov models (HMMs) with continuous emission probabilities for the chemical shifts of CA, HA, C, CB, and N nuclei. The HMM architecture accounts for the sequential nature of a protein's secondary structure. Multivariate Gaussian probability distributions model the observed chemical shifts and their distribution for particular amino acids and secondary structures. This approach has several advantages over existing methods: (i) it is highly accurate and competitive with leading methods such as DANGLE, TALOS+ and PsiCSI; (ii) it provides probabilistic output, i.e. a full distribution over secondary structure states; (iii) it can deal with missing data; (iv) it can incorporate predictions by PSIPred or other background information.

3.2 Generative model for chemical shifts

We employ a generative model to relate secondary structure and chemical shifts. In contrast to a discriminative model, a generative model entails a description of how the observed data was generated from the hidden labels. Furthermore, generative models can more easily deal with missing data and need fewer training data

to reach their asymptotic error rate (Jordan and Ng, 2002). We chose an HMM to explain the sequential data. An HMM needs two components: first an output distribution that models the probability of observing the data conditioned on the hidden states and second, a distribution to model the transitions between the hidden states. The hidden states, our objects of interest, are the secondary structure types (helix (H), extended (E) and coil (C)). This allows us to state a posterior probability distribution over the hidden states and search for the most probable sequence of hidden states or sample from the distribution of hidden states.

To model the chemical shift distributions, we use continuous multivariate probability distributions $p(\boldsymbol{x}|a, s)$ that describe the simultaneous measurement of C, CA, CB, HA, N shifts stored in the five-dimensional vector $\boldsymbol{x}$ conditioned on a given amino acid a and secondary structure s. Often, signal overlap and/or considerable missing resonance assignments lead to incomplete chemical shift measurements. These missing values can be dealt with by analytically integrating out unobserved chemical shifts. Thus, we can analytically derive partial chemical shift distributions for patterns of missing measurements and eliminate the need to estimate chemical shift distributions for all possible patterns.

The distributions $p(\boldsymbol{x}|a, s)$ can be combined with prior distributions for amino acids $p(a|s)$ to obtain joint probabilities for the co-occurrence of $\boldsymbol{x}$ and a, given s: $p(\boldsymbol{x}, a|s) = p(\boldsymbol{x}|a, s)\, p(a|s)$. The propensities $p(a|s)$ are estimated by counting co-occurrences of amino acids and secondary structure types in the PDBSelect25 (Griep and Hobohm, 2010).

We consider a multivariate Gaussian (MG) distribution to model the emission of chemical shifts $\boldsymbol{x}$ $p(\boldsymbol{x}|a, s)$. The probability density function of a multivariate Gaussian distribution for a d-dimensional measurement $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{3.1}$$

where $\boldsymbol{\mu}$ is a five-dimensional mean vector and $\boldsymbol{\Sigma}$ the covariance matrix quantifying correlations between the chemical shifts of different nuclei. We choose a Gaussian because quantities that are subject to a large number of additive and independent effects follow a Gaussian distribution. Even for data that are not strictly Gaussian, a Gaussian distribution can provide a reasonably good approximation. Moreover, Gaussian distributions allow for an analytical calculation of marginal distributions that is valuable when dealing with missing measurements. The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated by maximum likelihood. Essentially, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the sample mean

shifts and covariances. We explore the importance of correlation between nuclei by comparing different correlated and uncorrelated Gaussians. We use the symbol "UG" to denote an uncorrelated Gaussian distribution, in which set all covariances to zero ($\Sigma_{ij} = 0$ for $i \neq j$). This representation is equivalent to modeling the shift of each nucleus with a univariate Gaussian. We denote the model with full covariance matrix MG or multivariate Gaussian. UG comprises 10 free parameters that need to be estimated, while MG has 20 parameters.

For estimation of the Gaussian distributions, we extracted the experimental chemical shifts from the VASCO database (Rieping and Vranken, 2010), a curated database of chemical shifts referenced with regard to their three dimensional structure. The corresponding secondary structure states were assigned using DSSP (Kabsch and Sander, 1983), which classifies residues into eight secondary structures classes: α helix (H), π helix (I), 3_{10} helix (G), β sheet (E), β bridge (B), hydrogen-bonded turn (T), bend (S) and loops. We group these types into three larger states and map H, I, G to helix (H), E and B to extended (E), and all other states to coil (C). We filter the chemical shifts before the estimation of the chemical shift distributions and use only residues with measurements for all nuclei(C, CA, CB, HA, N). Chemical shifts flagged as outliers by VASCO were removed. Table 3.1 provides a detailed list of the number of training examples broken down into the amino acid and secondary structure states.

Table 3.1. Number of training examples broken down into amino acid and secondary structure state.

AA	H	E	C	AA	H	E	C
ALA	3675	1405	3254	LEU	4395	2829	3441
ARG	2030	1319	2387	LYS	3158	1671	3877
ASN	1234	616	2902	MET	887	505	856
ASP	1821	842	4364	PHE	1524	1640	1448
CYS	455	595	1006	PRO	552	424	3264
GLN	1985	840	2004	SER	1743	1260	3975
GLU	4005	1703	3885	THR	1323	1893	3114
GLY	872	955	5992	TRP	458	495	460
HIS	772	567	1163	TYR	1066	1376	1203
ILE	2236	2591	1880	VAL	2200	3856	2597

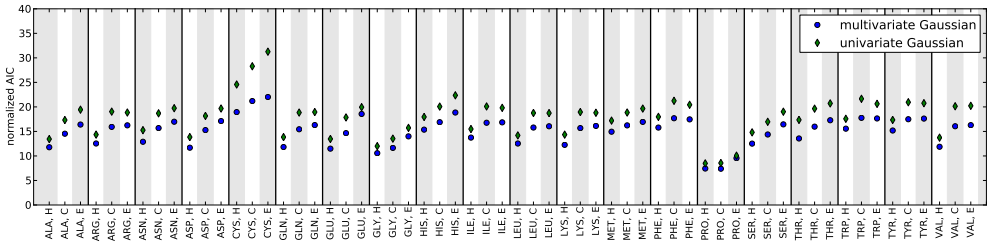


Figure 3.1. Normalized AIC values for chemical shift distributions fitted with univariate and multivariate Gaussian distributions and the MNIG distributions.

We estimated univariate and multivariate Gaussian distributions conditioned on secondary structure and amino acid type by maximum likelihood. To evaluate the quality of the emission probabilities, we applied Akaike's information criterion (AIC) (Akaike, 1974). The AIC is used for model selection, to balance the goodness-of-fit and complexity of the model. A lower AIC value, indicates a favourable model. We computed AIC values for the estimated univariate and multivariate Gaussian distributions (Figure 3.1). In all cases, the multivariate Gaussian model achieves a lower AIC than does the univariate model, which is not flexible enough to model the correlated chemical shift distributions accurately. Examples of the estimated multivariate Gaussian distributions are shown in Figure 3.2. The AIC values and visual inspection indicate that the distributions of sheet and coil shifts are not Gaussian but often skewed and heavy-tailed.

3.3 Hidden Markov models

The distributions that we have estimated do not capture the sequential nature of proteins and assume that neighbouring shifts are unrelated. In practice, this assumption does not hold (Wang et al., 2007b). For example, sheets and helices occur in segments joined by stretches of unstructured coiled regions and loops. To model these sequential correlations, we use a hidden Markov model (HMM) (Rabiner, 1989) to introduce dependencies between adjacent amino acids. An HMM is a probabilistic model, which assumes that a sequence of unobserved hidden states generated the observations. HMMs are a special case of Markov random fields (MRF), which have been applied before in both computational structural biology and in algorithms for biomolecular NMR (Zeng et al., 2011; Donald, 2011).

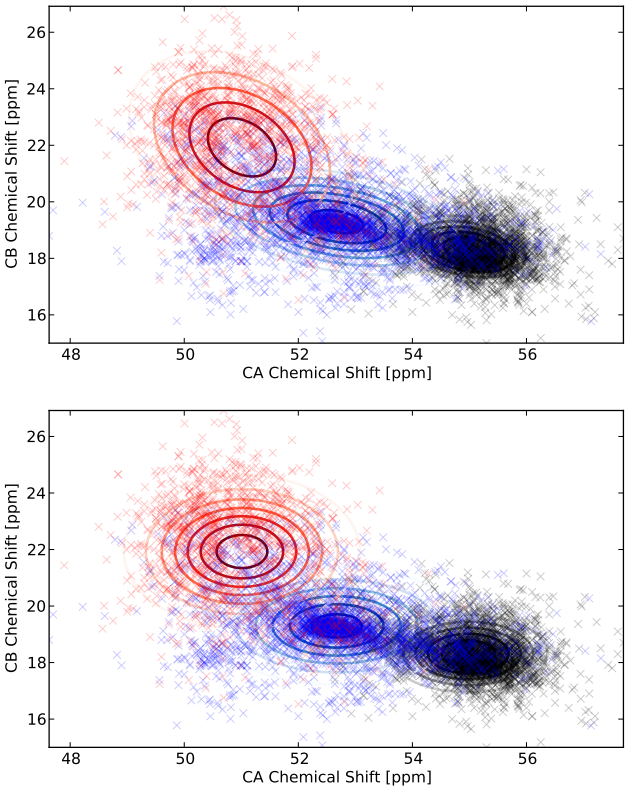


Figure 3.2. Chemical shift distributions for Alanine. Both panels show the observed CA and CB chemical shifts for the three secondary structure classes (H black, E red and C blue) as well as the fitted multivariate Gaussian (upper panel) and univariate Gaussian (lower panel) densities as contour lines.

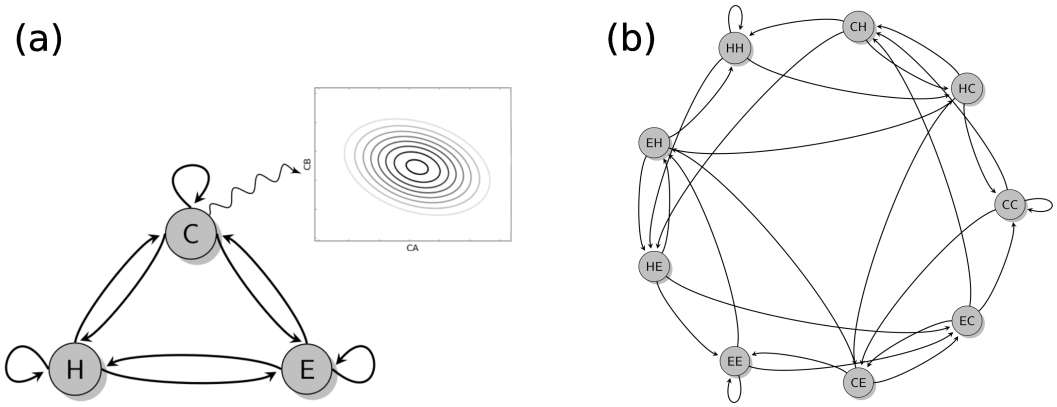


Figure 3.3. Architecture of the first- and the second-order HMM. A: Architecture of the first-order HMM. This is a fully connected model in which all transitions are possible. Every node emits chemical shifts conditioned on its secondary structure state and amino acid type. B: Architecture of the second-order HMM. We use a first order HMM with composite states to mimic the behaviour of a second-order HMM. Every node is labelled by the last and current state. Transitions are possible only if the current state of a node is the last state of the next node.

The hidden states are generated by a Markov process. In an Lth-order Markov process, the probability of the next hidden state depends only on the L last states.

In our specific application, we model the secondary structure types (H, E, C) as hidden states, which generated chemical shifts and amino acids with probability $p(\boldsymbol{x}, a|s)$. We use continuous multivariate Gaussians (Equation 3.1) as emission probabilities of the chemical shifts and discrete propensities $p(a|s)$ for the amino acids. Thus, the joint probability of a sequence of secondary structure elements $\{s_1, \ldots, s_N\}$ can be written as:

$$p(\{(\boldsymbol{x}_i, a_i, s_i)\}) = \prod_i p(\boldsymbol{x}_i|a_i, s_i)\, p(a_i|s_i)\, p(s_i|s_{i-1}, \ldots, s_{i-L}) \tag{3.2}$$

where $p(s_i|s_{i-1}, \ldots, s_{i-L})$ are the transition probabilities of the Markov chain. The architectures of the first- and the second-order HMM for secondary structure prediction are shown in Figure 3.3.

As we can assign the hidden states for the proteins in our training set, we do not need to use unsupervised training (Rabiner, 1989) and can estimate the emission and transition probabilities independently. We estimate the transition probabilities between the hidden states from PDBselect25 (Griep and Hobohm, 2010) by counting all secondary structure transitions in that database.

Many published methods for secondary structure prediction from chemical

shifts use stretches of up to five consecutive residues to make predictions for the central residue. This motivated us to use a second-order HMM, whose transition probabilities depend on the last two states. A second-order HMM offers a more precise model of the transitions between secondary structure state. We mapped the second-order to a first-order architecture by introducing composite states, which allows us to use the same algorithm irrespective of the order of the HMM. Again, the transition probabilities of the second order HMM are extracted from PDBselect25, thereby completing the HMM.

To infer the hidden states from chemical shifts, i.e. the unknown sequence of secondary structure types, we consider two options. The first option is the maximum a posteriori (MAP) estimate, which chooses a secondary structure sequence maximizing the joint probability (Equation 3.2) for given chemical shifts and amino acid sequence. The MAP sequence is obtained using the Viterbi algorithm (Rabiner, 1989). An alternative approach is to calculate the marginal probability of observing secondary structure s_i at position i:

$$p(s_i|\{\boldsymbol{x}_i, a_i\}) = \frac{\sum_{s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_N} p(\{(\boldsymbol{x}_i, a_i, s_i)\})}{\sum_{s_1,\ldots,s_N} p(\{(\boldsymbol{x}_i, a_i, s_i)\})} \tag{3.3}$$

We use the forward-backward (FB) algorithm (Rabiner, 1989) to calculate the marginal posterior probabilities (Equation 3.3).

3.4 Incorporation of PsiPred predictions

The use of evolutionary information, is available in the form of sequence profiles, has a long history in secondary structure prediction. A popular method is PsiPred (Jones, 1999), which as an intermediate step predicts secondary structure propensities ψ. To combine the output of PsiPred with chemical shifts, we treat the PsiPred propensities for each residue as uncertain observations.

We describe the probabilities $p(\psi|s, a)$ through a Dirichlet distribution that is estimated by maximum likelihood from a database of known structures The conditional distribution of observing the predicted secondary structure score vector ψ is given by

$$p(\psi|a, s) = \frac{\prod_{i=1}^{3} \Gamma(\alpha_i(a, s))}{\Gamma\left(\sum_{i=1}^{3} \alpha_i(a, s)\right)} \prod_{i=1}^{3} \psi_i^{\alpha_i(a,s)-1} \tag{3.4}$$

with parameters $\alpha_i(a, s)$ measuring the uncertainty of the PsiPred prediction. Since closed form maximum likelihood estimates for $\alpha_i(a, s)$ are not known, we use a fixed-point iteration method (Minka, 2000) to learn $\alpha_i(a, s)$ for all sixty combinations of amino acid and secondary structure.

3.5 Secondary structure prediction

We tested our HMMs by performing a 10-fold cross-validation on the VASCO database. The accuracy of our prediction is assessed by the Q_3-score, which is defined as the percentage of correctly predicted secondary structure across all residues of a protein. To gain insights into which aspects contribute most to the prediction performance, we tested various models of varying complexity. The prediction performance is shown in Table 3.2.

The conceptually simplest model neglects the correlations between chemical shifts of different nuclei and neighbouring positions in sequence. Despite these simplifications, the model predicts 72.5% of secondary structures found in the VASCO database correctly. This confirms that chemical shifts by themselves are a good predictor of secondary structure. In the next step, we impose sequential correlations and use first- and second-order HMMs. We can use two inference methods in the context of the HMM. First, the Viterbi algorithm provides the maximum likelihood solution. Second, the FB algorithm optimizes the marginal likelihood (Equation3.3). The first-order HMM, independent of the type of inference, predicts approximately 80% of all residues correctly. The second-order HMM does not improve the accuracy further. It seems that the boundary regions between secondary structure elements, which should be improved by the second order HMM, are not important for the prediction accuracy. The boundaries between secondary structures are, as previously noted by Rost and O'Donoghue (1997), difficult to define and often ambiguous. For this reason, we did not consider the use of a higher-order HMM.

The multivariate Gaussian distribution with a full covariance matrix provides a better approximation of the chemical shift distributions. It achieves a prediction accuracy of 74.7% if sequential correlations ($L = 0$) are ignored and rises above 82% if the multivariate Gaussians are embedded in a first- or second-order HMM.

The influence of the inference algorithm is small, although the marginal decoding of the first-order HMMs using the forward-backward algorithm leads to a minor

Table 3.2. Summary of secondary structure predictions on the VASCO database using 10-fold cross-validation. The table shows the prediction accuracies (in percentages) obtained with 10-fold cross-validation as measured by Q_3-score and SOV for all emission probabilities, all HMM architectures (order $L = 0, 1, 2$) and both inference methods (MAP using the Viterbi algorithm and the forward-backward algorithm, FB). The abbreviations, UG and MG denote uncorrelated univariate Gaussian emissions and correlated multivariate Gaussian emissions, respectively. We ran DANGLE and TALOS+ on the same data set. Columns headed by H, E, and C report the Q_3-score for individual secondary structure types. The overall Q_3-score is listed in the "All" column.

Order L	Emission $p(\boldsymbol{x}\|a, s)$	Method	H	E	C	All	SOV
0	UG	MAP	83.9	75.5	63.2	72.5	58.8
1	UG	MAP	86.0	79.1	75.2	79.5	74.4
1	UG	FB	85.9	76.1	78.1	80.1	74.7
2	UG	MAP	86.1	80.1	74.8	79.1	74.1
0	MG	MAP	84.7	77.7	66.3	74.7	61.4
1	MG	MAP	86.0	78.7	80.6	81.6	77.8
1	MG	FB	86.4	75.1	84.1	82.3	79.3
2	MG	MAP	86.1	80.3	80.0	81.9	78.8
1	PsiPred only	FB	72.6	77.9	80.0	76.9	69.2
1	MG + PsiPred	MAP	82.2	81.7	85.8	83.7	81.1
1	MG + PsiPred	FB	82.2	78.8	87.9	84.0	81.5
2	MG + PsiPred	MAP	83.0	82.8	85.2	83.9	82.0
DANGLE			90.9	75.7	77.2	80.7	72.8
TALOS+			79.3	71.1	83.4	78.5	68.1
PsiPRED			82.9	77.2	81.3	80.5	77.6

improvement in Q_3-score to 80.1% and 82.3% for univariate and multivariate Gaussian emissions, respectively. In the case of the second-order models, where some transitions are not allowed, the marginal decoding does not necessarily generate a valid path through the graphical model. Thus, we do not use decoding for the second-order HMMs.

Although the Q_3-score score has a long tradition in secondary structure prediction, it suffers from several shortcomings (Zemla et al., 1999). The Q_3-score only reports the average per-residue accuracy and does not take the segmentation

of the sequence into account. A high Q_3-score can be misleading, because it is biologically more relevant to predict the correct number of secondary structure segments. This problem is addressed by the segment overlap (SOV) score, which has been developed to provide a more realistic assessment of the prediction accuracy by focusing on the correct prediction of secondary structure segments.

Table 3.2 reports average SOV scores for our models. In the absence of sequential correlations, we achieve an SOV of roughly 60%. The introduction of first-order correlations leads to an increase in SOV to 79.3%, thereby confirming the previous observations regarding the importance of sequential correlations. As we already observed for the Q_3-score, the additional higher-order sequential correlations have no effect on the SOV score. The prediction accuracy differs between secondary structure elements, while all methods are able to predict helices reliably they struggle when it comes to sheets and coils. This observation is supported by confusion matrices (see Table 3.3); most misclassifications are between sheet and coil. To some extent, this can be ameliorated by including sequential correlations through HMMs.

The distribution of Q_3-score accuracy values for the entire VASCO database is shown in Figure 3.5 and it clearly highlights that an HMM leads to more accurate predictions.

3.6 Detection of residual secondary structure

We analysed the results in detail and found that some of the less accurate predictions stem from structures that, according to DSSP, are largely unstructured ,while our method clearly predicts regions of ordered secondary structure. It is not uncommon for disordered regions in proteins to exhibit residual secondary structure that can be implicated in molecular recognition. Based on the three examples studied by (Camilloni et al., 2012), we demonstrate that our HMM can detect weak secondary structure elements on the basis of chemical shifts. The predictions are presented in Figure 3.4.

The HMM prediction for the C-terminal domain of the sendai virus nucleoprotein comprises an α-helical region (Figure 3.4(a)) that is consistent with the experimental finding of a transient α-helical element with a core between residues 479–484 (Jensen et al., 2008). The C-terminus of the TFIIF-associating CTD phosphatase exhibits a 16-residue-long helical region that forms upon binding.

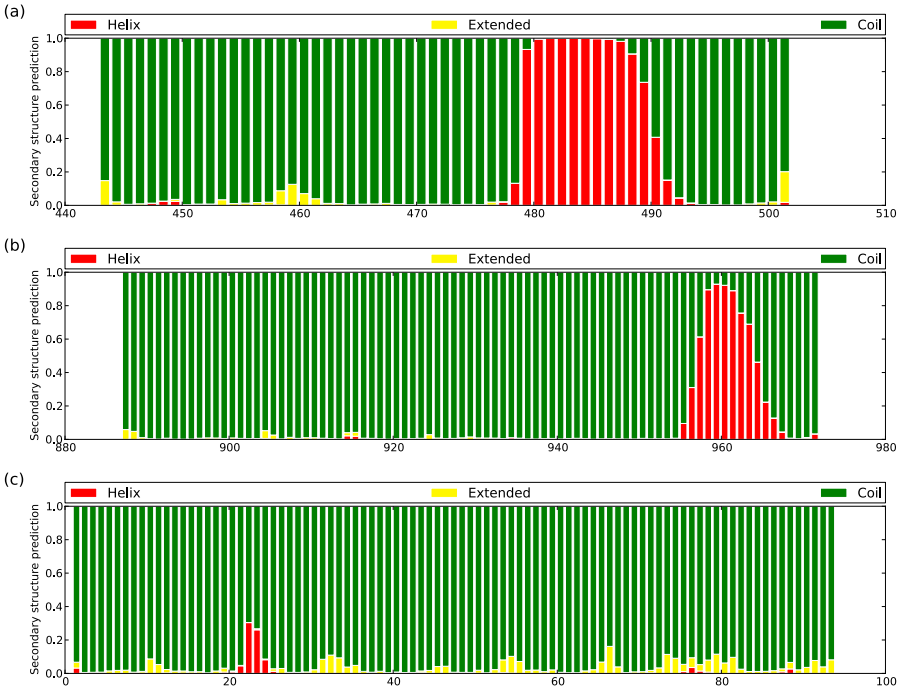


Figure 3.4. Detection of residual secondary structure. Secondary structure probabilities calculated using the forward-backward algorithm for (a) C-terminal domain of the sendai virus nucleoprotein (brmb 15123). (b) C-terminus of the TFIIF-associating CTD phosphatase (brmb 16296). (c) N-terminal region of p53 (brmb 17760).

Lawrence *et al.* (Lawrence et al., 2011) show that even in the disordered state, residues 945–955 form a nascent α-helical structure. Our HMM detects a strong α-helical signal in the same region (Figure 3.4(b)). The terminal domain of p53 contains a nascent helical turn between residues 18–25 (Wells et al., 2008) for which we observe a weak helix signal (Figure 3.4(c)).

3.7 Influence of individual chemical shifts on the prediction accuracy

For experimental reasons, chemical shifts are often incomplete. We test the influence of incomplete observations by systematically removing measurements. To this end, we extracted proteins with more than 90% complete measurements from

the VASCO database. This resulted in 1,121 proteins comprising 123,327 residues, from which we systematically removed selected nuclei and predicted the secondary structure using a first-order HMM with multivariate Gaussian emissions and Viterbi decoding. The results are presented in Table 3.4 (for the complete results, see Table A.1 of the Appendix).

For the complete set of chemical shifts, we achieve a prediction accuracy of 84.8%, which is close to the theoretical limit of the secondary structure prediction accuracy of $88 - 90$% (Rost, 2001). The systematic absence of a single chemical shift causes a small drop in prediction accuracy that ranges from 0.6% for missing C or CB shifts to 1.1% for missing CA or HA shifts. The accuracy of a single measured shift per residue depends largely on the nucleus and ranges between 80.1% for CA and 67.6% for N only, rendering the N shift the least informative. The HMM is not intended to work only on sequences, resulting in an accuracy of 49.5%, which is just slightly above chance. We reran the same analysis with correlated Gaussians as the model for chemical shifts, excluding sequential correlations and amino acid preferences $p(a|s)$. The results are shown in Table 3.4 (for the complete results, see Table A.1 of the Appendix). This analysis provides further support for our previous finding that the CA shifts are the most informative, while the N shifts are the least informative (Cheung et al., 2010). Moreover, the gain though the use of sequential information depends on the completeness and ranges from 5.% for complete measurements to 10.% for residues with only a single observed nuclei.

3.8 Incorporation of evolutionary information

Secondary structure prediction from sequences and multiple alignments is a long-standing problem in bioinformatics that has benefited from the inclusion of evolutionary information (Rost and Sander, 1993). To capture the evolutionary preference of a protein for secondary structures, we incorporate the predictions of PsiPred as a Dirichlet distribution conditioned on the secondary structure as additional emissions in our HMM. Thereby, we achieve a Q_3-score of up to 84% and a SOV score of up to 82%. Again, there is no benefit of using a second-order HMM. It should be noted that our HMM requires chemical shifts and that the prediction accuracy drops below that of PsiPred if no shifts are provided.

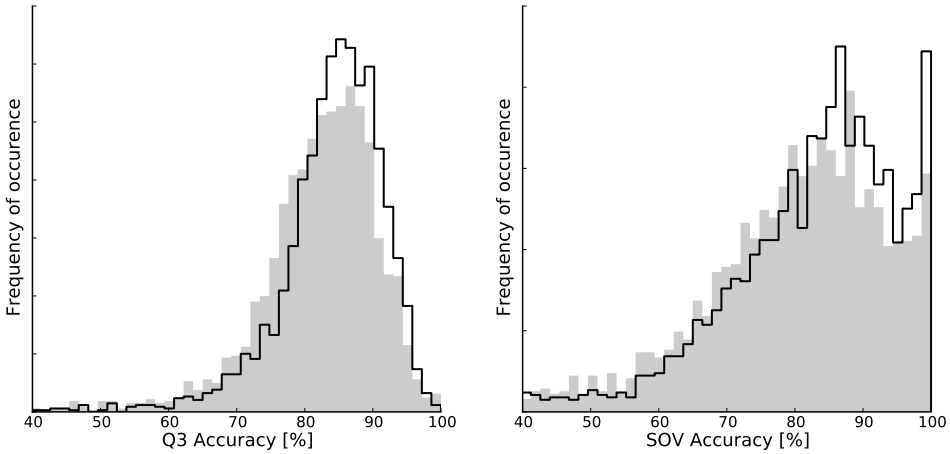


Figure 3.5. Impact of evolutionary information on the prediction accuracy. The left panel shows the accuracy as measured by Q_3-score; the right panel shows SOV. The grey histogram represents the FB-HMM with shift information only. The black line indicates the improvement when using the PsiPred output as additional information.

3.9 Comparison with other secondary structure prediction methods

To compare our results with other approaches, we made predictions for the VASCO database using DANGLE (Cheung et al., 2010) and Talos+, both of which are based on fragments. Since the fragment libraries are relatively small, we ignore that some VASCO entries will also contribute to the fragment libraries and compare the results directly to the cross-validated results obtained with our HMMs.

The methods differ most in SOV. Our first- and second-order HMMs achieve a significantly higher average SOV than do DANGLE and Talos+ (see Table 3.2). The higher accuracy can be traced to smoothing through the sequential correlation of the HMM. The difference in Q_3-score is less pronounced, for which DANGLE achieves an accuracy of 80%. In summary, correlations between neighbouring secondary structure states introduced by the HMM help to remove local artefacts and increase the accuracy compared to existing methods.

3.10 Analysis of prediction errors

The combination of chemical shift and evolutionary information leads a prediction accuracy of 84%. Although a promising result, it does not reach the theoretical upper limit of 90% as proposed by Rost and O'Donoghue (1997).

Table 3.3. Confusion matrices obtained with multivariate Gaussian emissions using a zero- and first- order HMM. Each row specifies the predicted secondary structure class and each column the actual class.

		0th order HMM			1st order HMM		
Actual		H	E	C	H	E	C
Inferred	H	65274	1242	16801	66678	160	8682
	E	1292	43895	21545	167	44441	13536
	C	10616	11314	76286	10337	11850	92414
Total:		77182	56451	114632	77182	56451	114632

With this result in mind, we set out to analyze the prediction errors. The boundary of secondary structure elements, which often cannot be defined unambiguously, is an obvious source of errors. To analyse this further, we excluded the residue preceding and following a change in secondary structure in the calculation of Q_3-score. If we include the PsiPred prediction in the inference we obtain an accuracy of 91.4% excluding the boundary regions. For the same regions, we still achieve a Q_3-score of 89.6% based on chemical shifts alone. Thus, a large number of prediction errors can be traced to the boundaries of secondary structures, where it is difficult to assign exact secondary structures.

Another source of errors are the DSSP assignments themselves. An example for which the DSSP assignment is misleading, is the budworm anti-freeze protein (PDB code 1N4I). Although the solution structure of this protein has a clearly defined beta-helix fold as shown in Figure 3.6, only four, very short strands out of the 10 present are found by DSSP. This discrepancy is also reflected in the predictions from the chemical shifts; the percentage of correctly predicted secondary structure depends on the assignments and ranges from 47.7% (DSSP assignment) to 80% (author assignment). For the same structure, there exists a crystal structure. The secondary structures predicted from chemical shifts agree in 84.7% of

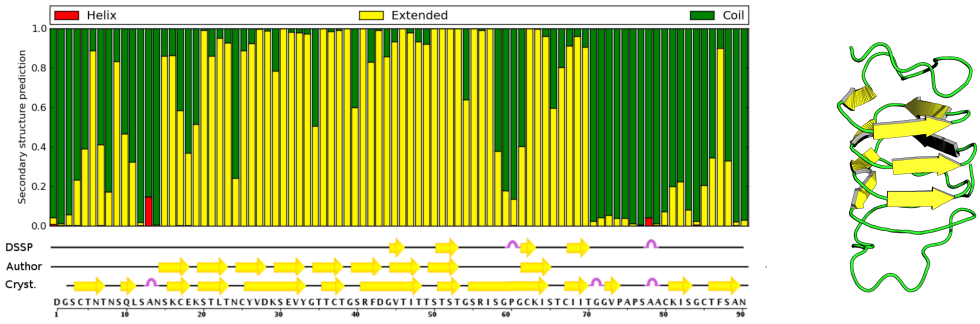


Figure 3.6. Comparison of assigned and predicted secondary structure. The left panel shows our secondary structure prediction as well as DSSP and author secondary structure assignments for budworm anti-freeze protein. The upper part of the left panel presents the marginal probabilities of the three secondary structure classes shown as stacked bars. The right panel is a cartoon representation of the NMR solution structure 1N4I.

residues, again showing that our approach is able to capture some of the secondary structure that DSSP fails to detect.

3.11 Prediction of φ/ψ angles from chemical shifts

We have so far demonstrated that we are able to extract secondary structure information from chemical shifts. Previous studies Cornilescu et al. (1999) have shown that it is also possible to predict dihedral angles from chemical shifts reliably. But expanding the existing hidden Markov models to predict dihedral angles poses several problems. Instead of the discrete secondary structure categories, the hidden states would be two continuous, angular variables. The inference in an HHMM with continuous hidden variables is intractable, except in the special case of Gaussian distributed states (Roweis and Ghahramani, 1999; Bishop, 1995). Even more severe, is the lack of a distribution that would allow us to model the five-dimensional chemical shifts conditioned on two angular variables.

To ameliorate these problems, we decided to use the probabilistic representation of the backbone angles put forward by Boomsma et al. (2008), and termed TorusDBN. Boomsma et al. represent local protein structure as a Markov model. The statistical model comprises 55 states that capture the local protein structure. Each state is associated with a distinct emission distribution over dihedral angles,

Table 3.4. Impact of incomplete chemical shift assignments on the prediction accuracy. In this experiment, we investigated the effect of incomplete chemical shift assignments on a subset of VASCO. We report in the leftmost column the observed chemical shift(s). The right columns show the result for a first order HMM using multivariate Gaussian emissions and Viterbi decoding. The left columns report Q_3-score when using only the conditional emission probabilities $p(\boldsymbol{x}|a, s)$ thereby neglecting sequential correlations and amino acid preferences. Each residue is classified by assigning the secondary structure state that maximizes $p(\boldsymbol{x}|a, s)$. Only proteins with more than 90% chemical shift completeness were used for testing.

	emission only				first order HMM			
Observed	H	E	C	All	H	E	C	All
C	76.5	65.4	63.6	67.6	84.2	65.5	77.4	76.9
N	62.8	50.2	52.7	54.9	85.6	57.8	61.9	67.6
CA	83.4	68.5	61.7	68.7	85.4	68.9	81.5	80.1
CB	78.7	67.2	41.3	58.3	84.8	73.0	66.1	73.3
HA	80.3	74.4	52.4	65.6	85.2	74.1	75.9	78.3
CA+CB	85.6	76.3	59.1	71.0	81.2	86.2	76.8	81.7
C+N+CA+CB+HA	88.0	83.6	68.7	79.1	88.6	84.2	82.1	84.8

amino acids and secondary structure. The states are connected by a transition matrix. Boomsma et al. demonstrated that this model captures the local sequence–structure preferences of proteins.

Thus, we can use the discrete TorusDBN states instead of backbone angles and use the mean torsion angles of each state as prediction. Now, instead of finding the most likely backbone angles, we need to find the most likely TorusDBN state given the observed chemical shift. To adapt our HMM to the new situation, we estimate the chemical shift distribution conditioned on the TorusDBN state. A difference to the secondary structure distributions is that the new emission distributions are no longer conditioned on the amino acid, because the TorusDBN states emit an amino acid. This poses a problem as the value of the chemical shift depends on the amino acid. To correct for the difference in chemical shift between amino acids, we use secondary shifts to predict the TorusDBN states. Secondary shifts are the differences between chemical shifts and their corresponding random coil values. We use the training set from the previous application and assign to each residue of the proteins within this set the corresponding TorusDBN state. Based on this training, set we use the standard estimators for the Gaussian distribution

to infer $p(\bar{X}|t)$, where $\bar{X}$ denotes the secondary shift and t one of the 55 TorusDBN states. In structure space, the states are well defined and share little overlap. To check weather this is still true with respect to the chemical shifts, we show the distance matrix based on the chemical shift distributions in Figure 3.7. As the distance measure between two probability distributions $p(x), q(x)$ we use the Hellinger distance $H(p, q)$ given by

$$H^2(p,q) = 1 - \int \mathrm{d}x \sqrt{p(x)q(x)}. \tag{3.5}$$

The Hellinger distance for two multivariate Gaussians with means $\boldsymbol{\mu}_p, \boldsymbol{\mu}_q$ and covariance matrices $\boldsymbol{\Sigma}_p, \boldsymbol{\Sigma}_q$ can be calculated analytically as

$$H^2(p,q) = 1 - \frac{|\boldsymbol{\Sigma}_p|^{\frac{1}{4}}|\boldsymbol{\Sigma}_q|^{\frac{1}{4}}}{|\frac{1}{2}\boldsymbol{\Sigma}_p + \frac{1}{2}\boldsymbol{\Sigma}_q|^{\frac{1}{2}}} \exp\left\{-\frac{1}{8}\boldsymbol{u}^T\left(\frac{1}{2}\boldsymbol{\Sigma}_p + \frac{1}{2}\boldsymbol{\Sigma}_q\right)^{-1}\boldsymbol{u}\right\} \tag{3.6}$$

where $\boldsymbol{u} = \boldsymbol{\mu}_p - \boldsymbol{\mu}_q$. The distance matrix in Figure 3.7 reveals three large blocks and several smaller ones. The large blocks correspond to the three main secondary structure states. The smaller blocks represent special amino acids, like proline and glycine. The high similarity of some distributions suggests that the number of states is not optimal and that some states, although different in structure, are indistinguishable in their chemical shift profiles.

The transition matrix of our HMM is taken directly from TorusDBN. Testing of the algorithm proceeded as before by tenfold cross-validation on the VASCO test-set. A prediction was deemed correct if both angles where within 30.0° of those of the reference structure, as proposed by Cornilescu et al. (1999). We compared both inference algorithms, the Viterbi algorithm and the forward-backward algorithm. We use TALOS+ as a reference, which is able to identify 61.0% of all angles correctly. The results of the tests can be seen in Figure 3.8. Using the Viterbi algorithm, we are able to predict 66.8% of all angles correctly. The Forward-Backward algorithm achieves an accuracy of 62.4%. Although the HMM is an improvement over TALOS+, we were still surprised by the low accuracy. This is particularly evident if we compare our results with those published by Cornilescu et al. (1999) and Shen et al. (2009). The distribution of the per protein accuracy is also striking, TALOS+ produces a large number of structures for which it is unableto predict any residue correctly, but for the reminder it bests the HMM. One major difference between this work and previous publications is that we do not exclude any residues from the final assessment and restrict ourselves to highly resolved structures. Furthermore,

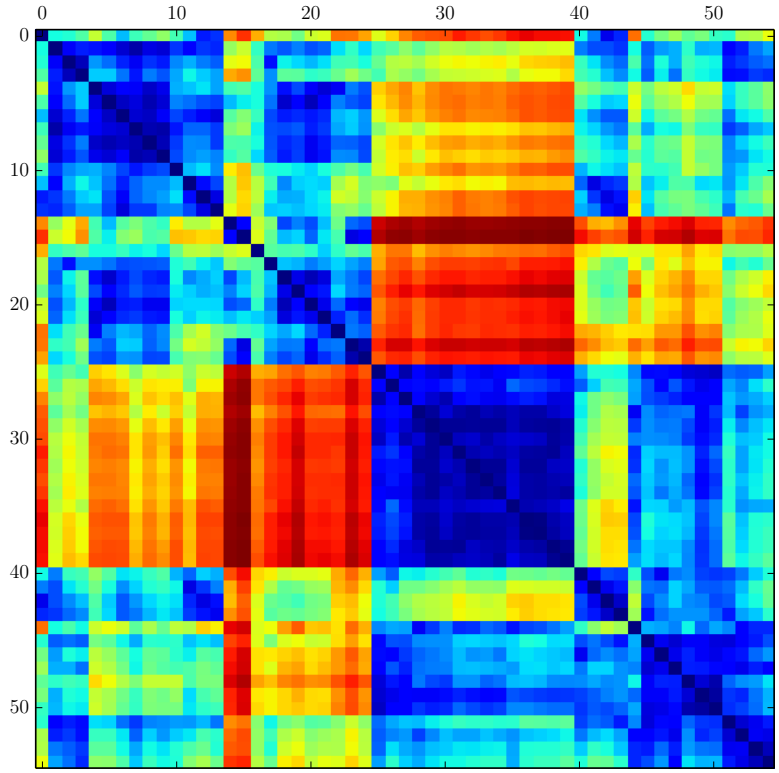
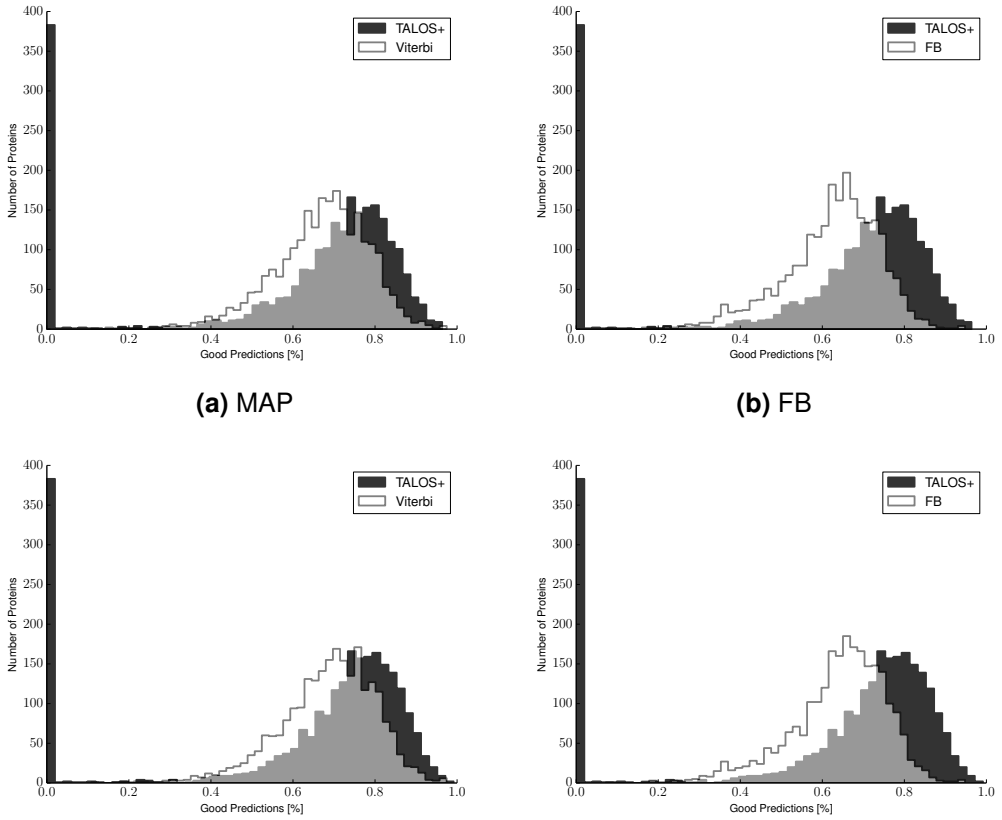


Figure 3.7. Hellinger Distance matrix. The figure shows the distance matrix between TorusDBN states clustered by their chemical shift emission distributions. The distance matrix contains three blocks 1–13, 17–39 and 40–54 that correspond to helices, coil regions and β sheets respectively.



(a) MAP **(b)** FB

(c) MAP filtered **(d)** FB filtered

Figure 3.8. Comparison of dihedral angle prediction accuracy with TALOS. We show histograms accumulated from the individual prediction accuracy per protein on the VASCO test set. The panels 3.8a and 3.8b show the HMM based prediction based on MAP and FB inference respectively, compared to the results of TALOS+. Panels 3.8c and 3.8d depict the results on those residues that are predicted to be reliable by TALOS+.

TALOS+ does not report a secondary structure for all residues that were identified as unreliable by the neural network. To exclude this as a possible source of bias, we calculated the accuracy again, but excluded the unreliable residues. The results on a per protein basis are shown in Figures 3.8c and 3.8d. The filtering done by TALOS+ leads only to a very minor improvement of about 2% for all algorithms. FB, MAP and TALOS now achieve an accuracy of 68.5%,64.0% and 62.7% respectively. The exclusion did not abolish the peculiar failures in the predictions of TALOS+. A more thorough inspection of the proteins in question reveals that they often miss a large number of chemical shift measurements or feature long unstructured patches, both of which lead to spurious matches in the TALOS+ database.

To conclude we can remark, that although the HMM performs better on average on our test set, TALOS+ remains the better choice for many reasonably well behaved proteins. Moreover, the prediction of dihedral angles is not as successful as the prediction of secondary structures. In particular, the large variance of the dihedral angle prediction – we considered in our test a divergence of up to 30.0° as a correct prediction – makes the result unsuited for use in structure calculation, where small deviations in the dihedral angles can lead to large deviations in atomic positions. One reason for the failure could be that a Gaussian is not flexible enough to describe the chemical shift distributions of the TorusDBN states. It is also uncertain, whether the TorusDBN states are an optimal representation of the structure space, but again learning the hidden states in an unsupervised fashion is computationally not feasible.

3.12 Conclusions

We introduced probabilistic models for secondary structure prediction from protein chemical shifts. Our model uses hidden Markov models with continuous emission probabilities to capture the connection between chemical shifts and secondary structure. We fit multivariate Gaussian probability densities to experimental chemical shifts of five nuclei (C, CA, CB, HA and N). Benchmark calculations show that first- and second-order HMMs achieve prediction accuracies of up to 82.3% and are competitive with current state of the art. Our findings suggest that higher-order sequential correlations are not able to improve the prediction accuracy substantially. The addition of evolutionary information can lead to an accuracy of up to 84.0%. This level of accuracy is on par with more complex approaches that tend to obscure the relationship between chemical shifts and secondary structure. Occam's razor, which states that we should prefer the more elegant, algorithmically parsimonious solution if it can describe the observations equally well, argues in favour of our approach, where the chemical shift of a nucleus depends on the secondary structure of that residue and the secondary structure only depends on the preceding one.

Our probabilistic formulation allows us to deal with missing chemical shift measurements gracefully by integration over the unobserved variables, which results in a distribution for the observed pattern of chemical shifts.

Another consequence of using an HMM is that we can calculate the probability

of each secondary structure for every residue. This full distribution over the secondary structure states provides us with a more realistic view of the dynamic nature of protein structures to the point where we are able to identify transient secondary structure elements in disordered regions.

Furthermore, we used an HMM to predict dihedral angles, where we used discrete states to model the torsion angles. Although the accuracy of our approach is comparable to competing approaches, there is still room for improvement.

Future work will focus on learning chemical shift distributions that discriminate better between secondary structure states and finding a more appropriate representation of dihedral angles. We will also focus of integrating the predicted secondary structure in the structure calculation process of ISD. Preliminary experiments indicate that using secondary structure can lead to improvement in structure calculation.

4

Weighting priors in Bayesian data analysis

In this chapter, we return to the topic of structure calculation from NMR. All NMR data are indirect measurements and, unlike X-ray densities, NMR data sets are insufficient for determining the structure on their own. They need to be interpreted in the light of an energy function to arrive at a set of three-dimensional structures. The resulting ensemble should be compatible with the energy function as well as with the observed data. But how do we know that the energy function is compatible with the data? And how can we quantify and adjust the compatibility? In this chapter, we will answer these questions within a Bayesian framework. To adjust the compatibility, we assign a linear weight to the energy function that can be increased if data and energy are compatible and decreased if not.

From a Bayesian point of view, each value of the weight can be seen as a different hypothesis. Bayesian inference stipulates determining the weight of an energy function based on the model evidence (see Chapter 2.3). This is challenging because the model evidence, a ratio of two high-dimensional normalization integrals, cannot be calculated analytically. All of the current approximations make assumptions of the functional form of the underlying problem. Here, we outline a replica-exchange Monte Carlo (REMC) scheme that allows us to estimate the model evidence through use of multiple histogram reweighting for a general class of data-analysis problems. The method is illustrated for examples in protein structure determination. The work presented here has been published in (Mechelke and Habeck, 2012).

4.1 Introduction

Let us have a more formal look at the theoretical problem. Recall that ISD (see Chapter 2.4) incorporates the energy function as Boltzmann prior, a trait shared with many complex data analysis problems. The Boltzmann prior $\pi(\boldsymbol{x}|\beta)$ is defined as:

$$\pi(\boldsymbol{x}|\beta) = \frac{1}{Z(\beta)}\, e^{-\beta E(\boldsymbol{x})} \tag{4.1}$$

where microstate $\boldsymbol{x}$ is the parameter of interest (for example, a protein structure), and E is an energy function that encodes our prior knowledge and guides the predictions towards plausible configurations. We assume, without loss of generality, that the energy is always greater than zero ($E > 0$). $Z(\beta)$ is the normalization constant or partition function, and is used to ensure that $\pi(\boldsymbol{x}|\beta)$ is a proper probability distribution ($\int \mathrm{d}\boldsymbol{x}\, \pi(\boldsymbol{x}|\beta) = 1$). The hyperparameter $\beta \geq 0$ is reminiscent of the inverse temperature in physical systems. It describes the influence of $\pi(\boldsymbol{x}|\beta)$ on the analysis and how strongly the degrees of freedom x are coupled. Even a Gaussian prior with a known mean can be interpreted as a special case of a Boltzmann prior. β than takes to role of the precision that controls the spread of the prior distribution. In ISD, the prior distribution is a simplified force field. Aside from ISD, prior distributions of the above type occur often in image analysis, where Ising models and Markov random fields are popular priors (Geman and Geman, 1984; Bishop, 1995).

In many data analysis problems, it is often unclear how much influence, controlled by the weight or inverse temperature β, the prior probability should have. If the system is at thermal equilibrium and the energy is an accurate description of the entire system, then $\beta = (k_B T)^{-1}$, where k_B is the Boltzmann constant and T the system's temperature. For protein structure determination, we would choose the temperature at which the sample was measured. However, the temperature might not always be known and not all force fields have a physical basis to choose the hyperparameter β. Unless there is a physically justifiable reason to set β, Bayesian reasoning dictates that we should treat it as a free parameter that is estimated from the data. Hence, it should be to be chosen depending on the quality of the data, the amount of data, and the type and size of the system.

An alternative view of β is that it is a regularization parameter controlling the flexibility of the model (Bishop, 1995). By increasing β, we increase the rigidity of

the model, which prevents overfitting of the data. Therefore, our Bayesian method of choosing β can be viewed as a probabilistic approach for determining the optimal strength of the regularizer in a consistent and data-driven way.

Such problems are encountered in many disciplines. For example, MacKay (1992) has pointed out the need to estimate hyperparameters in the context of Bayesian interpolation. To infer the strength of the prior he developed the "evidence framework". His framework relies on the assumption that the posterior distribution can be approximated by a Gaussian distribution around the maximum a posteriori estimate. But it also has a very severe shortcoming, as it is applies to only a very specific set of prior distributions.

Besides from the general insights of MacKay this problem has received some attention in the field of image analysis. So, we will proceed to review some of the more important work in image restoration and reconstruction. Here many common priors are intractable to the "evidence framework". At the same time, wrong prior assumptions and parameters can cause distortions (Li, 2009). A typical approach is to chose β by visual inspection, trial and error, or in the best case, by cross-validation (Johnson et al., 1991). Geman and McClure (Geman and McClure, 1987) developed a Bayesian approach to determine β. They use an expectation maximization algorithm to determine β iteratively in image restoration problems. This algorithm was later improved through the use of mean-field approximation (Pryce and Bruce, 1995; Zhou et al., 1997). Inoue and Tanaka (2002) and, more recently, Kiwata (2012) introduced methods that focus on temperature estimation for Ising and Potts model priors. But what all these methods have in common, is that they make some assumptions about the functional form of the prior and the likelihood function. Unfortunately, none of these approaches are applicable in the context of ISD.

We present in this chapter a general algorithm to select β that makes no assumptions on the form of prior distribution, likelihood function or configuration space. Another difference to existing approaches is that the method does not give a point estimate of β. Rather, following Bayesian reasoning, it infers the full posterior distribution of the hyperparameter. The full posterior allows us to characterize the uncertainty in our estimate. We illustrate our method on different data analysis problems. First, we make a short detour to image restoration, in order to demonstrate that this method is universally applicable, before we return to the original problem of protein structure determination.

4.2 Methods

Inverse temperature calibration by maximization of the model evidence

Let us recall Chapter 2.3, where we introduced Bayesian statistics. In Bayesian data analysis, the observed data $\mathcal{D}$ are interpreted by a likelihood function and combined with the data-independent prior distribution (4.1) that captures our prior knowledge. The likelihood function is a function of the configuration $\boldsymbol{x}$, which we denote by

$$L(\boldsymbol{x}) = \Pr(\mathcal{D}|\boldsymbol{x}) \tag{4.2}$$

to simplify the notation. The likelihood is the probability of observing the data $\mathcal{D}$ under the assumption that $\boldsymbol{x}$ is true. Bayes' theorem (Equation 2.7) states, that we can solve the inverse problem of estimating $\boldsymbol{x}$ from the observation $\mathcal{D}$ by multiplying the prior probability $\pi(x|\beta)$ and the likelihood function to obtain the posterior probability:

$$\Pr(\boldsymbol{x}|\beta, D) = \frac{1}{\Pr(D|\beta)}\, L(\boldsymbol{x})\, \pi(\boldsymbol{x}|\beta). \tag{4.3}$$

There is one caveat to this: the posterior $\Pr(\boldsymbol{x}|\beta, D)$, on which we base our estimate of $\boldsymbol{x}$, assumes that we know the true value of the hyperparameter β. As pointed out earlier, this assumption does not hold in many interesting application.

The model evidence or marginal likelihood (Equation 2.8), introduced in chapter 2.3 as

$$\Pr(\mathcal{D}|\beta) = \int L(\boldsymbol{x})\, \pi(\boldsymbol{x}|\beta)\, \mathrm{d}\boldsymbol{x}, \tag{4.4}$$

allows us to quantify the probability of the data for one particular value of β. An optimal value of β will scale the prior distribution (4.1) in such a way that models that are consistent with the observed data are more likely. We obtain the marginal posterior probability distribution of β by multiplying the evidence by a prior probability for β. In the following, we assume β to be uniformly distributed between 0 and a preset maximum $\beta_{\max}$.

The idea of using the model evidence to estimate hyperparameters is well established (Bishop, 1995). The real difficulty is the evaluation of the integral in Equation

4.4. In the case large systems like proteins, it is impossible to evaluate the integral directly. One of the main goals of this chapter is to arrive at a good approximation of this integral. To this end, we first introduced the quantity:

$$c_\lambda(\beta) = \int [L(\boldsymbol{x})]^\lambda \, e^{-\beta E(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x}. \tag{4.5}$$

This allows us to express $\Pr(\mathcal{D}|\beta)$ as $\Pr(D|\beta) = \Pr(D,\beta)/\Pr(\beta) = c_1(\beta)/c_0(\beta)$. Now the model evidence is given as the ratio of two high-dimensional integrals, which we later try to approximate. The nominator is the partition function of the ensemble at a given temperature. This integral measures the probability mass of the posterior. It is dominated by those configurations that are consistent with the prior as well as with the likelihood. The denominator is the partition function of the prior. Its volume increases with decreasing β. In the limit of $\beta \to 0$, the volume of $c_0(\beta)$ converges towards the total volume of the configuration space.

The behaviour of denominator and nominator becomes Occam's razor and optimally balances the complexity against the power of the model. The value of the nominator is highest when prior and data are consistent, emphasizing a small region of configurational space, while the denominator is low for prior models that are focused on a small subspace of configurational space. Thus, they can be seen as opposing forces that balance at an optimal β. The two integrals thus select a value of β that does not suffer from overfitting, yet has an positive effect on the inference process.

Still, the question of how we can evaluate the integrals, $c_1(\beta)$ and $c_0(\beta)$, remains. A feasible approach is to sample from $c_\lambda(\beta)$ and approximate the marginal likelihood from these samples. We use the ensemble for sampling

$$p(\boldsymbol{x}|\lambda,\beta) = \frac{1}{c_\lambda(\beta)} \, [L(\boldsymbol{x})]^\lambda \, e^{-\beta E(x)} \tag{4.6}$$

where we have two inverse temperatures, one for the prior probability and another for the data. If we set $\lambda = 1$, we obtain the posterior probability $p(x|\lambda = 1,\beta) = \Pr(x|D,\beta)$. If we set $\lambda = 0$, we obtain the prior distribution $p(x|\lambda = 0,\beta) = \pi(x|\beta)$. Values of λ between zero and one allow us to smoothly bridge between the prior and the posterior distribution.

We are interested in the β for which the model evidence $\Pr(D|\beta)$ is maximal. Instead of working with the model evidence directly, we take its logarithm, calculate the derivative and set it to zero:

$$\frac{\partial \log \Pr(D|\beta)}{\partial \beta} = \frac{c_1'(\beta)}{c_1(\beta)} - \frac{c_0'(\beta)}{c_0(\beta)} \stackrel{!}{=} 0.$$

Thus, we arrive at

$$-\frac{\partial \log c_\lambda(\beta)}{\partial \beta} = -\frac{c_\lambda'(\beta)}{c_\lambda(\beta)} = \langle E \rangle_{x|\lambda,\beta}$$

where $\langle \cdot \rangle_{x|\lambda,\beta}$ denotes an average over the extended ensemble (4.6). Therefore, the optimal prior weight, $\hat{\beta}$, is determined by the equality (Geman and McClure, 1987; Zhou et al., 1997):

$$\langle E \rangle_{\boldsymbol{x}|\lambda=1,\beta=\hat{\beta}} = \langle E \rangle_{\boldsymbol{x}|\lambda=0,\beta=\hat{\beta}}. \tag{4.7}$$

To put it simply, for an optimal choice of β according to Equation 4.7, the average energy of an ensemble generate from the prior ($\lambda = 0$) is the same as the average energy of a posterior ensemble ($\lambda = 1$). The result is intuitive. It states that for an optimal choice of β, the arrival of new data does not change the expected interaction energy. In the example of a protein, we would expect to sample energies close to those of the native ensemble from prior and posterior.

But how many maxima of the marginal probability can we expect? Is there at least one maximum? Let us first look at the expected interaction energy $U_\lambda(\beta) = \langle E \rangle_{x|\lambda,\beta}$. By taking the first derivative

$$\frac{\partial U_\lambda(\beta)}{\partial \beta} = \left\langle [E - U_\lambda]^2 \right\rangle_{x|\lambda,\beta} \geq 0 \tag{4.8}$$

into account, it is easy to see that the expected interaction energy is monotonically decreasing in β. Furthermore, we have $U_0(0) \geq U_1(0)$. If we choose $\beta = 0$, the prior becomes completely flat and configurations are drawn at random. The posterior, without the influence of prior, collapses to the maximum likelihood solution. Thus, all samples drawn from the posterior will scatter around the maximum likelihood estimate. We would expect any sensible energy function to assign an on-average lower energy to the maximum likelihood configuration, than to a configuration drawn at random from configurational space. Therefore, it is safe to assume that $U_0(0) \geq U_1(0)$.

Let us now take a closer look at the low temperature regime ($\beta \to \beta_{\max}$). There, the prior will only sample from its ground state, as the prior has collapsed to a sharp peak. The situation is different for the posterior. Only if posterior and prior perfectly coincide, we have $U_0(\beta_{\max}) = U_1(\beta_{\max})$ (a rare case, since a perfect prior precludes

the need to collect any data). Otherwise, the posterior will still feel some influence of the data, and seek a consensus between posterior and likelihood, thereby leading to higher average interaction energy. Therefore, $U_0(\beta_{\max}) \leq U_1(\beta_{\max})$. If we combine these results with the fact that $U_\lambda(\beta)$ is monotonically decreasing in β for all λ, we expect that both curves, $U_1(\beta)$ and $U_0(\beta)$ will cross at least once. Thus, there exists at least one β that satisfies our optimality criterion (Equation 4.7).

Replica-exchange Monte Carlo and multiple histogram reweighting

So far, we have presented only theoretical considerations on how to estimate β. On the more applied side, we need to solve how to sample from the distributions and, more importantly, approximate the high-dimensional integrals. For the sampling, we use replica-exchange Monte Carlo (REMC) (Swendsen and Wang, 1986; Habeck et al., 2005a) of the extended ensemble $p(x|\lambda, \beta)$ as implemented in ISD and introduced in Chapter 2.4. REMC, which is also known as "parallel tempering" (Geyer, 1991), is a variant of the Monte Carlo algorithm by Metropolis *et al.* (Metropolis et al., 1957), that simulates the joint distribution $\prod_{r=1}^{R} p(x_r|\lambda_r, \beta_r)$ of multiple configurations at different inverse temperatures (λ_r, β_r). Configurations can be exchanged between neighbouring systems according to the Metropolis criterion (Metropolis et al., 1957). For the prior ($\lambda = 0$) and the posterior ($\lambda = 1.$), we run two independent REMC simulations. The replicas in each of the simulations, $p(x_r|\lambda_r, \beta_r)$, are chosen in such a way that they bridge between the target distribution, e.g. $p(x|\lambda = 1, \beta)$, and a flattened version of the energy landscape, which is more suitable for sampling, e.g. $p(x|\lambda = 0, \beta = 0)$. We experimented with running only one, very long REMC simulation that bridges between $p(x|\lambda = 1, \beta)$ and $p(x|\lambda = 0, \beta)$. This has the downside that convergence time of a REMC simulation depends quadratically on the length of the chain, leading to a very slow convergence of the long chain. Furthermore, $p(x|\lambda = 1, \beta)$ and $p(x|\lambda = 0, \beta)$ often sample different regions of conformation space. It is more efficient to run two REMC simulations, one for the prior and the other for the posterior expectation. This has another advantage as the expectation of the prior only needs to be estimated once, as it is transferable between similar data sets (Geman and McClure, 1987).

A free parameter is the maximum inverse temperature $\beta_{\max}$. Both replica simulations will only explore $U_0(\beta)$ and $U_1(\beta)$ up to $\beta_{\max}$ In the REMC simulation of the prior ensemble $p(x|\lambda=0, \beta)$ we chose replicas ranging from $\beta\colon 0 \to \beta_{\max}$. The

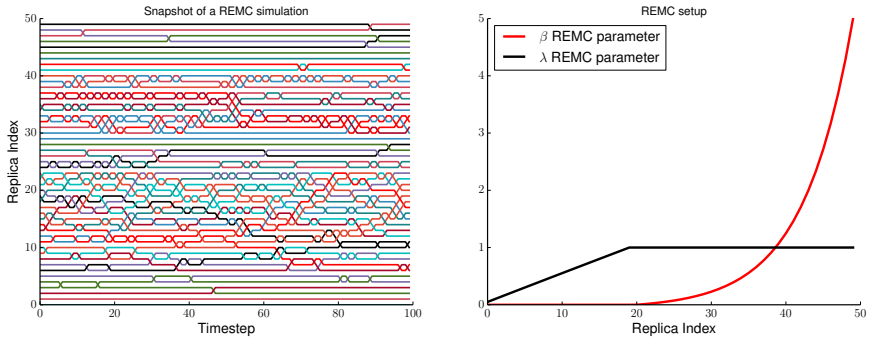


Figure 4.1. Example of an REMC Simulation. The left panel shows the diffusion of the Markov chains over time. The right panel shows corresponding replica parameters to simulate the posterior distribution

samples from all replicas are kept and used to estimate the expected interaction energy $\langle E \rangle_{x|\lambda=0,\beta}$ that we need to solve Equation 4.7. We proceed differently for the second REMC, which simulates the ensemble $p(x|\lambda=1,\beta)$. Here, we vary both λ and β in the REMC simulation to improve convergence. Figure 4.1 shows the evolution of the replica parameters. We start with $\lambda = 0.0$ and $\beta = 0.0$. At first, we switch on the data $\lambda\colon 0 \to 1$, but keep the prior switched off ($\beta = 0$). This is necessary to ensure proper sampling, as a Markov chain generated from the likelihood function alone is very likely to get stuck in a local mode. The additional heat baths allow a stuck system to escape from a local mode by diffusion to a higher temperature replica. If you look a the Markov chain of a particular replica, this will be visible as a "tunnelling" effect (see Figure 4.1). In a second step, we start increasing $\beta : 0 \to \beta_{\max}$ to study the influence of the Boltzmann prior. We use all samples of this simulation to estimate the expected interaction energy $\langle E \rangle_{x|\lambda=1,\beta}$. Instead of specifying a $\beta_{\max}$ it is also possible to update $\beta_{\max}$ interactively until a crossing of the interaction energies is found.

The states sampled by the two replica-exchange simulations are used to estimate the average interaction energies and the model evidence. Instead of focusing on a single replica, we pool all states of an REMC simulation and use multiple histogram reweighting (Ferrenberg and Swendsen, 1989) to estimate the density of states (DOS), defined as

$$g_\lambda(E) = \int \delta(E - E(x)) \left[L(x)\right]^\lambda \mathrm{d}x \tag{4.9}$$

where δ is the Dirac delta function. The DOS is a quantity from statistical physics that describes how many states can occupy a single energy level E. Typically, the

DOS is estimated using the Wang-Landau algorithm (Wang and Landau, 2001) or histogram reweighting (Ferrenberg and Swendsen, 1989). The latter was originally developed for discrete thermodynamic systems such as Ising models. As the name histogram reweighting suggests, the calculation depends on an iterative update of energy histograms that tend to the DOS. Histogram reweighting was later refined into the weighted histogram analysis method (WHAM) to investigate continuous systems like biomolecules (Kumar et al., 1992). This method is not without a drawback, however, as the spacing of the histogram in continuous systems introduces some arbitrariness and potential artefacts. The histogram methods suffer from poor scalability in large systems or for multiple energies. Habeck (2012) developed a nonparametric version of WHAM that eliminates the need to bin the energies. This algorithm estimates the density of states $\hat{g}_i$ associated with the energy of each of the sampled configurations x_i in such a way that $g(E) \approx \sum_i \hat{g}_i \, \delta(E - E(x_i))$. To eliminate the need for binning, Habeck (2012) uses a "infinitely resolved" histogram $H(E) = \sum_i \delta(E - E_i)$. Using the implicit representation of the DOS we have

$$\begin{aligned} Z(\beta) &= \int dE \, g(E) \, e^{-\beta E} \\ &= \int dE \, \frac{H(E)}{\sum_j N_j e^{-\beta_j (E - f_j)}} \, e^{-\beta E} \\ &= \sum_i \frac{e^{-\beta E_i}}{\sum_j N_j e^{-\beta_j (E_i - f_j)}} \end{aligned}$$

From the above equations, we obtain a nonparametric estimate of the DOS that does not require binning of the data:

$$\hat{g}(E_i) \equiv g_i = \frac{1}{\sum_j N_j e^{-\beta_j (E_i - f_j)}} \tag{4.10}$$

such that $Z(\beta) = \sum_i g_i e^{-\beta E_i}$. In the nonparametric version, we perform the following iterative updates until convergence to arrive at an estimate of $g(E)$:

$$\begin{aligned} g_i^{(t+1)} &= \frac{1}{\sum_j N_j e^{-\beta_j (E_i - f_j^{(t)})}} \\ f_j^{(t+1)} &= -\frac{1}{\beta_j} \log \left\{ \sum_i g_i^{(t+1)} \, e^{-\beta_j E_i} \right\} \end{aligned}$$

We reconstruct $g_{\lambda=1}$ and $g_{\lambda=0}$ for both replica simulations. Now the model evidence as a function of β can be expressed as

$$\Pr(D|\beta) = \frac{\int g_1(E)\, e^{-\beta E}\, \mathrm{d}E}{\int g_0(E)\, e^{-\beta E}\, \mathrm{d}E}. \tag{4.11}$$

Also the average interaction energies can be obtained from the density of states as

$$U_\lambda(\beta) = \frac{\int E\, g_\lambda(E)\, e^{-\beta E}\, \mathrm{d}E}{\int g_\lambda(E)\, e^{-\beta E}\, \mathrm{d}E}. \tag{4.12}$$

We evaluate these expressions in the log domain to avoid overflows. The samples from all replicas are combined into the density of states. This allows us to give a more accurate estimate of the expectations (4.12) and of the model evidence (4.11), by using an estimate from a single replica. Other methods like Gibbs sampling (Geman and Geman, 1984) of the posterior and prior distributions, as proposed in (Geman and McClure, 1987), result in significantly less accurate estimates of the expectations. But not only are the estimates more accurate, using the density of states we can quantify the distribution of the model evidence, which allow us to calculate the uncertainty of the estimate.

4.3 Applications

Calibration of the Ising model in image reconstruction

Our first application will be image reconstruction. We will restrict ourselves to a simple toy problem, that demonstrates all the complexity of real world applications. We use the two-dimensional Ising model on a $L \times L$ lattice with $L = 32$ as prior probability. The prior on a 32×32 black and white images is given by

$$\pi(x|\beta) = \frac{1}{Z(\beta)} e^{\beta \sum_{i\sim j} x_i x_j}$$

where $\sum_{i\sim j}$ indicates a sum over the neighbours of x_i and $x_i \in \{-1, 1\}$. The Ising model and its generalization, Markov random fields, are common priors in image analysis (Li, 2009; Wang et al., 2013). The parameter β controls the coupling of neighbouring pixels; for a small β neighbouring pixels are uncorrelated, whereas for a high β we observe a strong coupling and the prior emphasizes large patches of the same colour.

The goal of the inference is to reconstruct a black-and-white image from a series of noisy observations of the same image. The image we use shows a chequerboard pattern with $(L/K)^2$ blocks of K^2 spins bearing the same colour. We add noise by flipping each pixel with probability $1-\theta$, $\theta \in [0,1]$. This allows us to construct the likelihood of observing $y_i \in \{-1,1\}$, assuming that the original colour of the pixel is x_i.

$$\Pr(y_i|x_i,\theta) = \frac{1}{2}\sqrt{\theta(1-\theta)}\left(\frac{\theta}{1-\theta}\right)^{x_i y_i/2}. \tag{4.13}$$

The posterior probability for N observations is then given by

$$\Pr(x_i|y_i,) = \frac{1}{Z(\beta)} e^{\beta \sum_{i\sim j} x_i x_j + 1/2 \log(\theta/(1-\theta)) \sum_{j=1}^{N} x_i \left(y_i^j\right)}. \tag{4.14}$$

We can rewrite the negative logarithm of the posterior probability, in such a way that it that is equivalent to the energy function of a standard Ising model (Inoue and Tanaka, 2002) with a local magnetic field h_i

$$-\log \Pr(x|\beta, D) = -\beta \sum_{i\sim j} x_i x_j - \sum_i h_i x_i$$

where the local external magnetic field is determined by the data

$$h_i = N \log(\theta/(1-\theta))\, \overline{y}_i/2$$

with $\overline{y}_i \in [-1,1]$ referring to the mean value for pixel y_i averaged over N observations. The likelihood function also gives rise to the maximum likelihood estimator $\hat{x}_i = \mathrm{sign}\{h_i\}$.

We generated $N = 5$ noisy images for $K = 4$ at $\theta = 0.65$. In Figure 4.2, we show the true image together with the averaged observations and the maximum likelihood reconstruction.

An optimal estimate of β would match the inverse temperature of the original image. Based on the configurational temperature formalism (Rugh, 1997), we can calculate the inverse configurational temperature of x as

$$\beta(x) = \left.\frac{\mathrm{d}\, s(E)}{\mathrm{d}\, E}\right|_{E=E(x)} \tag{4.15}$$

where $s(E) = \log g_0(E)$ is the microcanonical entropy and $g_0(E)$ the density of states (4.9). An advantage of the Ising model is that density of states is well

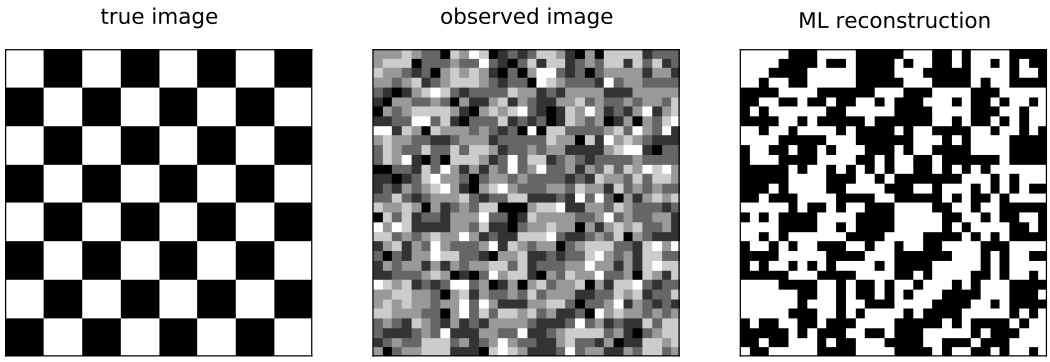


Figure 4.2. Data for image reconstruction with an Ising model. Left: true image; middle: average observed image with $\theta = 0.65$ averaged over $N = 5$ observations; right: maximum likelihood reconstruction.

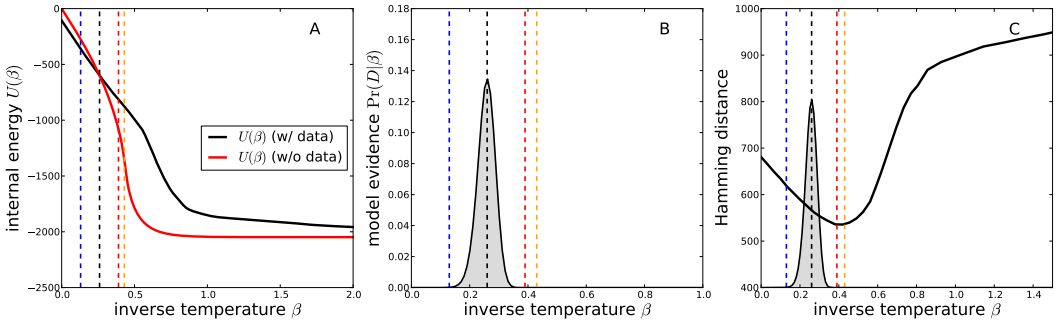


Figure 4.3. Estimation of the optimal temperature of the Ising model in image reconstruction. A: Internal energy with and without data as a function of inverse temperature. Dashed vertical lines indicate the inverse configurational temperature of the true image (red) and of the maximum likelihood image (blue), the critical temperature (orange), and the optimal inverse temperature obtained from Equation 4.7. B: Model evidence as a function of β. C: Hamming distance of posterior mean images and the true image.

known (Beale, 1996). The true image has an inverse configurational temperature of $\beta(x_{\text{true}}) = 0.38$. The inverse configurational temperature of the ML estimate is lower at $\beta(\hat{x}) = 0.13$. We apply our approach and ran two replica-exchange Monte Carlo simulations as outlined above. We explored the temperature up to $\beta_{\max} = 2$. Based on both REMC runs, we estimated the densities of states, $g_0(E)$ and $g_1(E)$, by multiple histogram reweighting.

The results are detailed in Figure 4.3. The first panel (Figure 4.3A) shows the internal energy curve $U(\beta)$ based on the replica runs with and without data. The curve obtained from $p(x|\lambda = 1, \beta)$ starts at a lower energy, as the data imposes restrictions beyond the prior energy. As these restrictions are absent for the curve obtained from $p(x|\lambda = 0, \beta = 0.)$, it sets in at a higher value. The phase transition of

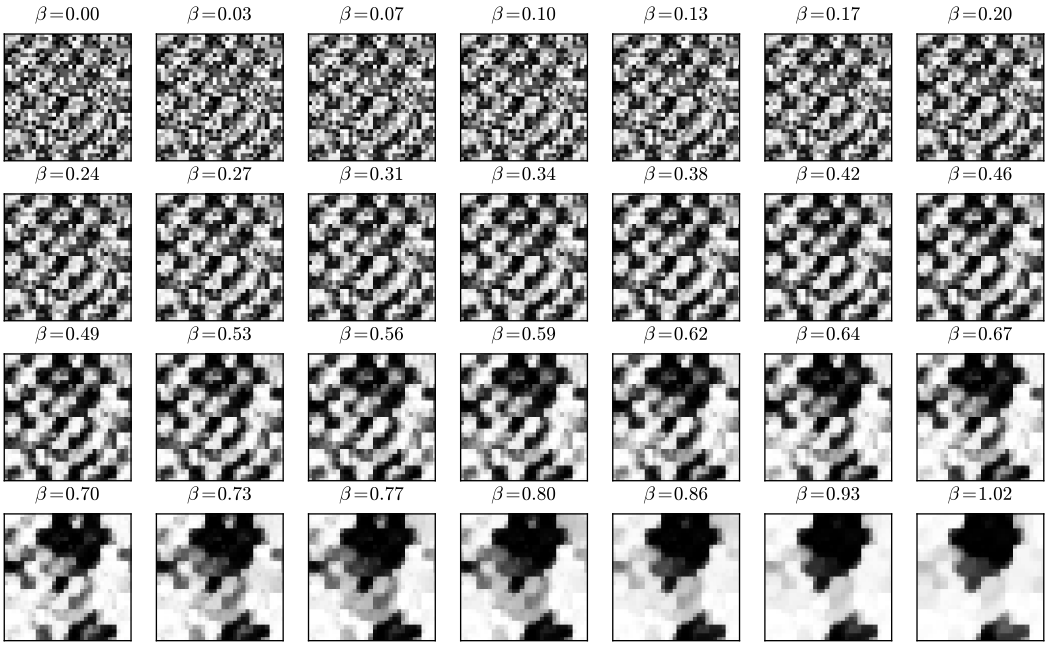


Figure 4.4. Posterior mean images for different choices of β. β values between 0.0 and 2.0 were probed during REMC simulation. However, we only show the restored images for $\beta \leq 1$, because no significant changes are observed for larger β.

the Ising model with its characteristic sigmoidal form occurs at $\beta_c \approx 0.44$.

As expected, the Bayesian estimate is somewhere between the two extrema given by the configurational temperature of the true image and of the maximum likelihood reconstruction. Indeed, the model evidence peaks at 0.26, with a width of 0.03 (Figure 4.3B). We measure the accuracy of the reconstruction by the Hamming distance $H(x, y) = \sum_{x_i \neq y_i} 1$.

In Figure 4.3C, we show the Hamming distance between the mean posterior image and the original image for the whole range of β. According to this metric, the optimal choice is $\beta = 0.39$, near the inverse temperature of the original image of $\beta = 0.38$. Still, the Bayesian estimate of $\beta = 0.26$ improves the accuracy considerably compared to the maximum likelihood estimate at $\beta = 0$. The plot of the Hamming distance in Figure 4.3C also demonstrates the risk of emphasizing the prior too much. A Boltzmann weight greater then 0.6 will lead to a reconstructed image, whose accuracy is lower than that of the maximum likelihood estimate.

The mean posterior images for all simulated values of β are shown in Figure 4.4. This figure nicely illustrates the phase transition, where we go from speckled images to reconstructions with large patches of similar colour. The Ising model allows us to test further the behaviour of the Bayesian inverse temperature estim-

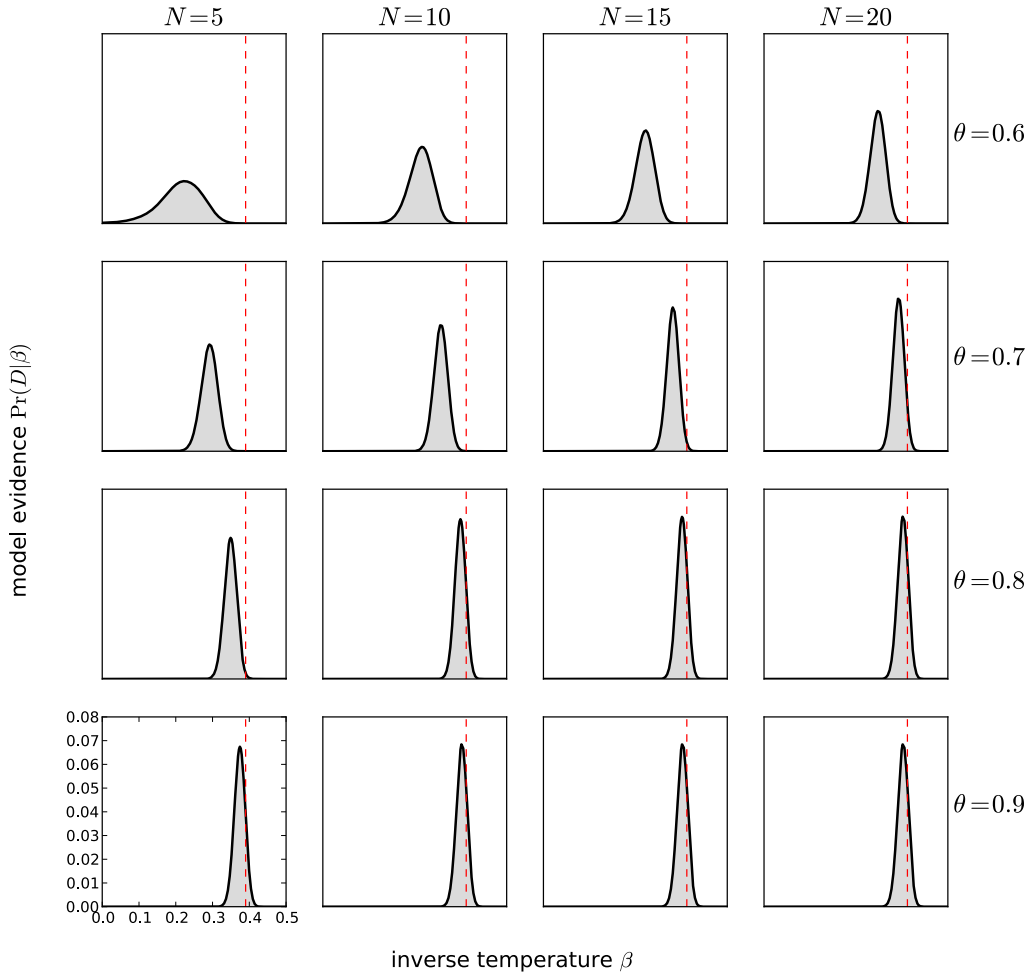


Figure 4.5. Model evidence for decreasing noise level $\theta = 0.6, 0.7, 0.8, 0.9$ and increasing amount of data $N = 5, 10, 15, 20$. The dashed vertical line indicates the configurational temperature of the true image.

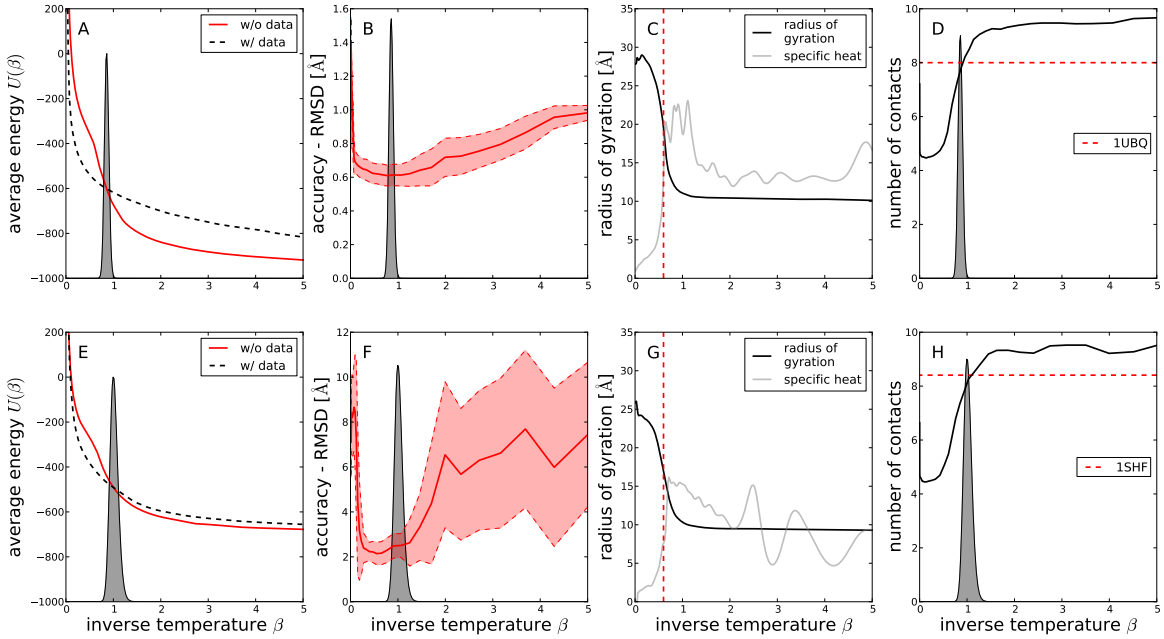


Figure 4.6. Calibration of the Boltzmann prior in protein structure calculation from NMR data. An approximate Lennard-Jones potential is weighted for high quality (top row) and a sparse data set (bottom row). The model evidence (panels A and E) peaks at $\beta \approx 1$. Panels B and F show the accuracy of the structure ensemble obtained for different choices of the inverse temperature; the filled region indicates one standard deviation. Also shown is the compaction of structures in terms of the radius of gyration (panels C and G); the dashed vertical line marks the critical value $\beta = 0.6$ at which the compaction sets in. Another measure of compactness is the average number of nearest-neighbour contacts (panels D and H); the dashed horizontal lines indicate the average number of contacts in the crystal structures 1UBQ and 1SHF, respectively. The radius of gyration and average number of contacts (black lines) are calculated over samples from the prior distribution, i.e. without using the data.

ates. If we increase the parameter θ, we decrease the noise in the observations. Furthermore, we can increase the number of observations N to add more data. Figure 4.5 shows the distribution of the evidence for all the different combinations of $\theta = 0.6, 0.7, 0.8, 0.9$ and $N = 5, 10, 15, 20$. An increase in the number of observed noisy images N shifts the estimate of $\hat{\beta}$ closer to the configurational temperature of the true image. In the case smaller data sets, β drops and the distribution becomes wider. The same observation can be made, when we increase the noise level (decreasing θ). A lower θ will result in a lower estimate of β. As the quality and amount of data deteriorates, Bayesian inference tells us to be cautious and to introduce only weak prior correlations.

Optimal weighting of force fields in protein structure calculation

As a second application, we focus on protein structure calculation using NMR distance measurements. Biomolecular structure calculation by ISD (see Chapter 2.4) is based on minimalist force fields that are considerably less complex and realistic compared to modern molecular dynamics force fields (Brünger and Nilges, 1993; Rieping et al., 2005). Typically, only van der Waal's contributions are considered as non-covalent interactions; electrostatic and solvent interactions are ignored (Linge and Nilges, 1999). In this application, we use the van der Waal's energy function from the Rosetta structure prediction software (Kuhlman et al., 2003). The energy function is a Lennard-Jones potential of the form $E(r_{ij}) = 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$ where r_{ij} is the distance between atoms i and j, ϵ denotes the depth of the energy well and σ_{ij} is the distance at which the inter-particle potential for these atom types is zero. To speed up calculation, interactions between atoms that are more than 5.5 Å apart are set to zero.

We will use the algorithm presented earlier to determine the optimal temperature of this force field. We expect that the modes of the β distributions are close to 1.0, as the samples were measured at room temperature. The first data set are high resolution data for ubiquitin (PDB code 1D3Z) (Cornilescu et al., 1998). Ubiquitin has 76 amino acids and adopts a beta-barrel structure that is closed by an α helix. The data set for ubiquitin comprises 1,444 NOE measurements that were converted to distance restraints. As described in the algorithm section, we simulated two REMC chains and computed the model evidence conditioned on the inverse temperature. Figure 5.5A depicts the expected prior energies curves as well as the resulting distribution of the model evidence $\Pr(D|\beta)$. In the case of ubiquitin we find that an optimal model evidence is attained at $\hat{\beta} = 0.87 \pm 0.05$, a little short of the expectation of 1.0 Another encouraging result is that the peak of the model evidence is within the region of smallest root mean square deviation (RMSD) from the X-ray structure 1UBQ. The RMSD depending on the inverse temperature and the distribution of model evidence is shown in Figure 5.5B.

As our second data set, we choose a sparse data set of the Fyn-SH3 domain (Mal et al., 1998; Rieping et al., 2005), a small beta-barrel domain. This data set contains only 154 distance measurements for a protein with 54 residues. Of these only 60 restraints are long-range restraints that are important for defining the architecture of the protein fold. We use the proposed algorithm to calculate the inverse temperature. For this protein, we estimated the inverse temperature at

$\hat{\beta} = 1.02 \pm 0.09$ (Figure 5.5E). The larger variance of $\hat{\beta}$ is the effect of fewer data for SH3 compared to ubiquitin. Again, the model evidence peaks in the β region that results in the most accurate structures (RMSD to crystal structure 1SHF; Figure 5.5F). The effects of a wrong estimate of β are more severe for the sparse data set. Already for $\beta = 2.$, the average RMSD to the crystal structure is 6.0Å. This highlights the importance of choosing β in a data-driven manner. If we put too low or too high a weight on the Lennard-Jones potential, the ensemble comprises multiple conformers, e.g. mirror images of the correct structure, leading to a large RMSD.

We performed REMC simulations of the prior ensembles of both proteins without any data. We use these to gain further insights into the properties of the energy function. One interesting observation is that the prior energy for both protein chains experiences a phase transition at $\beta \approx 0.6$ The Lennard-Jones potential favours compact structures and hence we expect to a see an effect of the radius of gyration R_g. At the critical point, the $R_g \approx 27/22$ Å drops to $R_g \approx 10/9$ Å for ubiquitin/SH3, respectively (Figures 5.5C,G). Experiments with freely jointed polymer chains indeed indicate that the temperature of this collapse transition is length dependent (Baumgärtner, 1980). We also looked at the average number of contacts depending on β (Figures 5.5D,H). We define a contact as two, non-neighbouring C_α atoms, that are closer than 7.5 Å. Figures 4.6D,H show the average number of contacts depending on the inverse temperature β. The phase transition that we observed for the radius of gyration and prior energy is also visible for the average number of contacts. At the critical point, the structures collapse and form densely packed structures. The contact number of the prior at the estimated inverse temperature almost matches that of the crystal structure of 8.0 (1UBQ) and 8.4 (1SHF) contacts on average. In contrast to the radius of gyration, the number of contacts does not plateau; it continues increasing until the structures are on average a lot denser the crystal structure. Whether the temperature of the Lennard-Jones potential is universal needs to be elucidated in further studies. Our test on two structures seems to indicate that the temperature of the Rosetta software is consistent.

4.4 Conclusion

In this chapter, we introduced a new fully Bayesian method to estimate the inverse temperature of Boltzmann-type priors in data analysis problems. This method is

applicable in the absence of a physical foundation that could allow us to determine the inverse temperature. Our temperature estimates rely on a replica-exchange Monte Carlo scheme and on nonparametric histogram reweighting to obtain accurate estimates of the density of states, rendering the process independent of the functional form of prior and likelihood, and permitting widespread application. Another strength is the probabilistic formulation, as we are able to express the uncertainty in our estimates.

We demonstrate the usefulness of the method in image analysis and protein structure determination from NMR data. An interesting result from the latter application indicates the existence of a critical temperature, at which a sudden compaction of the structures occurs.

In the next chapter, we will present application in protein structure calculation as well as generalization to multiple energy terms.

5

Optimal combination of statistical potentials in NMR structure calculation

The main objective of this chapter is to explore further the use of Bayesian model comparison in structure calculation. In the previous chapter (Chapter 4), we introduced a method to estimate the weight of a Boltzmann prior and used it to estimate the force constant of a potential function. Here, we focus on multiple potential functions and use our method to find an optimal combination of different potential functions with the aim of increasing the accuracy of structures from NMR data. In principle, this would enable us to tailor an optimal energy function for each structure calculation project. The relevance of the combination of different kinds of energy functions was highlighted recently by Pande (2011).

We demonstrate that an optimally weighted potential leads to an improvement in the accuracy and quality of the final structure, especially if the data are incomplete or noisy.

5.1 Introduction

An objective of this thesis is to improve the accuracy of the structures calculated from NMR experimental data. The data collected by NMR spectroscopy are indirect and incomplete measurements of the three-dimensional structure. In our

Bayesian framework, we combine the experimentally derived, pairwise distances derived from NOEs with our prior knowledge of protein structures, which is typically encoded in a potential function. This combination results in ensembles that are consistent with experiments and the laws of physics. A good energy function should guide us to the high-quality structures and provide an objective representation of our knowledge about protein structure, but it should not introduce artefacts that contradict the data. The better that potential function approximates the free energy of a protein structure, the more it will improve the generated structures. In order to avoid biasing the calculation process, but also for reasons of computational efficiency, one tends to use minimalist physical force fields that ignore complex effects such as electrostatic screening or solvent interactions in structure calculation. This strategy works well with high-quality data. But if we are faced with few experimental observations, we need expressive potential functions to fill the gaps. Potential functions can be grouped into two different classes (Skolnick, 2006): physics-based force fields (Ponder and Case, 2003) aim to approximate the underlying physical laws, whereas statistical or knowledge-based potentials (Sippl, 1995) are extracted from a structure database and describe the effective forces resulting from all interactions and do not necessarily have a physical basis. For example, the Lennard-Jones potential introduced in Chapter 4 belongs to the first class. Both approaches have their unique advantages and disadvantages. While physical approaches are transferable, the calculation is often time-consuming and it is challenging obtain accurate parameters. On the other hand, statistical approaches are able to capture interactions that are often difficult to describe using physics-based potentials; but they are limited by the data used to parameterize the potential, prone to overfitting, and not necessarily transferable (Das, 2011). Therefore, it seems attractive to combine different potentials.

But how can we combine physics- and knowledge-based potential functions? Several aspects should be considered when combining different potential functions. First, there is the risk of double counting interactions, as some facets of statistical potentials are already captured by physical potentials. For example, aspects of the Ramachandran plot, which is often employed as potential in structure calculation, can be explained by Lennard-Jones interactions. Second, knowledge-based potentials, which are averaged over large structural databases, are not universally transferable and may not represent the preferences of a particular structure.

To alleviate these problems and objectively combine different potentials we introduce an additional weighting factor to the statistical potentials, akin to the inverse temperature. This weight is then estimated by Bayesian model compar-

ison, as presented in Chapter 4. We demonstrate our methodology on a newly derived Ramachandran potential and estimate the inverse temperature of the Ramachandran potential in the presence of a Lennard-Jones potential as a function of the experimental data.

5.2 Dihedral angle potential

In his renowned paper, Ramachandran et al. (1963) introduced the parameterization of the protein backbone using φ/ψ angles. The two-dimensional scatterplot of these angles, which is now called the Ramachandran plot, was first used to predict the possible conformations of the protein backbone. Ramachandran et al. postulated, based on hard-sphere interaction in peptides, that some regions of the φ/ψ-space are not accessible to the protein backbone. These predictions, made just before the publication of the first protein structure, were later validated and proved to be accurate. Later studies (Hovmöller et al., 2002; Hooft et al., 1997), supported by crystallographic data, showed that the overall features of the original plot are correct, but many finer details, like the orientation of the α helix area, differ.

Over time, the Ramachandran plot has become one of the standard tools for analysing protein structures. Moreover, empirical energy functions for backbone dihedral angles derived from structural databases, "Ramachandran potentials", have been used in biomolecular structure calculation for almost two decades. Programs like Procheck (Laskowski et al., 1993), WHATCHECK (Hooft et al., 1996) and Molprobity (Davis et al., 2007) use empirical Ramachandran potentials to assess the quality of experimentally determined structures. The Ramachandran potential also play an important role in structure prediction (Rohl et al., 2004) and molecular dynamics (Buck et al., 2006). Several statistical models have been proposed ranging from two-dimensional histograms (Gong et al., 2007) to continuous representations based on linear interpolation, cubic splines and Gaussian mixtures (Mardia et al., 2007; Boomsma et al., 2008; Ting et al., 2010). These models ignore two fundamental aspects of statistical distributions over angular variables, namely their periodicity and their smoothness, which can result in artefacts during the refinement process (Kuszewski and Clore, 2000). Over the course of this section, we proceed to derive a parametric description of the distribution of backbone dihedral angles derived from a structural database, that we use as an energy function to guide structure calculation.

maximum entropy distribution for backbone dihedral angles

We employ a non-parametric approach based on the principle of maximum entropy (Jaynes, 1957) to estimate the density of the Ramachandran potential. As proposed by Pertsemlidis et al. (2005), we use a Fourier basis to represent the joint distribution of the φ and ψ backbone dihedral angles. This representation is inherently smooth and periodic, and has the advantage that it can easily cope with multi-modality as opposed to the unimodal von Mises or Kent distribution, which need to be combined into mixtures (Mardia et al., 2007; Boomsma et al., 2008; Ting et al., 2010) in order to represent the multi-modal Ramachandran plot. The functional form of the distribution of backbone dihedral angles is given by:

$$p(\varphi,\psi) = \frac{1}{Z(a,b,c,d)} \exp\{-E(\varphi,\psi)\}, \quad Z(a,b,c,d) = \int \exp\{-E(\varphi,\psi)\} \, \mathrm{d}\varphi \mathrm{d}\psi \tag{5.1}$$

where the Ramachandran potential $E(\varphi,\psi)$ is given by:

$$\begin{aligned} E(\varphi,\psi) = \sum_{i=0}^{k}\sum_{j=0}^{k} & \quad a_{ij}\cos(i\varphi)\cos(j\psi) + b_{ij}\cos(i\varphi)\sin(j\psi) \\ + & \quad c_{ij}\sin(i\varphi)\cos(j\psi) + d_{ij}\sin(i\varphi)\sin(j\psi) \end{aligned} \tag{5.2}$$

$Z(a,b,c,d)$ normalizes the dihedral angle distribution; k is the order of the Fourier expansion; a,b,c,d are the parameters we need to estimate.

We fit the expansion coefficients a,b,c,d to pairs of φ and ψ angles. During the optimization procedure, we need to evaluate the normalization constant $Z(a,b,c,d)$. Since there is no closed form expression for this integral, we evaluate it numerically using the two-dimensional trapezoidal rule. To circumvent overfitting of the data and to encourage a sparse solution, we introduce a Gaussian prior probability with unknown precision λ over the expansion coefficients a,b,c,d:

$$p(a,b,c,d|\lambda) = \left(\frac{\lambda}{2\pi}\right)^{2k(k-1)} \exp\left\{-\frac{\lambda}{2}\sum_{i=0}^{k}\sum_{j=0}^{k}(a_{ij}^2 + b_{ij}^2 + c_{ij}^2 + d_{ij}^2)\right\} \tag{5.3}$$

The precision of the prior λ is not known and is estimated simultaneously with the expansion coefficients. We use an iterative scheme in which we cycle through conditional updates of the expansion coefficients and of the precision. For fixed

precision, the log-posterior probability of the expansion coefficients is a convex function, which we optimize using the Powell minimizer (Press et al., 1989). The update of the precision for given a, b, c, d can be calculated analytically.

Backbone dihedral angle distributions

We extracted the backbone dihedral angles from the PDBselect25 (Hobohm et al., 1992), a database of representative protein structures with less than 25% sequence identity to estimate the Fourier coefficients in Eqn. 5.3. For reasons of computational efficiency and to avoid overfitting, we truncate the Fourier series at order k, which is selected by the Bayesian information criterion (BIC) (Schwarz, 1978). The BIC is an approximation of the marginal likelihood given by

$$\mathrm{BIC}(m) = -2 \log \mathcal{L}_{max} + m \log n$$

where $\mathcal{L}_{max}$ is the maximum of the likelihood of the data set for m parameters (here $m = 4k(k-1)$) and n data points. The BIC curves for all amino acid are shown in Figure 5.1. After a steep decline, the BIC value of almost all amino acids plateaus at $k = 5$. Based on the BIC analysis, we selected the Fourier expansion with an order of 5 The corresponding distributions are shown in Figure 5.2. They provide a good fit and capture important features such as the α-helical peak at $\varphi = -60^{\mathrm{o}}, \psi = -40^{\mathrm{o}}$. In addition, the backbone dihedral angle distributions of glycine and proline, which deviate from the standard Ramachandran plot, are captured well by the Fourier model. To compare our parametric estimates visually, we show histograms of the φ and ψ angles extracted from the PDBselect25 in Figure 5.3.

5.3 Data-driven weighting of the backbone potential

We would like to use the dihedral angle potential in ISD (Chapter 2.4) to encourage more protein-like backbone conformations. But integrating an additional potential into ISD is not completely trivial. How can we add the potential without introducing a bias? Recall that the combined potential function is $E = (w_{\mathrm{phys}} E_{\mathrm{phys}} + w_{\mathrm{rama}} E_{\mathrm{rama}})$, where $w_{\mathrm{phys}} = 1/k_{\mathrm{B}}T$ is the reciprocal temperature involving Boltzmann's constant

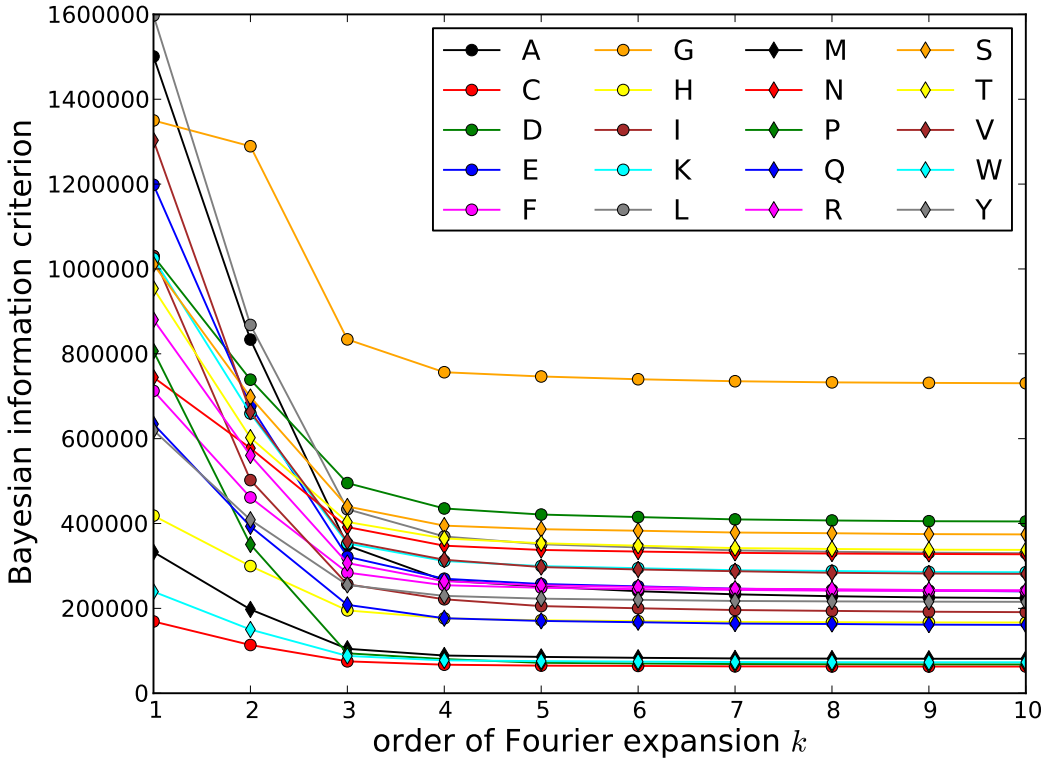


Figure 5.1. BIC analysis to determine the optimal order of the Fourier expansion. The BIC curves of the backbone dihedral angle distributions of all 20 amino acids (indicated as one letter codes in the legend) is shown in increasing Fourier order.

k_{B} and the absolute temperature T, and where E_{phys} is the Lennard-Jones potential adapted from the Rosetta software (Kuhlman et al., 2003).

Naively, we would set the weight of the backbone potential w_{rama} to one. But this is problematic because some aspects of the Ramachandran plot are already captured by the Lennard-Jones potential. To show this dependence, we simulated tripeptides using the Lennard-Jones potential. Each peptide comprises an amino acid flanked by two Alanine as termini. From these simulations, we collected the dihedral angles and estimated the backbone dihedral angle distributions, which are shown in Figure 5.4. The distributions show an outline that is roughly similar to the Ramachandran distributions in Figure 5.2, but they differ largely in the finer details and positions of the modes. The more subtle aspects, such as optimal hydrogen-bonding geometry (Porter and Rose, 2011) that result in pronounced peaks, cannot be reproduced by the Lennard-Jones potential alone. We analysed the connection between both potentials further by plotting the correlation between

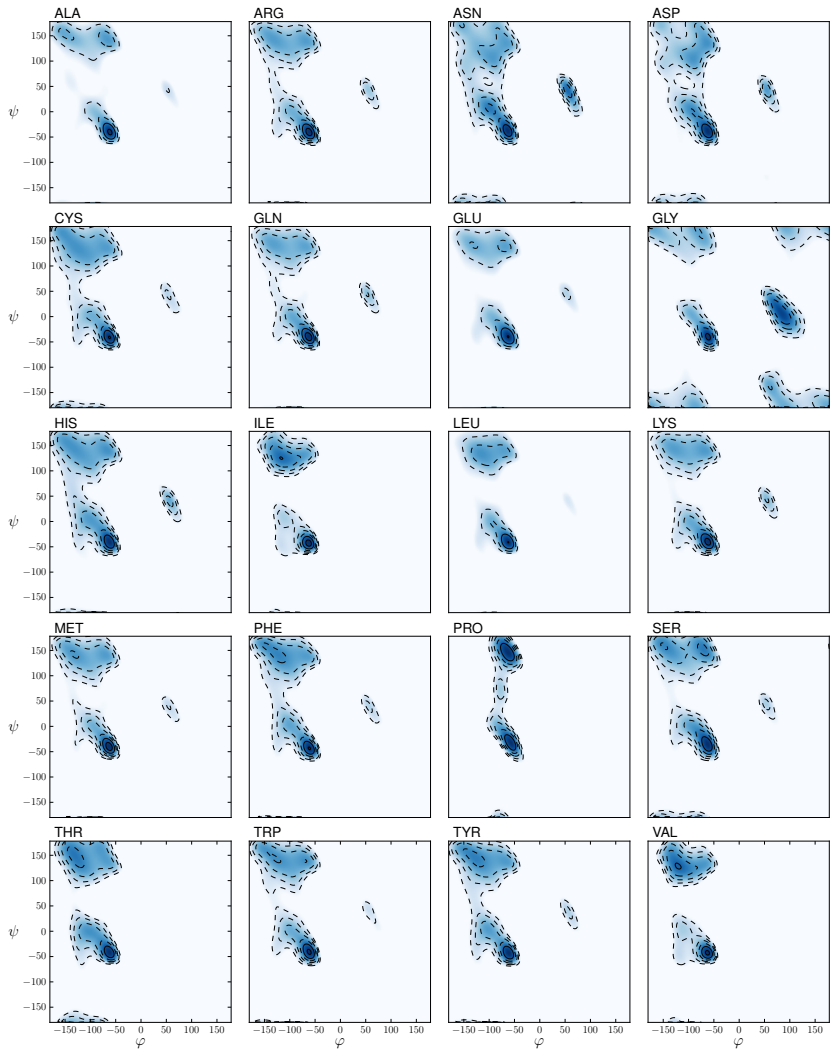


Figure 5.2. Backbone dihedral angle distributions of all amino acids estimated from high-resolution crystal structures. Heat maps of all φ/ψ distributions as approximated by the maximum entropy distribution outlined in this paper.

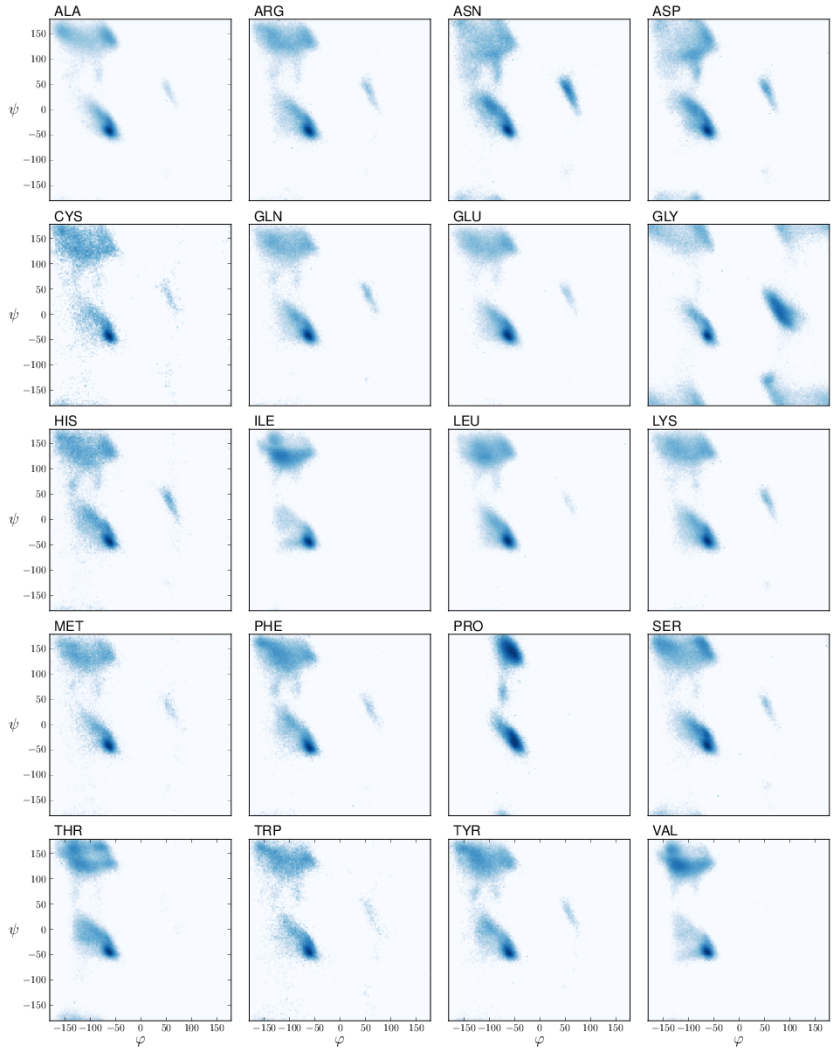


Figure 5.3. Empirical backbone dihedral angle distributions. Histogram of the empirical φ/ψ distribution extracted from high-resolution crystal structures.

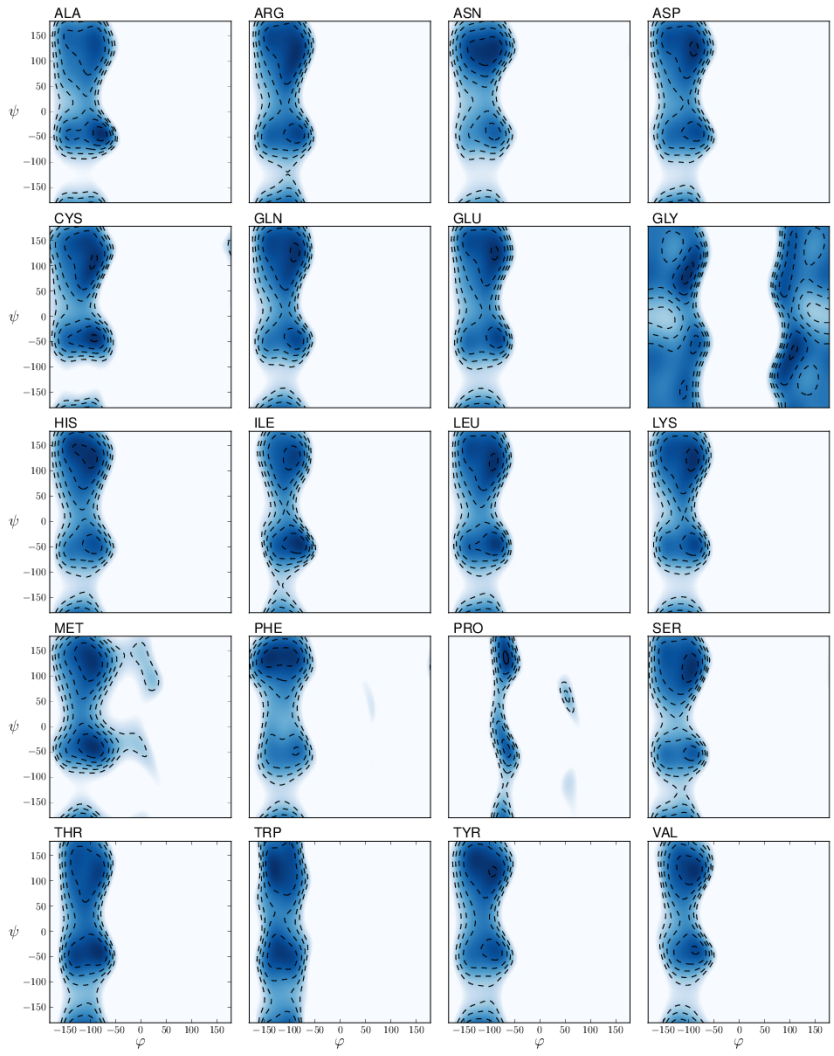


Figure 5.4. Backbone dihedral angle distributions implied by the physical energy. Heat maps of all φ/ψ distributions generated by a physical prior alone approximated by a maximum entropy distribution.

the Ramachandran potential and the Lennard-Jones potential for a ubiquitin ensemble in Figure 5.5A. To generate this plot, we ran a REMC simulation of ubiquitin controlled by the Lennard-Jones and Ramachandran potentials. We combined the samples from different replicas by the weighted histogram analysis method to take all available information into account and show the likelihood of encountering a sample with a given combination of energies. Depending on the energy range, the potentials can be positively and negatively correlated. For the high energy regime we see a positive correlation, as both potentials prefer the same areas of the Ramachandran plot. But as we get closer to the low energy structures, the different potentials cannot be reconciled and we observe a negative correlation. These observations stress the importance of a careful combination of different potentials. The difficulty of estimating w_{rama} stems from its being an ensemble average. Hence, we need to assess how well entire ensembles agree with the potential, rather than examine only a single structure. Since we have to deal with ensembles, we need to use computationally intensive simulations to estimate w_{rama}.

Recall our Bayesian approach to estimate the temperature of a potential based on the experimental (Chapter 4). To estimate w_{rama}, we need to look at the expected backbone energy $\langle E_{\mathrm{rama}} \rangle$, where $\langle \cdot \rangle$ denotes an ensemble average. To obtain this ensemble average, structures are sampled from the combined energy $w_{\mathrm{phys}} E_{\mathrm{phys}} + w_{\mathrm{rama}} E_{\mathrm{rama}}$. It should be noted that this simulation is not influenced by the experimental data and that the sampled structures are not necessarily close to the native conformation of the protein. The resulting average $\langle E_{\mathrm{rama}} \rangle_{\text{no data}}$ summarizes how the force field and the backbone potential are correlated for a particular setting of w_{phys} and w_{rama}. This value is contrasted with the expected backbone energy obtained with data $\langle E_{\mathrm{rama}} \rangle_{\text{data}}$. To calculate this ensemble average, structures are sampled based on the full energy $w_{\mathrm{data}} E_{\mathrm{data}} + w_{\mathrm{phys}} E_{\mathrm{phys}} + w_{\mathrm{rama}} E_{\mathrm{rama}}$ where the cost function E_{data} assesses the fit with the data. According to the principle of maximum entropy (Jaynes, 1957), the averages are equal at the optimal w_{rama}

$$\langle E_{\mathrm{rama}} \rangle_{\text{data}} = \langle E_{\mathrm{rama}} \rangle_{\text{no data}}. \tag{5.4}$$

In the previous chapter we devised these rules based on the maximization of the *model evidence*, $\Pr(D|w_{\mathrm{rama}})$, which is the probability of observing the data for a particular value of w_{rama} and whose computation involves an ensemble average. In essence, our procedure of finding w_{rama} involves Bayesian model comparison (MacKay, 2003) for a continuous family of models that differ in the weight of the backbone potential. As before, we have $\Pr(D|w_{\mathrm{rama}})$ as

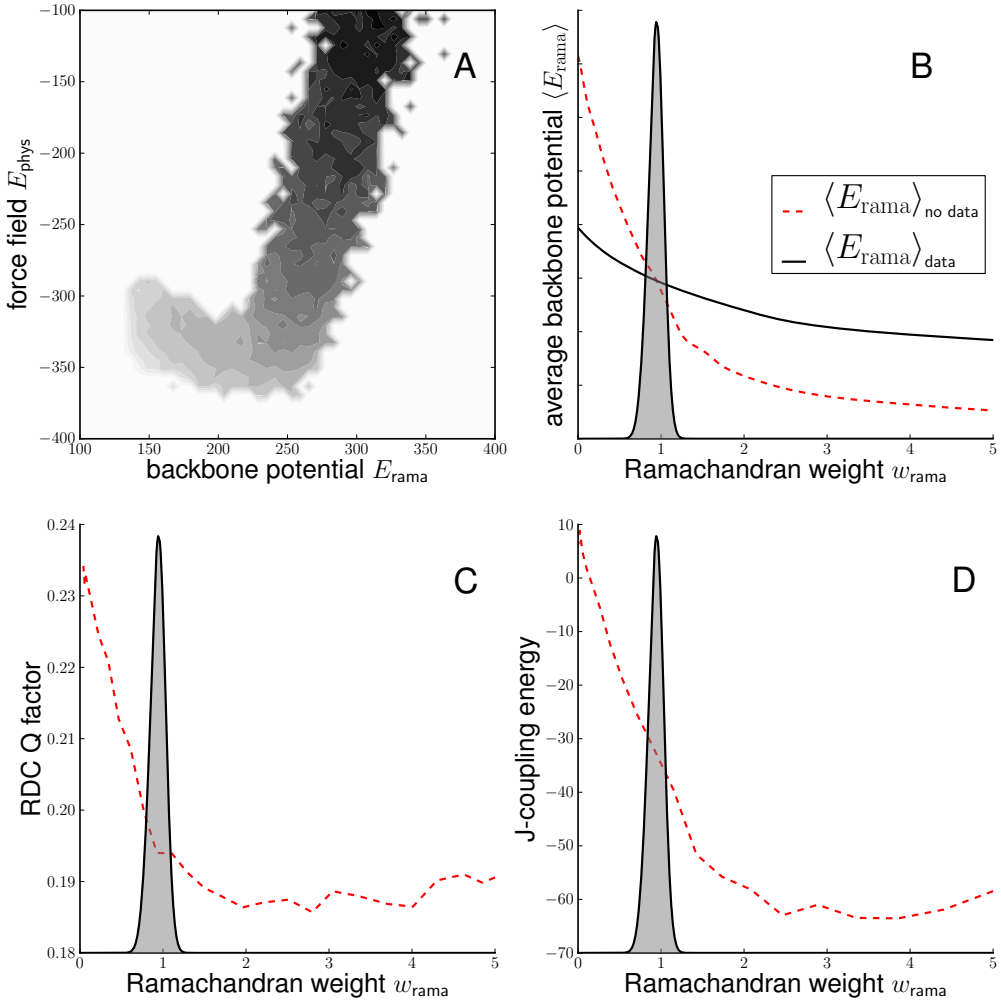


Figure 5.5. Bayesian weighting of the backbone potential for ubiquitin inferred from distance data. A: Correlation between backbone potential and non-bonded force field. Shown is the joint distribution of physics- and knowledge-based contributions in the absence of any structural data. B: Model evidence $\Pr(D|w_{\text{rama}})$ as a function of the Ramachandran weight w_{rama}. C: Influence of the Ramachandran weight on the average Q-factor (dashed line) calculated for 11 RDC data sets that were not used in the structure calculation. The Q-factor reflects the agreement between experimental and calculated RDCs. D: Influence of the Ramachandran weight on the fit with scalar coupling measurements (dashed line). Six three-bond scalar coupling data sets are available for ubiquitin and have not been used in the structure calculation. The grey distribution indicates the model evidence $\Pr(D|w_{\text{rama}})$.

$$\Pr(D|w_{\text{rama}}) = \int \Pr(D|\theta, \alpha, w_{\text{rama}})\ \Pr(\theta|w_{\text{rama}})\ \Pr(\alpha)\ \mathrm{d}\theta\, \mathrm{d}\alpha.$$

where $\Pr(\theta|w_{\text{rama}}) \propto \exp\{-w_{\text{phys}}E_{\text{phys}}(\theta) - w_{\text{rama}}E_{\text{rama}}(\theta)\}$ is the combined prior probability of conformation θ for a given weight w_{rama}. We can reduce the computation to a low-dimensional integral by using the density of states as described in Chapter 4

$$\begin{aligned} g_\lambda(E_{\text{rama}}) \quad &= \quad \int \delta(E_{\text{rama}} - E_{\text{rama}}(\theta))\, [\Pr(D|\theta, \alpha)]^\lambda \\ &\qquad \times \Pr(\alpha)\, e^{-w_{\text{phys}}E_{\text{phys}}(\theta)}\, \mathrm{d}\theta\, \mathrm{d}\alpha \end{aligned}$$

where $\delta(\cdot)$ denotes the Dirac delta function.

We can now express the model evidence as the ratio of two integrals.

$$\Pr(D|w_{\text{rama}}) = \frac{\int g_1(E_{\text{rama}}) e^{-w_{\text{rama}}E_{\text{rama}}}\ \mathrm{d}E_{\text{rama}}}{\int g_0(E_{\text{rama}}) e^{-w_{\text{rama}}E_{\text{rama}}}\ \mathrm{d}E_{\text{rama}}}$$

which requires two densities of states, $g_0(E_{\text{rama}})$ and $g_1(E_{\text{rama}})$, to describe how the backbone energy E_{rama} is distributed without and with data, respectively. Estimates of the density of states are obtained by applying multiple histogram reweighting (Ferrenberg and Swendsen, 1989; Habeck, 2012) as outlined in Chapter 4. The replica schedules were also adapted to account for the additional force field and provide adequate sampling.

5.4 Application to a single degree of freedom

We illustrate the ideas outlined in this article by revisiting the example we used to introduce the concepts of ISD in Chapter 2.4. Recall tbat the goal is to determine this angle on the basis of a simulated three-bond scalar coupling, J, between atoms HN-N-CA-HA (see Figure 2.6). Figure 5.6 illustrates the method for that simple toy system. The Karplus relation (Karplus, 1963) and a Gaussian error model (Habeck et al., 2005b) result in a trimodal likelihood function for the φ angle; one probability peak is located in the largely disallowed positive domain. Switching only the physical potential on ($w_{\text{rama}} = 0$) eliminates the positive peak but cannot resolve the ambiguity in the remaining two modes favouring the wrong solution at $\varphi = -144^{\text{o}}$. Incorporation of a knowledge-based dihedral angle potential helps to resolve this

ambiguity. But it is unclear how strongly the knowledge-based potential should contribute. If the weight is too small ($w_{\mathrm{rama}} = 0.7$), the ambiguity remains unresolved; if the weight is too large ($w_{\mathrm{rama}} = 7.0$), the posterior distribution peaks at an angle that is systematically shifted away from the correct value ($\varphi = -75^{\mathrm{o}}$). The optimal weight is determined by maximizing the overlap between the probability of the data and of the prior. This choice compromises the data as to the lowest extent possible, but still tries to utilize of the information contained in the knowledge-based contribution.

To find the optimal weight, we calculate the *marginal likelihood*, $\Pr(D|w_{\mathrm{rama}})$, which is the probability of observing the data for a particular value of w_{rama}. We then choose the weight that is most compatible with the given data by maximizing $\Pr(D|w_{\mathrm{rama}})$. This provides a data-driven way of controlling the influence of the knowledge-based potential function relative to the physical energy and the data. In essence, our procedure of finding w_{rama} is Bayesian model comparison (MacKay, 2003) for a continuous family of models that differ in their weight of the statistical potential.

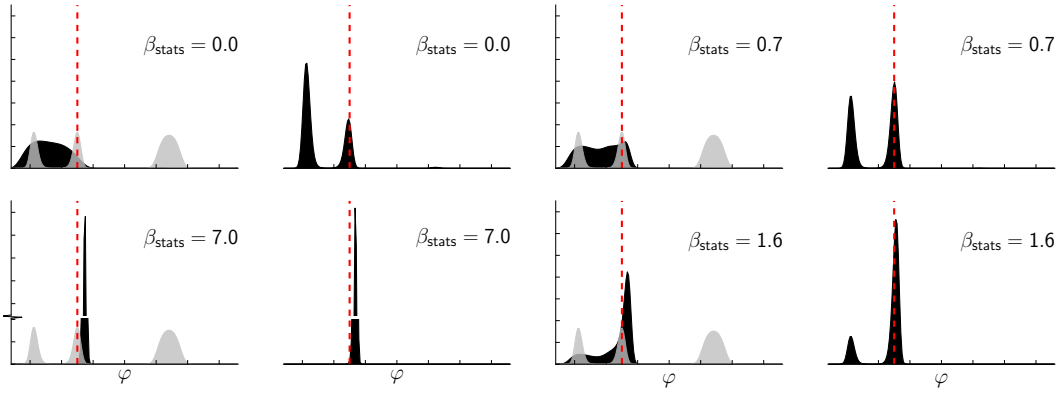


Figure 5.6. Example highlighting Bayesian weighting of the statistical potential. Shown is the effect of different choices of the weight of the statistical potential, w_{rama}, for a toy system with one angular degree of freedom φ: $w_{\mathrm{rama}} = 0.0$ (no statistical potential), $w_{\mathrm{rama}} = 0.7$ (too small weight), $w_{\mathrm{rama}} = 7.0$ (too large weight), and $w_{\mathrm{rama}} = 1.6$ (optimal weight obtained by Bayesian weighting). (For $w_{\mathrm{rama}} = 7.0$, the y-axis has been broken to show the full posterior distribution.) The left column shows the likelihood resulting from a scalar coupling measurement in grey and the combined prior in black. The right column shows the posterior probability (i.e. the product of the likelihood and the prior) in black. The correct angle is indicated with a red dashed line ($\varphi = -75^{\mathrm{o}}$).

5.5 Bayesian weighting with high-quality data

We investigate the effect of Bayesian weighting on the example of the ubiquitin data introduced in the previous chapter (PDB code 1D3Z), comprising 1,444 non-redundant distance restraints. The available scalar dipolar couplings are not included in the calculation. We used the method outlined above to calculate the model evidence $\Pr(D|w_{\mathrm{rama}})$, as shown in Figure 5.5B. The curve shows a clear peak at 0.94 ± 0.09. At this weighting, the ensemble averages of prior energies are equal regardless of whether we include the data or not, thereby satisfying Equation 5.4. The RMSD to the crystal structure is largely unaffected by the choice of β for this high-quality data set as seen in Figure 5.7A. This is not completely surprising, as the fold is well defined by the restraints. The residual dipolar couplings and scalar coupling measurements that were not used as restraints are more sensitive to the local structure and are used to validate our approach. To assess the agreement with the 11 available RDC sets, we calculated the Q-factor for ensembles of 100 structures obtained from the replica sampling at different values of w_{rama}. The ensemble averages of the Q-factor in Figure 5.5C show that the Ramachandran weight, which optimizes the marginal probability, almost achieves the minimum Q-factor (0.19). FThe picture is slightly different for the scalar couplings; although the Bayesian choice improves the agreement with the observed RDCs, a setting of $\beta > 2.$ would have minimized the RDC energy.

5.6 Bayesian weighting with incomplete data

One focus of this thesis is the calculation of structures from sparse data. The completeness of the ubiquitin dataset can be reduced to observe the influence of Bayesian weighting on sparse data. We achieve this effect by using a completeness parameter λ that decreases the influence of the data. This parameter was introduced by Habeck (2011) to investigate structure calculation from a statistical mechanics point of view. For example, if we set $\lambda < 1$, the effective number of observations is reduced; for $\lambda = 1$ the original dataset is recovered. Based on our observations in the previous chapter, we expect the marginal likelihood to favour smaller values of β as the number of effective observations is reduced. Figure 5.7 shows the marginal likelihoods for $\lambda = 1., 0.1, 0.01$ and 0.005 with $w_{\mathrm{rama}} = 0.94 \pm 0.09$,

$w_{\mathrm{rama}} = 1.00 \pm 0.09$, $w_{\mathrm{rama}} = 0.92 \pm 0.1$, and $w_{\mathrm{rama}} = 0.84 \pm 0.15$ respectively. For lower values of λ, the posterior is no longer able to identify the native ensemble. The Ramachandran potential allows us to decrease further the number of effective observations per residue that still results in a native ensemble (Habeck, 2011) to less then 0.1 distances per residue. However, this value should be taken with a grain of salt it since only decreases the influence of the distances, and not their actual number. We also analysed the effect of an optimal w_{rama} on the RMSD of the ensemble to the crystal structure in Figure 5.7. The Bayesian choice of w_{rama}, which maximizes the marginal likelihood, is constantly located in the region of minimal RMSD. Especially at higher values of w_{rama} and increasing sparsity, we observed a detrimental effect of the Ramachandran potential on the average RMSD.

5.7 Impact on structure ensembles from sparse and noisy NMR data

So far, we have studied how Bayesian weighting of the backbone potential impacts the conformational ensemble under artificially sparsified data. The sparse data set of the Fyn-SH3 domain introduced before (Mal et al., 1998; Rieping et al., 2005) and a set of noisy distance bounds measured by solid-state NMR of the α-spectrin SH3 domain (Castellani et al., 2002) provide challenging real-world applications. We show the ensemble average of the backbone energy depending on the choice of w_{rama} and the resulting marginal likelihood for both datasets in Figure 5.8. The marginal likelihood peaks at $w_{\mathrm{rama}} = 0.76 \pm 0.11$ and $w_{\mathrm{rama}} = 0.3 \pm 0.13$ for the sparse and noisy distances, respectively.

We assess the accuracy of the ensembles by the RMSD to the crystal structure. The results shown in Figure 5.9 demonstrate that the maximum of the marginal likelihood is located in regions of low RMSD. Moreover, the RMSD distributions of the ensembles with optimal w_{rama} show less variance and no additional minima. We also observe that setting $w_{\mathrm{rama}} > 2.0$ leads to less accurate structures. Furthermore, the average structure of each ensemble is even more accurate than the individual structures (see Table 5.1). This indicates that the structure ensembles are better defined when using the backbone potential.

Figure 5.10 shows the structural ensembles of the sparse SH3 data set at dif-

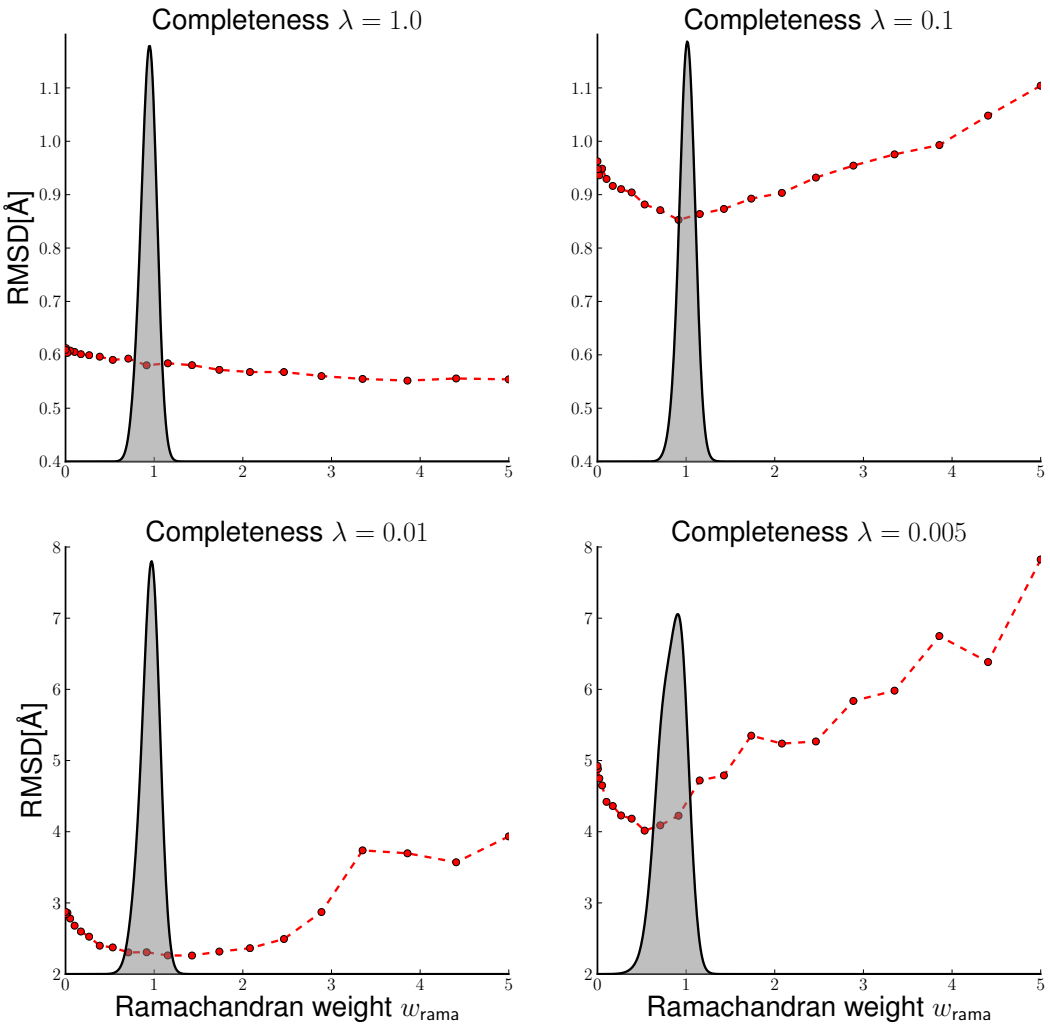


Figure 5.7. Impact of incomplete ubiquitin data on w_{rama}. Shown is the model evidence as a function of w_{rama} (grey) and the average RMSD (dots). The sparsity increases from the top left panel to the bottom right panel.

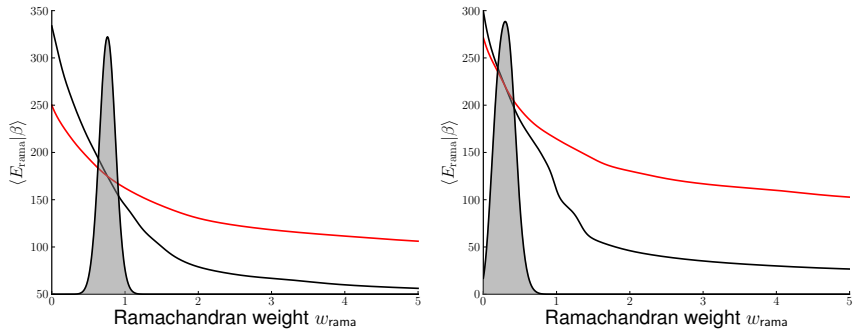


Figure 5.8. Bayesian weighting with sparse and noisy NMR data. Shown is the average backbone energy $\langle E_{\mathrm{rama}}\rangle$ with (black) and without data (red) for the Fyn-SH3 (left) and α-spectrin SH3 domain (right). The model evidence peaks where the two curves cross.

ferent values of w_{rama}. A small weight leads an ensemble that has a high variability, while the introduction of the backbone potential with an optimal weight results in an regular and accurate ensemble. If we keep increasing w_{rama}, we start to introduce additional helical regions, that disrupt the structures.

Our observations are supported by the Ramachandran plots as well. For $w_{\mathrm{rama}} = 5$, the Ramachandran plot becomes artificially narrow and peaks in the helical region. But the Ramachandran plot of the optimal w_{rama} is very close to that of the crystal structure.

Often, energy functions are assessed on how well they correlate with the RMSD. We plot the negative log posterior probability, comprising the Lennard-Jones potential, the Ramachandran potential and a data-dependent term against the RMSD for values of w_{rama}. The funnel guiding the simulation towards the native ensemble is much more pronounced with an optimally weighted Ramachandran potential. We even observe that RMSD and energy become anti-correlated at a large value of w_{rama}

5.8 Impact on structure quality

Besides the accuracy, we also analysed the effected of the marginal likelihood maximization on the quality of the different structures. We assessed the quality using Procheck (Laskowski et al., 1993) and WhatCheck(Vriend, 1990), and show the values of several validation criteria in Figure 5.12. A complete assessment by

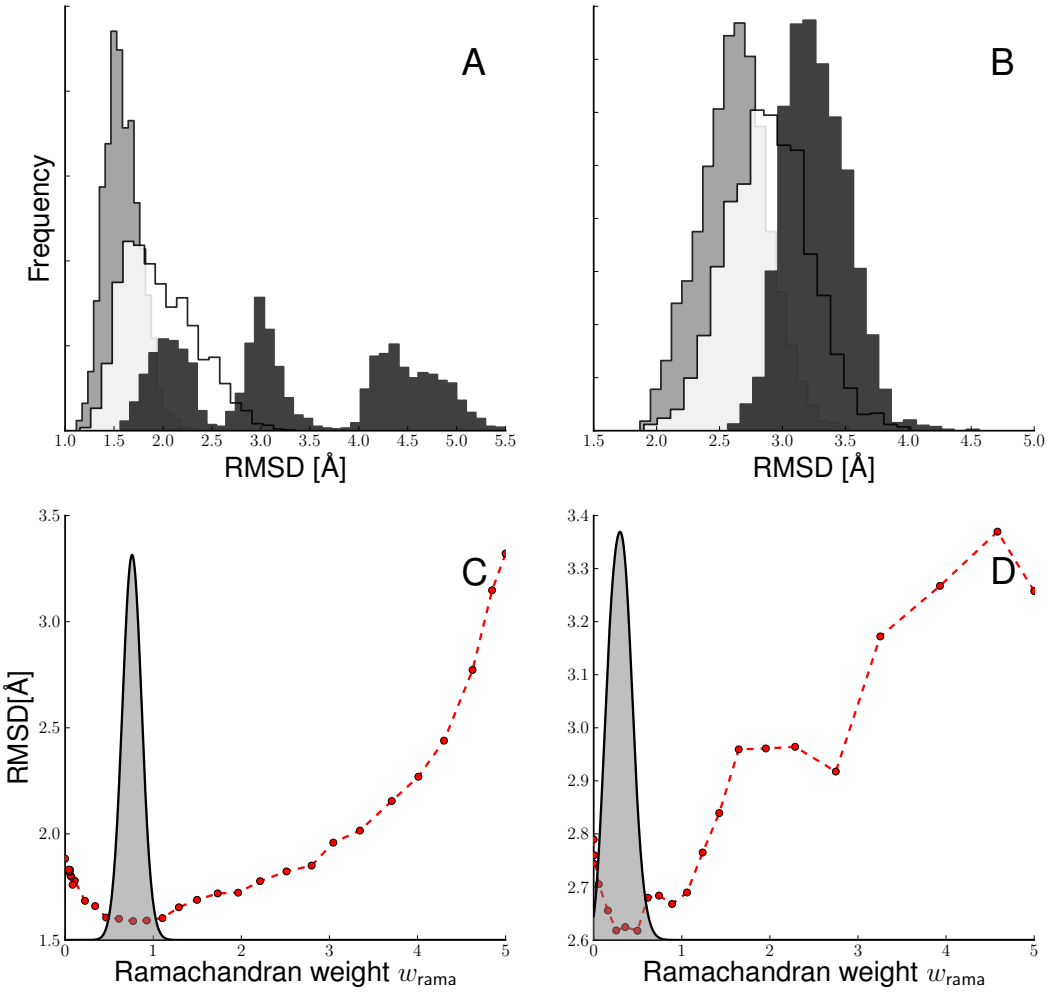


Figure 5.9. Impact on structure ensembles from sparse and noisy NMR data. Panels A, C show the results for the sparse Fyn-SH3 data set. Panels B, C show the results for the solid-state data. The top row displays the RMSD distributions with $w_{\text{rama}} = 0$ (white), $w_{\text{rama}} = 5$ (black) and optimal w_{rama} (grey). The grey distribution shown in the bottom panels is the model evidence as a function of the weight w_{rama}.

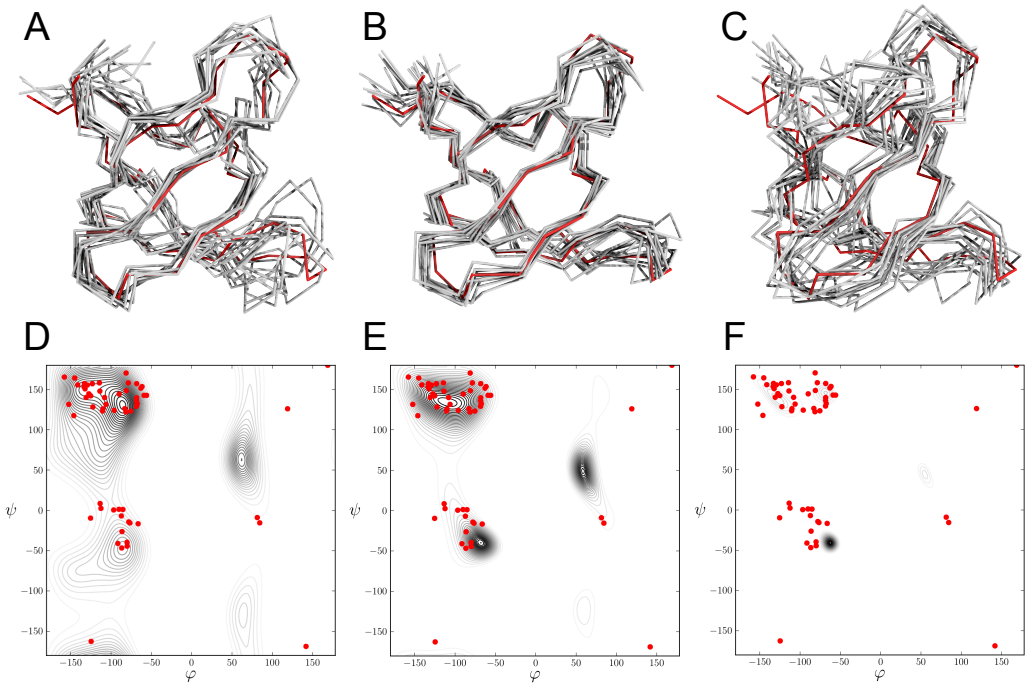


Figure 5.10. Influence of the weight w_{rama} on the structural ensemble of Fyn-SH3 inferred with sparse NMR data. Shown are the conformations and backbone dihedral distributions generated with different w_{rama}. Panels A-C display structure ensembles comprising ten randomly selected conformations (grey) superimposed onto the crystal structure (red). Panels D-F show in black a maximum entropy distribution fitted to the backbone torsion angles of the structures generated with ISD. The backbone dihedral angles of the crystal structure are marked by red dots. Panels A and D show the results for $w_{\mathrm{rama}} = 0.0$, panels B, E: $w_{\mathrm{rama}} = 0.76$ (optimal weight), panels C and F show $w_{\mathrm{rama}} = 5.0$ (maximum weight probed during replica-exchange simulations).

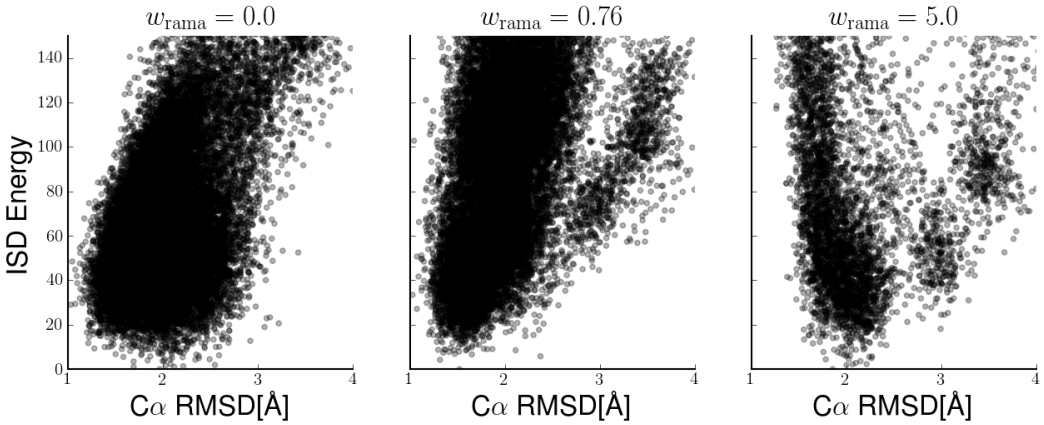


Figure 5.11. Energy funnels obtained with the sparse Fyn-SH3 data at different Ramachandran weights (left: $w_{\mathrm{rama}} = 0.0$, middle: $w_{\mathrm{rama}} = 0.76$, right: $w_{\mathrm{rama}} = 5.0$). The full ISD energy (negative log-posterior probability) is plotted against the RMSD to the Fyn-SH3 crystal structure.

Procheck and WhatCheck can be found in Table 5.1. The values are averages of 100 structures per ensemble. All ensembles, even the high-quality ubiquitin data, are improved though the addition of the Ramachandran potential. The ensemble favoured by the marginal likelihood represents a compromise between different validation scores. While the Ramachandran score (RAMCHK) continuously rises, others decrease with increasing w_{rama} . In general, however, no score is an accurate indicator of high accuracy structures; none correlates with the RMSD.

We also examine the goodness of fit with the data. The Ramachandran potential could introduce bias into the calculation process that increases the model error. Figure 5.13 shows that the estimated model parameters of the data (Habeck et al., 2006) are largely unaffected if we incorporate the knowledge-based contribution. The Bayesian choice of w_{rama} does not increase the model error. At higher values of w_{rama}, the variance and average model increases due to the bias introduced through the Ramachandran potential.

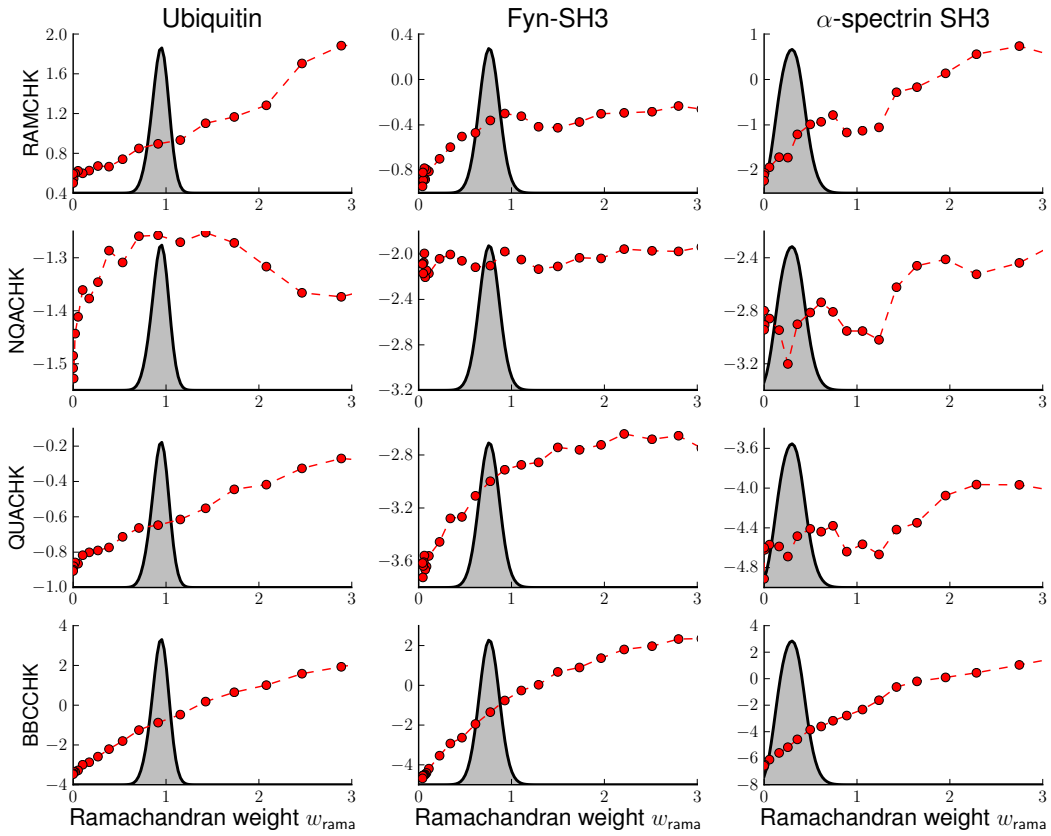


Figure 5.12. Influence of the Ramachandran weight on various quality criteria. Shown is the impact of w_{rama} on WhatCheck validation criteria. Each column reports the results for a different data set (left column: ubiquitin, middle column: Fyn-SH3 domain, right column: α-spectrin SH3 domain). Each row shows the evolution of a quality score with increasing w_{rama} (each dot marks the average of over 100 structures that were randomly selected from the ISD ensemble, dashed lines are added to guide the eye). The first row reports the Ramachandran appearance as assessed by RAMCHK. The second and third rows show WhatCheck's packing scores. The last row reports the regularity of the backbone (BBCCK). The grey distribution indicates the model evidence $\Pr(D|w_{\text{rama}})$ as a function of w_{rama}.

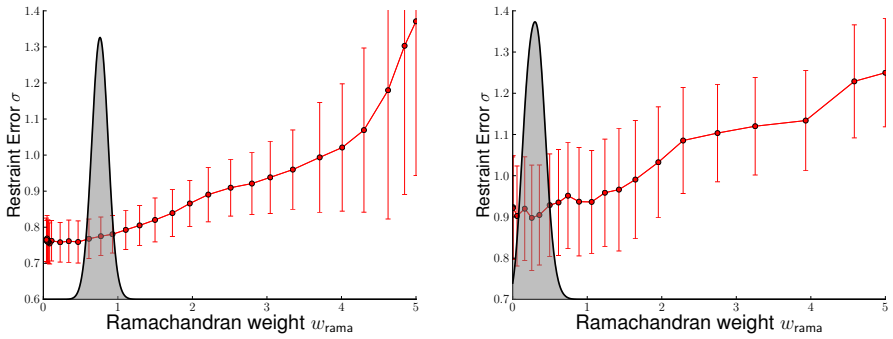


Figure 5.13. Impact on the estimation of model parameters. The red dots and error bars indicate the estimated error σ of the distance restraints for the sparse distance data (left) and the noisy restraints measured with solid-state NMR (right). The model evidence is shown as grey distribution.

Table 5.1. Quality and accuracy of posterior ensembles. The table shows various validation scores for ubiquitin, Fyn-SH3 and α-spectrin SH3. PDB entries 1ubq, 1shf and 1shg serve as reference crystal structures. The reported values are averaged over 100 conformations generated with ISD. In every sub-table, the first four rows show the Ramachandran appearance (as percentages) performed with Procheck. The next four rows list the WhatCheck Z-scores assessing the quality of packing, the dihedral angle statistics and the regularity of the backbone. The accuracy of the ensemble is measured as Cα RMSD between the crystal structure and individual ensemble members or the ensemble mean.

	$w_{\text{rama}} = 0.0$	$w_{\text{rama}} = w_{\text{opt}}$	$w_{\text{rama}} = 5.0$	**X-ray**
ubiquitin (1d3z)				
Procheck				
Core	79.71 ± 3.78	90.38 ± 2.84	97.53 ± 1.21	95.5
Allowed	19.21 ± 3.82	9.56 ± 2.82	2.48 ± 1.21	4.5
Generous	1.00 ± 1.10	0.06 ± 0.29	0.00 ± 0.00	0.0
Disallowed	0.06 ± 0.29	0.00 ± 0.00	0.00 ± 0.00	0.0
WhatCheck				
QUACHK	-0.90 ± 0.25	-0.66 ± 0.21	-0.10 ± 0.20	1.19
NQACHK	-1.51 ± 0.34	-1.26 ± 0.34	-1.33 ± 0.24	-1.66
RAMCHK	0.59 ± 0.46	0.85 ± 0.41	1.88 ± 0.27	0.85
BBCCHK	-3.43 ± 0.57	-1.25 ± 0.70	3.16 ± 0.39	1.91

Table5.1 – continued from previous page

	$w_{\mathrm{rama}} = 0.0$	$w_{\mathrm{rama}} = w_{\mathrm{opt}}$	$w_{\mathrm{rama}} = 5.0$	**X-ray**
Accuracy				
RMSD [Å]	0.62 ± 0.06	0.58 ± 0.06	0.56 ± 0.04	–
RMSD (mean) [Å]	0.50	0.43	0.46	–
Fyn-SH3 (1zbj)				
Procheck				
Core	70.17 ± 5.36	85.22 ± 3.67	97.60 ± 1.74	98.0
Allowed	28.75 ± 5.37	14.64 ± 3.61	2.32 ± 1.64	2.0
Generous	1.07 ± 1.20	0.14 ± 0.51	0.06 ± 0.34	0.0
Disallowed	0.02 ± 0.20	0.00 ± 0.00	0.02 ± 0.20	0.0
WhatCheck				
QUACHK	-3.28 ± 0.45	-3.11 ± 0.39	-3.65 ± 0.66	-0.81
NQACHK	-2.00 ± 0.54	-2.12 ± 0.55	-3.08 ± 0.83	2.52
RAMCHK	-0.60 ± 0.43	-0.47 ± 0.39	0.34 ± 0.55	0.22
BBCCHK	-2.93 ± 0.73	-1.95 ± 0.73	2.33 ± 0.68	-0.91
Accuracy				
RMSD [Å]	1.97 ± 0.28	1.59 ± 0.18	3.33 ± 1.08	–
RMSD (mean) [Å]	1.39	1.05	2.53	–

Table5.1 – continued from previous page

	$w_{\mathrm{rama}} = 0.0$	$w_{\mathrm{rama}} = w_{\mathrm{opt}}$	$w_{\mathrm{rama}} = 5.0$	**X-ray**
α-spectrin SH3 (1m8m)				
Procheck				
Core	32.70 ± 5.79	64.66 ± 5.43	97.78 ± 1.44	95.5
Allowed	43.16 ± 6.90	33.28 ± 5.55	2.22 ± 1.44	4.5
Generous	15.84 ± 4.48	2.02 ± 1.56	0.00 ± 0.00	0.0
Disallowed	8.30 ± 3.84	0.04 ± 0.28	0.00 ± 0.00	0.0
WhatCheck				
QUACHK	-4.60 ± 0.46	-4.59 ± 0.39	-3.50 ± 0.35	-0.44
NQACHK	-2.80 ± 0.51	-2.95 ± 0.57	-3.00 ± 0.40	2.83
RAMCHK	-2.11 ± 0.42	-1.71 ± 0.45	-0.06 ± 0.44	-0.90
BBCCHK	-6.63 ± 0.50	-5.60 ± 0.58	1.96 ± 0.60	-0.14
Accuracy				
RMSD [Å]	2.84 ± 0.29	2.63 ± 0.24	3.27 ± 0.25	–
RMSD (mean) [Å]	2.51	2.16	3.25	–

5.9 Conclusion

In this chapter, we introduced a new method that enables the combination of different potential functions in structure calculation. Our approach builds upon the Bayesian principles and is an application of the algorithm introduced in the previous chapter. The Bayesian formalism finds an optimal combination of the different potentials on the basis of the experimental data, which results in more accurate structures and does not introduce model bias. An optimal weighted Ramachandran potential can significantly improve the quality of the calculated ensembles as well as the similarity to the crystal structure. These advantages are especially important when dealing with noisy and sparse data, where the positive effects are more pronounced. We found that no universal weight exists that could be applied to all

data sets. Instead it is advisable to estimate the weight of a potential in the course of the structure calculation.

In the future, we plan to extend our method to weight multiple statistical energy terms simultaneously in the course of a structure calculation. The final goal is to design an efficient and unbiased but highly expressive conformational prior distribution that allows the calculation of high-quality ensembles from very sparse data sets.

6

Estimating energy functions from Boltzmann ensembles

In the following, we focus on the estimation of new potential functions. In the previous chapter, we studied the influence of potential functions on the structure calculation process, and found that the addition of a single potential significantly improved structure calculation. Furthermore, we observed that combining different potential functions does not necessarily lead to improved accuracy.

The latter point motivated us to investigate new techniques to estimate potentials, whereby physical interactions are used to define a model, but then the model parameters are estimated from a database of known structures. More formally, given a set of three-dimensional structures, the task is to find the most likely potential function that gave rise to the observed structures. In statistical physics, this problem is also known as the inverse problem of statistical mechanics. The solution to this inverse problem is involved as frequently repeated evaluations of the partition function are needed, where a single evaluations is often intricate. We introduce an extension of the configurational temperature that allows us to infer a parameterized approximation of the potential function. Parts of this chapter are published in Mechelke and Habeck (2013a).

6.1 Introduction

As outlined in the previous chapters, potential functions are integral to structure calculation from NMR data. But this hardly the only application. Potential functions are at the heart of Molecular dynamics and Monte Carlo simulations that allow scientists insights into biomolecular systems at a level of detail unmatched by experimental methods (Lane et al., 2012). But already small inaccuracies in the potential function can effect the results and bias simulations towards incorrect conformations (Freddolino et al., 2009; Wroblewska and Skolnick, 2007; Das, 2011).

The construction of new force fields is an art in itself; the final force field comprises several different potential functions with hundreds of interdependent parameters. Each of the parameters needs to be carefully tuned to arrive at a good approximation of the physical reality. Some of the difficulties are related to the tight coupling of the individual force field terms (Wang and Wade, 2006; Best and Hummer, 2009). Often, a change in one parameter makes a readjustment necessary in several other.

The goal of force field development is to reproduce experimental measurements like vibrational frequencies and spectra of small molecules (Mackerell, 2004). Quantum mechanical calculations are used to complement the experimental information, when needed. Regardless of whether quantum mechanical calculations or experimental measurements are used, there is a discrepancy between the size of the system on which the force field is estimated and the size of the proteins to which the force field is later applied. This discrepancy has sparked a debate on whether molecular dynamics force fields are really universally transferable (Feig, 2008). Only recently has it become possible to compare force fields in terms of their ability to fold a protein chain to its native state (Lange et al., 2010; Lindorff-Larsen et al., 2012).

Statistical potentials follow a different strategy, as they aim to exploit the wealth of information contained in the thousands of experimentally determined protein structures (Tanaka and Scheraga, 1976; Miyazawa and Jernigan, 1985; Muñoz and Serrano, 1994). Although the underlying idea is sound, most approaches make the assumption that individual potential functions are independent. From what we have learned in the previous chapters this assumption is not realistic. If a Ramachandran and Lennard-Jones potential show high correlation, it would be naive to assume that the Lennard-Jones terms between different atom types are not entangled. Nonetheless, many current approaches rely on this assumption Hamelryck (2009b),

they collect database statistics for certain geometric descriptors, such as distances and dihedral angles, and compute a statistical potential by direct inversion of the histogram. But the resulting energy functions are, at best, potentials of mean force and generally differ from the potential energy function.

Rather than to compute statistical potentials, we aim to extract force field parameters directly from configurations drawn from the canonical ensemble:

$$p(\boldsymbol{x}|\beta) = \frac{1}{Z(\beta)} \exp\left\{-\beta E(\boldsymbol{x})\right\} \tag{6.1}$$

where $\boldsymbol{x}$ is the configuration of the system, $E(\boldsymbol{x})$ is the system's potential energy, β the inverse temperature and $Z(\beta)$ the partition function. The problem of estimating $E(\boldsymbol{x})$ from a set of configurations is also known as the inverse problem of statistical mechanics. Statistically speaking, we try to estimate a parametric distribution $p_{\text{approx}}(\boldsymbol{x}|\lambda)$ that approximates the probability density $p(\boldsymbol{x}|\beta)$. Although density estimation is an extensively problem, the partition functions $Z(\beta)$ and $Z(\lambda)$ cannot be computed analytically and a numerical approximation is often intractable for this class of problems.

Previous approaches to solve the inverse problem of statistical mechanics encompass iterative Boltzmann inversion (Reith et al., 2003), Force Matching (Izvekov et al., 2004a,b) and reverse Monte Carlo (McGreevy and Pusztai, 1988; Soper, 1996; Lyubartsev and Laaksonen, 1995; Savelyev and Papoian, 2010) techniques. These methods do not work with $p(\boldsymbol{x}|\beta)$ but wth related quantities like the distribution of pairwise distances or similar correlation functions. They differ in terms of which distribution is used and how the parameters are changed depending of the differences. Thus, in every iteration, an ensemble simulation is necessary to compute the momentary distribution functions. All methods can be seen as special cases of the relative entropy formalism (Shell, 2008) as they minimize an entropic divergence between the true and estimated potential function. The generalized Yvon-Born-Green (gYBG) method developed by Mullinax and Noid (2009) provides a powerful alternative. It builds upon the works of Yvon (1935) and Born and Green (1946) for monatomic fluids, and generalizes the theory to arbitrary potential functions. One advantage of the gYBG method is that it calculates the potential energy directly from structural correlation functions and does not involve repeated ensemble simulations. A drawback of gYBG is that this method needs a large volume of samples to arrive at accurate estimates.

In this chapter, we extend the configurational temperature formalism (Rugh, 1997; Jepps et al., 2000) to derive a potential energy function. We treat the force

constants and other force field parameters as generalized temperatures and derive a system of linear equations to estimate them from a set of molecular configurations. Furthermore, we study different variants of the configurational temperature equations and discuss means to improve the transferability of the estimated potentials. We demonstrate the efficacy and accuracy of the approach for simple systems such as Lennard-Jones fluids as well as a coarse-grained protein model and a non-bonded knowledge-based potential

6.2 Configurational temperature

Rugh (1997) introduced the configurational temperature to express the thermodynamic temperature of a system. He showed that this measure is related to the curvature of the energy surface. This means that we can estimate the temperature of a system without knowledge of the kinetic energy. Jepps et al. (2000) and Rickayzen and Powles (2001) extended this proof to yield a new expression for the temperature of a system based on configurations only.

They defined the temperature in terms of a vector field $B(\Gamma)$:

$$\langle B(\Gamma) \cdot \nabla_\Gamma h(\Gamma) \rangle_\Gamma = \langle \nabla_\Gamma \cdot B(\Gamma) \rangle_\Gamma \tag{6.2}$$

where Γ and ∇_Γ denote the vector of phase space variables and their derivatives. h refers to the reduced Hamiltonian, $h(\Gamma) = -\log p(\Gamma) + \text{const.}$. This expression for the temperature (Equation 6.2) is valid for both, the micro- and macrocanonical ensembles. The requirements for the vector field are fairly relaxed: it has to hold that $0 \leq \|\nabla_\Gamma h(\Gamma) B(\Gamma)\| \leq \infty$ and $0 \leq \|\nabla_\Gamma B(\Gamma)\| \leq \infty$.

If we restrict B to vector fields that do not depend on the momenta and are zero in all directions of the momenta, we arrive at a general formula of Jepps et al. (2000) for the configurational temperature,

$$\langle B \cdot \nabla v \rangle = \langle \nabla \cdot B \rangle \tag{6.3}$$

where $v(\boldsymbol{x})$ is the reduced potential energy, and $B(\boldsymbol{x})$ is that vector field that only depends on the configuration and for which the above restrictions apply. The operator $\boldsymbol{\nabla}$ is the gradient with respect to $\boldsymbol{x}$ and $\boldsymbol{\nabla} \cdot B$ is the divergence of the vector field. The brackets $\langle \cdot \rangle$ denote the configuration average:

$$\langle f \rangle = \frac{1}{Z} \int f(\boldsymbol{x})\, e^{-v(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x} \quad \text{where} \quad Z = \int e^{-v(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x}.$$

If we choose $v(\boldsymbol{x}) = \beta E(\boldsymbol{x})$ with $\beta = (k_B T)^{-1}$ we arrive at the configurational temperature:(Jepps et al., 2000)

$$\beta = \frac{\langle \boldsymbol{\nabla} \cdot \boldsymbol{B} \rangle}{\langle \boldsymbol{B} \cdot \boldsymbol{\nabla} E \rangle}. \tag{6.4}$$

Now, the only choice left is the vector field $\boldsymbol{B}$. The typical choice is $\boldsymbol{B}(\boldsymbol{x}) = \boldsymbol{\nabla} E(\boldsymbol{x})$, which gives the temperature as an average of the configurations independent of the kinetic energy Jepps et al. (2000)

$$\frac{1}{k_B T_{\text{config}}} = \frac{\langle \boldsymbol{\nabla} \cdot \boldsymbol{\nabla} E \rangle_{\boldsymbol{x}}}{\langle \boldsymbol{\nabla} E \cdot \boldsymbol{\nabla} E \rangle_{\boldsymbol{x}}}.$$

6.3 Estimation of interaction potentials

We now generalize the configurational temperature to estimate potentials. For this we assume that the configurations follow a Boltzmann distribution:

$$p(\boldsymbol{x}|\lambda) = \frac{1}{Z(\lambda)} \exp\left\{-v(f(\boldsymbol{x});\lambda) - f_0(\boldsymbol{x})\right\}, \quad Z(\lambda) = \int \exp\left\{-v(f(\boldsymbol{x});\lambda) - f_0(\boldsymbol{x})\right\}\, \mathrm{d}\boldsymbol{x} \tag{6.5}$$

with a interaction potential v. We assume that v depends on a configuration $\boldsymbol{x}$ through K features $f_k, k = 1, \ldots, K$, for example, angles or pairwise distances. The parameter λ controls the shape of the potential and $Z(\lambda)$ is the partition function or normalization constant. The reference potential $f_0(\boldsymbol{x})$ is an arbitrary energy that we assume to be known; for example, we could let $f_0(\boldsymbol{x})$ be constant, or $f_0(\boldsymbol{x})$ could be a reference distribution that already accounts for some interactions. Further, we assume that the potential function v is a linear combination of the parameters: $v(f(\boldsymbol{x});\lambda) = \lambda^T f(\boldsymbol{x}) = \sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{x})$. Although this choice seems restrictive at first, it still allows us to include all of the commonly used force field terms. The canonical ensemble (6.1) is a special case of this formulation with a single feature

f that comprises the force field and a scalar λ that represents the inverse temperature.

For this special case, the connection to the configurational temperature formalism is clear. For a vectorial f_k and λ_k Equation 6.3evaluates to the identity.

$$\langle B \cdot \nabla(\lambda^T f + f_0)\rangle_\lambda = \langle \nabla \cdot B\rangle_\lambda$$

where we use the angle brackets $\langle\cdot\rangle_\lambda$ to denote the ensemble average (6.5) and the subscript to indicate the dependence on λ. Again, for the choice of vector field $B(\boldsymbol{x})$ the same weak conditions apply (Jepps et al., 2000). We choose a series of vector fields B_k, one for each expansion coefficient λ_k, to obtain

$$\sum_{l=1}^{K} \lambda_l \langle B_k \cdot \nabla(f_l + f_0)\rangle_\lambda = \langle \nabla \cdot B_k\rangle_\lambda, \quad k = 1, \ldots, K \tag{6.6}$$

This system of linear equations is a multi-temperature generalization of the configurational temperature relation (6.4) and determines the parameters λ by solving:

$$\sum_{l=1}^{K} A_{kl}\lambda_l = b_k \quad \text{with} \quad A_{kl} = \langle B_k \cdot \nabla f_l\rangle_\lambda,\ b_k = \langle \nabla \cdot B_k - B_k \cdot \nabla f_0\rangle_\lambda. \tag{6.7}$$

We propose thee vector fields B_k to estimate potentials:

$$B_k(x) = \nabla f_k(x)/\|\nabla f_k(x)\|^n, \quad n = 0, 1, 2 \tag{6.8}$$

where $\|\cdot\|$ indicates the Euclidean norm of configuration space vectors. In order to evaluate the configurational temperature equations (6.7), we need to compute the divergence of B_k:

$$\nabla \cdot B_k(x) = \frac{\Delta f_k(x)}{\|f_k(x)\|^n} - n\,\frac{\sum_{i,j}(\partial_i f_k(x))\,(\partial_i\partial_j f_k(x))\,(\partial_j f_k(x))}{\|\nabla f_k(x)\|^{n+2}} \tag{6.9}$$

where indices i, j enumerate all configurational degrees of freedom and ∂_i indicates a partial derivative along the ith coordinate (i.e. $\partial_i\partial_j f_k(x)$ is the Hessian matrix of the kth feature); Δ is the Laplace operator. We turn the above equations into estimators of the force field parameters λ from sampled configurations $x^{(t)}$, by replacing the ensemble averages in equation (6.7) with sample averages:

$$A_{kl} \approx \frac{1}{T}\sum_{t=1}^{T} B_k(\boldsymbol{x}^{(t)}) \cdot \boldsymbol{\nabla} f_l(\boldsymbol{x}^{(t)}), \quad b_k \approx \frac{1}{T}\sum_{t=1}^{T} \nabla \cdot B_k(\boldsymbol{x}^{(t)}) - B_k(\boldsymbol{x}^{(t)}) \cdot \boldsymbol{\nabla} f_0(\boldsymbol{x}^{(t)} \tag{6.10}$$

If we chose the vector field $B_k = \nabla f_k$, we find similarities to other methods in statistical physics and machine learning. Equation 6.10 can also be derived from score matching Hyvärinen (2005). Score matching is a method that allows one to estimate continuous probability densities with intractable normalization constants. For the same choice of B_k, the configurational temperature equation (6.7) can be derived from the gYBG equations ((Mullinax and Noid, 2009, 2010).

Lennard-Jones fluid

We demonstrate the effectiveness of our method on well understood systems. The first model system is a monatomic Lennard-Jones fluid. The simulation was carried out using the Molecular Modelling Toolkit (Hinsen, 2000). The system comprises 864 argon atoms that were simulated in an NVT ensemble at 86.0 K with periodic boundary conditions. After equilibration, we simulated the system for 2000 timesteps of 10 fs each. For the analysis, we considered every 100th sample to minimize the correlation between the samples. Recall that all interactions within this system are described by a Lennard-Jones potential:

$$E_{\text{L-J}}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{6.11}$$

where σ determines the location of the minimum and ϵ its depth. Thus, the total potential energy is $E(\boldsymbol{x}) = \sum_{i<j} E_{\text{L-J}}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$ where $\boldsymbol{x}_i$ is position of the ith atom. We aim to estimate the parameters ϵ and σ. Using some simple algebra we rewrite the potential 6.11 and use r^{-6} and r^{-12} as basis functions f_k

$$f_1(\boldsymbol{x}) = \sum_{i<j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^{-6} \quad \text{and} \quad f_2(\boldsymbol{x}) = \sum_{i<j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^{-12} \tag{6.12}$$

with the corresponding parameters:

$$\lambda_1 = -4\beta\epsilon\sigma^6 \quad \text{and} \quad \lambda_2 = 4\beta\epsilon\sigma^{12}. \tag{6.13}$$

From λ_1, λ_2 and the known temperature T, we can compute ϵ and σ.

We estimated the parameters ϵ and σ from 20 configurations using the outlined approach. Furthermore, we used each of the configurations to get an estimate of ϵ and σ from a single configuration In Figure 6.1A we illustrate the estimated potentials. The different choices for the vector fields B_k (Equation 6.8) result in very similar estimates. However, for reasons of clarity, we only show the results for $B_k = \nabla f_k$. The estimate obtained from a single configuration of the system already provides a good approximation of the potential.

In this experiment we used the same features for simulation and estimation. For practical purposes it is unrealistic to assume knowledge of the "correct" set of basis functions. As solution we propose using a weighted sum of Laguerre polynomials. These functions form an orthonormal system and constitute a basis for all functions on the non-negative axis. The Laguerre polynomials $L_k(r)$ are defined as

$$L_k(r) = \frac{e^r}{k!} \frac{\mathrm{d}^k}{\mathrm{d}r^k} \left(e^{-r} r^k \right).$$

We choose

$$f_k(r) = e^{-r/2} L_k(r)$$

because then our features are orthonormal $\int_0^\infty f_k(r)\, f_l(r)\, \mathrm{d}r = \delta_{kl}$.

We truncated the Laguerre polynomials after the first 20 elements $K = 20$. The potential energy of a configuration is now expressed by

$$E_{\text{Laguerre}}(x) = \sum_{k=1}^{K} \lambda_k \left\{ \sum_{i<j} L_k(\|x_i - x_j\|)\, e^{-\frac{\|x_i - x_j\|}{2}} \right\} \tag{6.14}$$

We use this representation to estimate the corresponding $\lambda_{1\ldots20}$ from the argon simulation. The estimated Laguerre potentials are shown in Figure 6.1B. Although the estimated potential show a higher variance, they are still able to approximate the underlying Lennard-Jones potential accurately. This becomes more visible for estimates from a single configuration, in which the potential exhibits a higher spread.

As a reference, we provide the corresponding potential of mean force (PMF) as proposed by Miyazawa and Jernigan (1985) in Figure 6.1C. The PMF potential is computed through inversion of the radial distribution function. PMFs are still commonly used as potential function in biomolecular simulation and prediction (Miyazawa and Jernigan, 1985; Sippl, 1995). However, PMFs technically represent

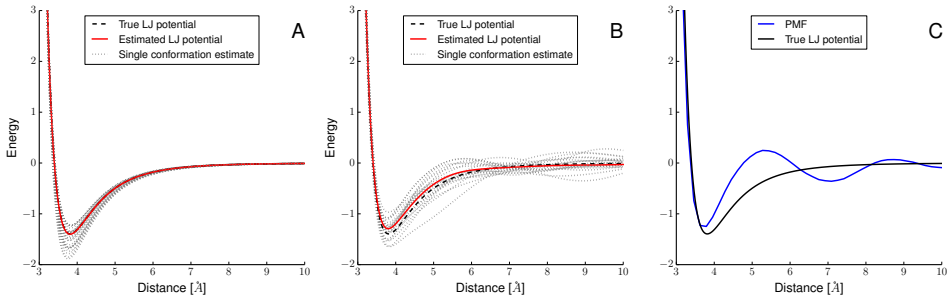


Figure 6.1. Lennard-Jones potential estimated from a simulation of liquid Argon. A: Configurational temperature estimates using the correct features of the Lennard-Jones potential. B: Configurational temperature estimate based on a Laguerre representation of the potential using the 20 first Lagurre polynomials. C: PMF obtained by the inverse Boltzmann law applied to the radial distribution function. The black solid lines indicate the true potential used in the simulation of the Argon fluid. In Panels A and B, the grey dashed curves indicate the potentials recovered from single configurations, the red solid line is the result based on all 20 structures.

the mean energy of changing a single particle in a multi-particle system (Chandler, 1987).

Impact of simulation temperature

The liquid argon lends itself for further analysis. Next, we wish to investigate how the fluctuations of the thermal noise influence the prediction accuracy. To this end, we simulated argon at temperatures starting from $5.0\,\mathrm{K}$ and increasing to $140.0\,\mathrm{K}$ in $5.0\,\mathrm{K}$ steps. For estimation, we used 20 conformations per temperature. The accuracy of the potential depending on temperature and the choice of the vector field B_k is show in in Figure 6.2. The three different vector fields $B_k(x) = \nabla f_k(x)/\|\nabla f_k(x)\|^n, \quad n = 0, 1, 2$ show a similar accuracy. At higher temperature the error increases, due to thermal noise. Only at very low temperatures, when we only explore the ground state of the system, does the estimation of the potential break down.

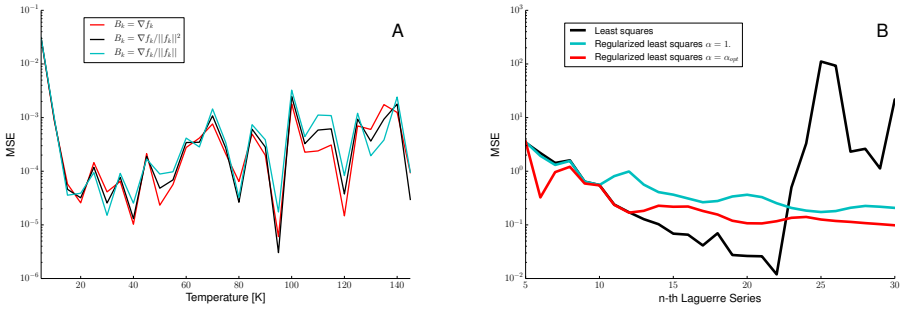


Figure 6.2. Prediction accuracy dependence on system temperature. We show the mean squared error (MSE) between the true potential function in the range of 3 to 10 Å for r^{-6} and r^{-12} basis functions. The figure on the right shows the influence of a regularization term on the accuracy of the Laguerre basis.

Impact of basis functions

We introduced Laguerre polynomials to fit arbitrary distance-dependent potentials. An important parameter is the number of Laguerre polynomials used. If we use too many polynomials, the linear equation (Equation 6.7) becomes increasingly ill-conditioned. Furthermore, it becomes increasingly likely to overfit the data. But if we break off the series expansion too early, the power of the approximation is limited. Therefore, we need a rational method to choose the number of Laguerre polynomials.

Estimating the potential boils down to solving of a system of linear equations (6.7). We recast this as a least squares optimization problem $\min \|A - b\lambda\|^2$, for reasons of numerical stability. This opens new venues, as it is common to use regularization techniques in least squares problems. Adding a regularization term, can also be interpreted as imposing a prior probability on λ. Here, we want to encourage a sparse encoding with as few parameters as possible. Thus, we use a Gaussian prior, where the variance is inverse proportional on k. The functional we minimize is

$$\|\boldsymbol{A}\lambda - \boldsymbol{b}\|^2 + \alpha \sum_{k=1}^{K} k^2 \lambda_k^2$$

where α is the strength of the regularizer. Instead of choosing α *ad hoc* we use an iterative scheme (Besag, 1986) in which we cycle through conditional updates of λ and α. For fixed α we use the LSQR-Algorithm to update λ; for fixed λ we treat α as

the precision of a Gaussian distribution and calculate the update analytically. Figure 6.2 compares the accuracy of the optimized regularization scheme denoted by α_{opt} to the arbitrary choice of $\alpha = 1$. The optimization of α increases the accuracy with little computational cost.

We test this procedure on the 20 snapshots of the liquid argon simulation. The accuracy depending on expansion coefficient is shown in Figure 6.2.

The accuracy of the reconstructed Lennard-Jones potential strongly depends on the maximum number of Laguerre features. In the absence of the regularizer, the problems quickly become ill-conditioned, while the accuracy with the additional regularization term is lower on average. Such a trade-off is commonly observed if a regularization parameter is introduced.

Diatomic fluid

Next, we increase the complexity of the system under investigation by introducing a second atom species. The system now consists of 432 argon and 432 neon atoms, whose dynamics are described using Lennard-Jones potentials with different parameters for each pair of atom types. All other parameter are held fixed. To recover these potentials, we need three different interactions (argon-argon, argon-neon, neon-neon) and two features per interaction r^{-6} and r^{-12}. The parameters of the potentials are recovered as before by solving $\boldsymbol{A} \cdot \boldsymbol{\lambda} = \boldsymbol{b}$ for $\boldsymbol{\lambda}$.

The results for diatomaic fluid are illustrated in Figure 6.3. From 50 samples of the systems, the method is able to recover the parameters of the Lennard-Jones potentials with good accuracy. A single snapshot gives a rough estimate of the potential form.

In order to determine the potential function, we solve $\boldsymbol{A}\boldsymbol{\lambda} = \boldsymbol{b}$ by least squares minimization. From a probabilistic perspective, this is similar to fitting a multivariate Normal distribution with a covariance matrix $(\boldsymbol{A}\boldsymbol{A}^T)^{-1}$. A graphical representation of this matrix can be found in Figure 6.3D. This matrix represents our initial assumption that all parameters of this model are interdependent.

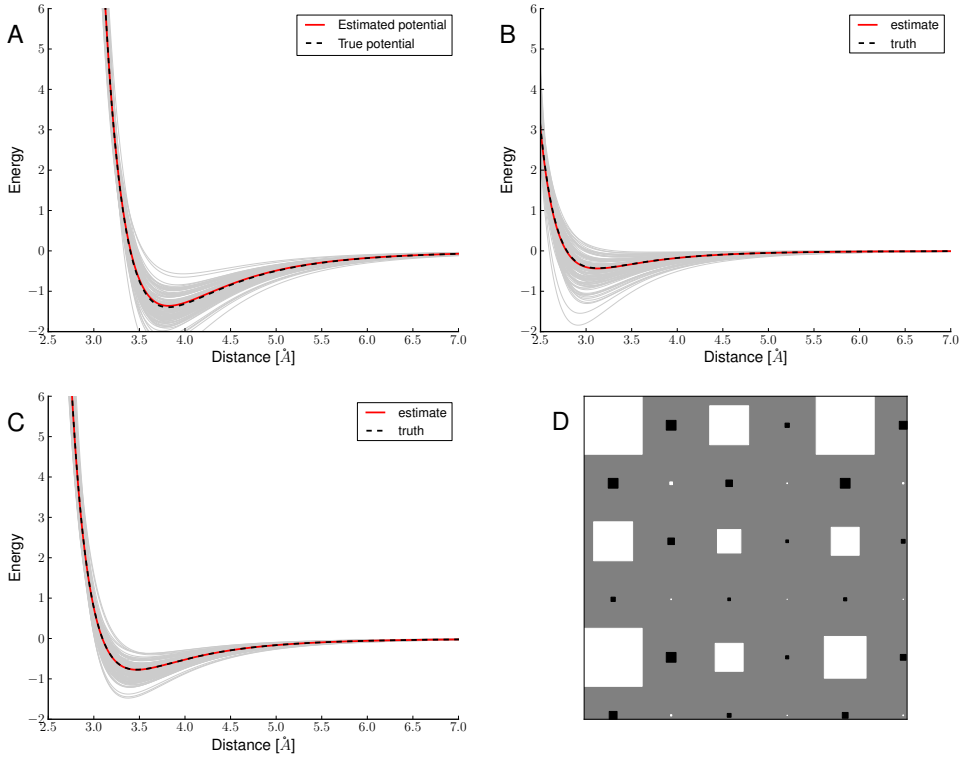


Figure 6.3. Estimated potentials (red broken line) compared to the true potentials (black line) and estimates from single conformations(grey) for the diatomic fluid. A: argon-argon potential. B: argon-neon potential. C: neon-neon potential. D: Hinton diagram where the colour (black/white) indicates the sign (negative/positive) of the covariance and the area of the rectangles is proportional to the magnitude.

Coarse-grained protein model

The next model system is a simplified protein model that is simulated using a coarse grained force field (Honeycutt and Thirumalai, 1990; Sorenson and Head-Gordon, 2002). Each amino acid is represented by a single bead, that is either hydrophobic (B), hydrophilic (L), or neutral (N). The beads are connected by bonds that form the protein backbone. An attractive Lennard-Jones potential between the hydrophobic beads drives compaction of the polypeptide chain, whereas the non-bonded interaction between all other beads is purely repulsive. To emphasize the chain geometry a harmonic potential on bond lengths and angles is used. Sec-

ondary structure elements are not the result of hydrogen bonds, but depend on the sequence and are enforced through a dihedral angle potential. The potential energy $E(\boldsymbol{x})$ is given by:

$$\begin{aligned} E(\boldsymbol{x}) &= \sum_{r \in \text{bonds}} k_{\text{bonds}}(r - r_0)^2 \\ &+ \sum_{\theta \in \text{bond angles}} k_{\text{angles}}(\theta - \theta_0)^2 \\ &+ \sum_{\varphi \in \text{dihedrals}} \left[a\left(1 + \cos(\varphi)\right) + b\left(1 + \cos(3\varphi)\right)\right] \\ &+ \sum_{non-bonded\ r_{ij}} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] \end{aligned}$$

We use a protein with the sequence $\mathtt{B}^9\mathtt{N}^3(\mathtt{LB})^4\mathtt{N}^3\mathtt{B}^9\mathtt{N}^3(\mathtt{LB})^5$, that forms a barrel-like tertiary structure. We used the GROMACS 4.0 (Hess et al., 2008; Berendsen et al., 1995) software to carry out the simulations. All simulations were run using a stochastic thermostat and a step size of 0.01ns. The system was simulated at reduced units with all particles having unity mass and a temperature of $T = 0.28$ As initial configuration an extended conformation is used. After an equilibration phase of 50000 steps, we considered every 500th configuration for further analysis.

Even though we use a simplified model, the potential functions have the same complexity as an all-atom force field. The Honeycutt-Thirumalei model (HT model) even shows an energy landscape comparable to real proteins Brown et al. (2003). We will now use our configurational temperature framework to estimate the parameters of the HT model. But we must first to rearrange the potential of the HT-model into the linear features needed for the configurational temperature estimation. We already demonstrated this for the Lennard-Jones model and use two features, r^{-6} and r^{-12} to estimate the parameters of the hydrophobic interactions and r^{-12} for all other non-bonded interactions. The gradient of the features is non-zero only for the correct interaction type, or if the interaction beads are less then three bonds apart. The bond and bend potentials can be expressed as r^2, r and θ^2, θ, respectively. The original parameters are recovered as $k_{\text{bonds}} = \lambda_{r^2}$ and $r_0 = -\lambda_r / 2\lambda_{r^2}$. We use the same rearrangement for k_{angles} and θ_0. For the parameters for the torsion angle potential, a and b, no transformation is necessary; $\cos(\varphi)$ and $\cos(3\varphi)$ can directly be used as features.

We generated 5000 conformations of the β-barrel-like structures comprising 76 beads. The estimated and true potential, together with the differences between

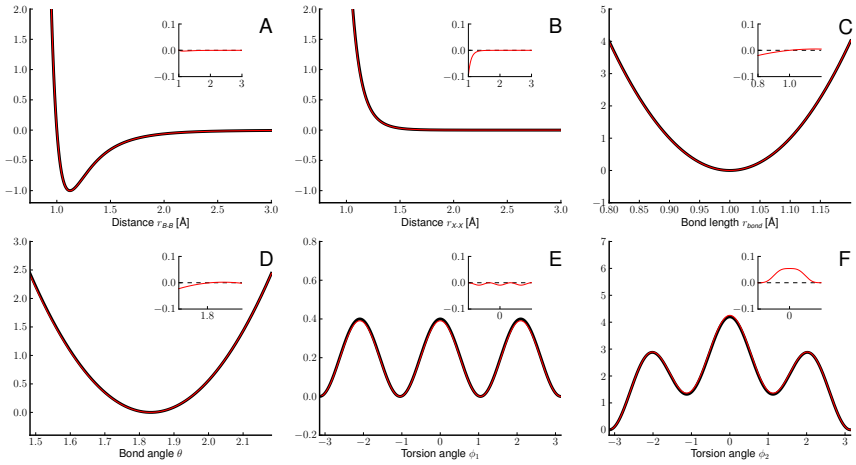


Figure 6.4. Estimated potentials (red broken line) compared to the true potentials (black line) with insets showing the error.

both, are shown in Figure 6.4. Overall, the estimated potentials provide a very close fit, with only minimal error in sparsely sampled regions of the torsion potential.

Again, we solve $\boldsymbol{A\lambda} = \boldsymbol{b}$ by least-squares minimization. The covariance matrix, given by $(\boldsymbol{AA}^T)^{-1}$, provides insight into the dependence of the force field terms. A graphical representation of this matrix can be found in Figure 6.5. It shows that there is considerable interdependence between the features. In particular, the bond and bend potentials are coupled tightly with the Lennard-Jones features.

We test the importance of the covariances by a simple experiment; to recover the individual potentials we use the part of the matrix, that corresponds to the parameters of the potential. For example, to estimate the B-B potential we use only the $r_{\mathrm{B-B}}^{-6}$ and $r_{\mathrm{B-B}}^{-12}$ entries of A and b, which reduces our problem to a system of linear equations given by a 2×2-matrix and a vector of size 2. The potential function estimated with reduced interdependence shows systematic errors in the non-bonded potential (Figure 6.5). In addition, the location of the minima of the bond length and bond angle terms as well as the strength of the dihedral potentials show a significant error.

Figure 6.6 demonstrates how the number of conformations influences the accuracy of the reconstructed potential function. Even with as few as 500 conformations, we are already able to reconstruct a good approximation of the force field, except in the case of dihedral potentials that suffer from poor sampling.

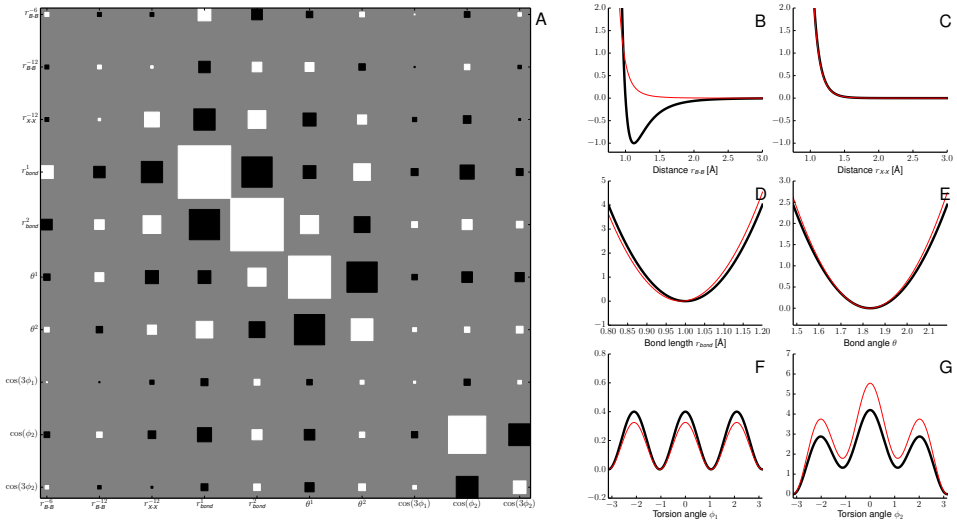


Figure 6.5. A: Hinton diagram showing the covariances $(AA^T)^{-1}$ between the force field terms. The colour (black/white) indicates the sign (negative/positive) of the covariance, and the area of the rectangles is proportional to the magnitude. The features are indicated as axis labels. B-G: Reconstructed potentials that ignore the correlations between the features.

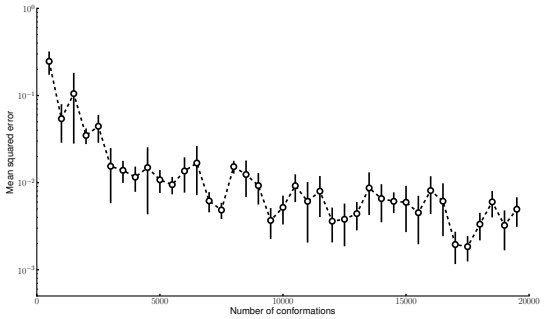


Figure 6.6. Mean squared error of the estimated force field depending on sample size. Mean error and standard deviation were determined based on five non-overlapping sets of structures.

Cβ potential for proteins

Next, we employ the configurational temperature method to infer a distance-dependent potential between the Cβ of proteins. In this practical application, we know neither the correct potential nor the basis function. We chose

$r^{-4}, r^{-6}, r^{-8}, r^{-10}$ and r^{-12} as basis functions f_i. This choice is flexible enough to model functions and quick to implement. Furthermore, we used a r^{-4}, r^{-6} interaction potential, a variant of the well known Lennard-Jones potential to model hydrogen bonds (Fabiola et al., 2002). The centre of the interaction is the Cβ atom of each residue, except for glycine where the Cα atom is used instead. We exclude all 1-2, 1-3 and 1-4 atom pairs from the interactions. Furthermore, we use a Lennard-Jones potential as implemented in the Rosetta structure prediction software (Kuhlman et al., 2003) as reference potential f_0. We can estimate four different potentials, $4-6$ and $4-6-8-10-12$ with and without the reference potential f_0.

The potentials are estimated from the PDBselect25 (Griep and Hobohm, 2010) comprising 3119 chains with 356088 residues. One key assumption that we must additionally, is that all these structures were generated by the same underlying energy function at similar temperatures. Whether this assumption is valid is an ongoing debate Ben-Naim (1997).

The resulting interaction potentials for alanine are shown in Figure 6.7. The reference potential f_0 has only a minor effect on the configurational temperature potentials; it decreases the depth of the potential slightly. The difference between the $4-6$ and $4-6-8-10-12$ potentials is more articulated; $4-6$ potentials are either purely repulsive or unimodal; $4-6-8-10-12$ potentials have up to two minima with very steep potential wells. The functional form of some $4-6-8-10-12$ potential seems to mimic the strong anisotropic effects observed in side chain interactions (Buchete et al., 2004). To include the orientation-dependent interaction a basis set that uses more interaction centres per amino acid is needed. For example, the MARTINI force field (Monticelli et al., 2008) uses up to four interaction centres per amino acid.

The accuracy of statistical potentials is often measured by their ability to identify the native structure as the lowest energy structure among a set of decoys. We compare our conformational temperature potentials to six all-atom knowledge-based potentials; the Random walk potential (RW) (Zhang and Zhang, 2010), the RAPDAF potential (Samudrala and Moult, 1998), the Dope potential (Shen and Sali, 2006), the HA_SRS potential (Rykunov and Fiser, 2010), the Dfire potential (Zhou and Zhou, 2002) and the quasi-chemical knowledge-based potential (KBP) (Lu and Skolnick, 2001). We also include the Lennard-Jones reference potential f_0 in the comparison. We used the CASP decoy set (Rykunov and Fiser, 2010), which comprises 148 targets collected from previous CASP competitions with 2,628 decoys, and the *fisa* decoy set (Samudrala and Levitt, 2000) with five targets and

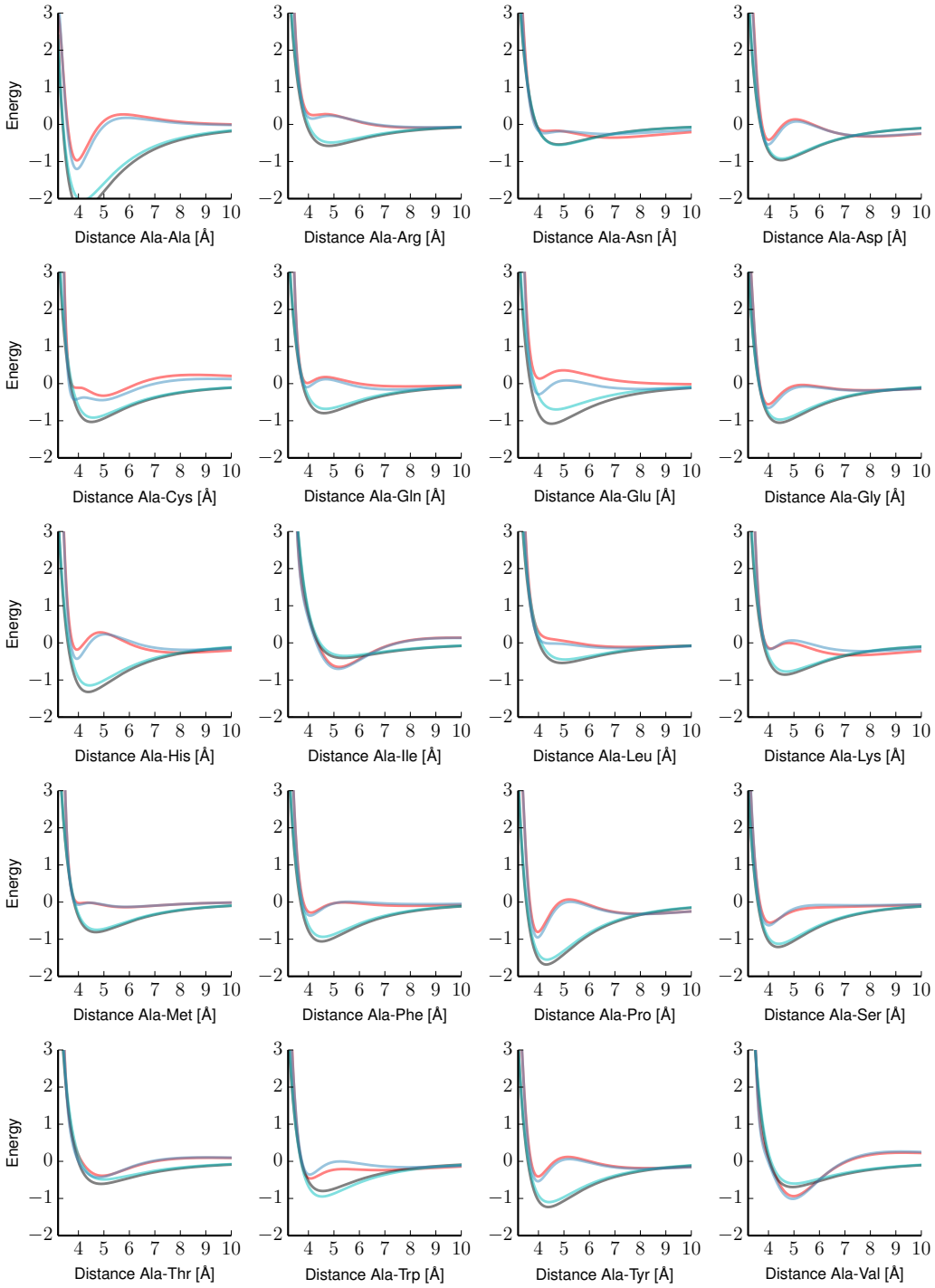


Figure 6.7. Estimated Cβ potentials for Alanine. We show the estimated $4-6$ potentials with (cyan) and without (black) reference potential and $4-6-8-10-12$ potentials with (red) and without (blue) reference potential.

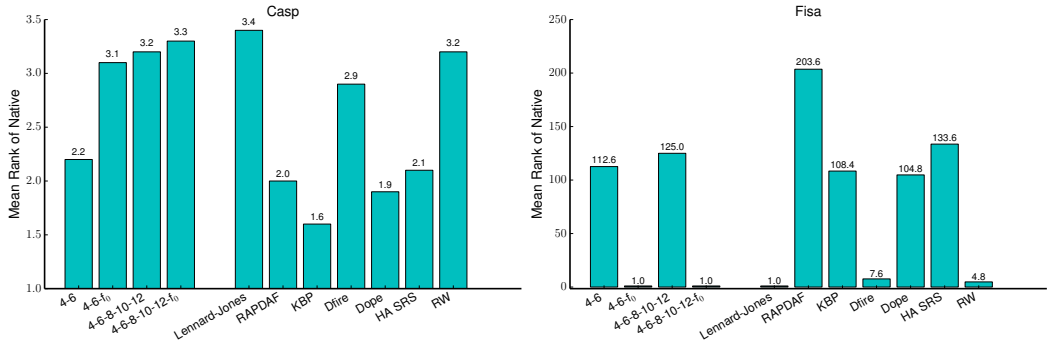


Figure 6.8. Performance of the CT potential in the CASP and fisa benchmarks

more than 1,400 decoys as test sets. The results shown in Figure 6.8 illustrate that the conformational temperature potentials are competitive with the all-atom knowledge-based potentials. Surprisingly, the $4-6$ potential outperforms the more flexible $4-6-8-10-12$ potential. Whether this can be attributed to overfitting of the $4-6-8-10-12$ potential, shortcomings in the assessment or unjustified assumption assumptions will have to be investigated in future research. In general, it seems that decoy discrimination tests, although common in the field, are often unable to discern good and bad potential functions, as the decoys differ substantially from the native protein.

The results on the *fisa* decoy set highlight the ability of the reference potential to take additional physics-based interactions into account. In this test set, all decoys show severe clashes that are easily detected by the reference force field but overlooked by the statistical potentials.

6.4 Conclusion

We present an extension of the configurational temperature formalism that allows us to estimate potential functions from ensemble averages. Our method is applicable to a wide range of systems and does not involve additional simulations. It scales well with the number of configurations and is amenable to standard regularization techniques to avoid overfitting and ill conditioning.

Future research will focus on the derivation of a non-bonded potential for proteins as well as its application to protein simulation and structure determination.

7

Conclusions

The goal of this thesis was to develop and apply Bayesian methods that facilitate the interpretation of data obtained from NMR experiments and allow new insights into protein structures.

In the first chapter, we tackled the prediction of secondary structure elements from assigned chemical shifts. We tried to avoid the pitfalls of "black box"-like approaches that obfuscate the relationship between chemical shift and secondary structure through complex algorithms, and attempted to develop a principled and transparent algorithm for solving this well-studied problem. Using hidden Markov models to represent the relationship between shifts and secondary structure, we are able to predict secondary structure with high accuracy from chemical shifts. The use of probabilistic models is a natural choice as they are able to deal with missing chemical shifts, incorporate evolutionary information and still achieve a higher performance than competing algorithms. We noticed that a second-order HMM, in which the current state depends on the preceding two states, does not lead to a higher prediction accuracy. Thus, it seems that the chemical shift of a residue largely depends on the secondary structure of that residue with very little interference from neighbouring residues. Aside from regular secondary structure, the probabilistic nature of our algorithm enables us to identify and quantify transiently forming secondary structure elements in intrinsically unstructured proteins.

Encouraged by this success, we extended the HMM to predict backbone torsion angles from chemical shifts, but achieved mixed results. Although our approach is competitive with similar algorithms, predicting approximately 65% of all residues within 30° of the observed dihedral angle, there is a lot of room for improvement. It

is left to future projects to investigate whether more powerful methods like Markov random fields (Kindermann et al., 1980) or Hilbert space embedding (Song et al., 2010) are able to solve the problem more satisfactorily. In general, it should be possible to predict dihedral angles accurately from chemical shifts; for instance, the approaches of Cavalli et al. (2007) and Shen et al. (2008) can accurately fold small protein structures close to their native state guided by potential functions and chemical shifts.

Potential functions are often used in computational structure biology as prior information to interpret experimental data. But in structure prediction, it is often unclear what temperature or weight we should assign to the potential function, as many potential functions lack a physical basis for choosing a temperature. This motivated us to investigate the general problem of how to assign an optimal weight to a prior distribution in Bayesian data analysis. Although the prior distribution is an integral part of Bayesian statistics, the weight of the prior is usually inferred via cross-validation or set ad hoc. To offer a more objective choice, we introduce a method to find the optimal weight of the prior in a data-driven way based on replica exchange Monte Carlo algorithms and histogram reweighting techniques. A strength of our method is that it makes no assumptions on the functional form of the prior and likelihood, and is applicable to all data analysis problems in Bayesian statistics that have a closed form expression of the posterior. Nevertheless, the versatility of the algorithm comes at a price: for our algorithm to work, we need to generate samples from tempered posterior and prior distributions. The computational cost of the extended sampling makes the algorithm prohibitive slow for some applications. The approach could be improved in terms of computational efficiency by using different sampling algorithms like Hamiltonian annealed importance sampling (Sohl-Dickstein and Culpepper, 2012) and nested sampling (Skilling, 2004). Despite these hurdles, we applied our algorithm successfully to structure calculation and image analysis. The latter is an especially interesting topic for further applications of our algorithm. In particular, it has been show that the prior distribution in imaging applications can drastically improve reconstruction and superresolution methods (He et al., 2011).

Another application of this algorithm to structure determination is shown in Chapter 5, where we infer the weight of a new Ramachandran potential in the presence of a Lennard-Jones potential. The optimal combination of both force fields leads to more accurate structures from noisy and sparse NMR data. Furthermore, our findings suggest that no universally optimal weight exists and that the weight should be determined based on the experimental data. We found that, in the case

of low-quality data, the weight of the statistical force field should be decreased because the forces that guide the ensemble towards the correct structure are weaker with low-quality data than with high-quality data. Our approach gives the data a chance to speak for itself and not be overwhelmed by the prior. Although the runtime requirements make widespread use of Bayesian model comparison difficult, one should keep these conclusions in mind when dealing with noisy data. Future applications will include the combination of additional potentials that model hydrophobic interactions (Lazaridis and Karplus, 1999; Ferrara et al., 2002) and hydrogen bonding (Kortemme et al., 2003), both being important forces in protein folding (Dill, 1990). The optimization of more than one potential is challenging, as we need to use multidimensional replica-exchange methods (Sugita et al., 2000) to sample configurations for a predefined set of weights. The sampled energies are used in a later maximization step that determines the optimal weight. However, in the case of multidimensional replica-exchange methods the number of ensembles from which we need to sample grows exponentially with the number of potentials. A less demanding alternative, that does not suffer from the ”curse of dimensionality”, is the exploration of the extended ensemble using different combinations of techniques like Gaussian Process upper-confidence bounds (Srinivas et al., 2009) or nested sampling (Skilling, 2004).

Many biological processes take place at different length scales; changes at a molecular level are propagated upward in scale to effect large biological units. The analysis of large proteins by accurate physics-based potential function is often intractable due to the large number of degrees of freedom involved. If we want to calculate structures of large molecular assemblies with ISD, we need potential functions that provide an accurate description of biological processes across different scales and are based on a reduced set of variables. To this end, we introduced an extension of the configurational temperature formalism that allows us to estimate coarse-grained and atomic interaction potentials from molecular configurations. Over the course of this project, we uncovered several interesting connections to existing methods in machine learning (Hyvärinen, 2005) and coarse graining (Mullinax and Noid, 2010) that are special cases of the configurational temperature formalism we introduced. The comparison with several potentials of mean force indicates that our method is competitive in accuracy, yet does not suffer from the idiosyncrasies (Ben-Naim, 1997) of potentials of mean force and leads to analytically tractable potential functions that can be used in molecular dynamics and ISD. Based on the configurational temperature formalism, we will further explore the use of new coarse-grained and atomic-resolution potentials in structure calculation.

In summary, we hope that our contributions taken together will improve structure calculation and, ultimately, help biologists to gain new insights into protein structure and function.

Appendix A

Chemical shift

A.1 Influence of individual chemical shifts on the prediction accuracy

Incomplete chemical shift assignments are often encountered in chemical shift lists. A good prediction method should also work with partial chemical shift assignments. To study this issue, we systematically removed measurements from a nearly ($\geq$ 90%) complete test set and predicted the secondary structure using a zero and first order HMM with Gaussian emissions. The influence of missing nuclei on the prediction accuracy is listed in Tables A.1 and A.2.

Table A.1. Influence of missing chemical shifts for all combinations of nuclei on the prediction accuracy of the zero order HMM.

Observed chemical shift(s)	Q_3-score	H	E	C
C	67.6	76.5	65.4	63.6
N	54.9	62.8	50.2	52.7
CA	68.7	83.4	68.5	61.7
CB	58.3	78.7	67.2	41.3
HA	65.6	80.3	74.4	52.4
C N	69.4	79.9	69.7	62.9
C CA	73.0	84.9	73.8	66.1
C CB	69.5	83.0	72.3	58.2
N HA	68.8	84.2	79.4	51.3
CA N	71.1	85.7	72.2	62.7
CB N	62.0	83.2	77.9	37.2
C HA	74.5	84.6	77.5	65.7
CA CB	71.0	85.6	76.3	59.1
CB HA	67.4	81.9	77.4	53.3
CA HA	73.5	86.1	75.7	64.4
C CB N	72.0	84.9	79.8	56.2
C N HA	76.3	86.1	81.5	64.9
C CA N	74.9	86.4	77.7	66.7
CA CB N	74.0	87.4	82.4	59.4
C CA HA	76.5	87.1	78.0	67.9
C CB HA	75.1	85.0	78.9	65.8
CB N HA	70.5	84.8	82.4	52.9
C CA CB	73.8	86.5	76.1	63.2
CA N HA	75.9	87.7	80.4	64.4
CA CB HA	74.7	86.1	78.0	65.9
C CB N HA	76.9	86.5	83.3	64.8
C CA CB N	76.4	87.8	82.1	63.0
C CA N HA	78.4	88.1	81.8	68.2
C CA CB HA	77.2	86.9	79.5	68.8
CA CB N HA	77.2	87.7	82.6	66.2
C CA CB N HA	79.1	88.0	83.6	68.7

Table A.2. Influence of missing chemical shifts for all combinations of nuclei on the prediction accuracy of the first order HMM.

Observed chemical shift(s)	Q_3-score	H	E	C
-	49.5	91.0	11.4	43.2
C	76.9	84.2	65.5	77.4
N	67.6	85.6	57.8	61.9
CA	80.1	85.4	68.9	81.5
CB	73.3	84.8	73.0	66.1
HA	78.3	85.2	74.1	75.9
C N	78.5	86.3	71.3	77.0
C CA	81.5	86.2	73.4	82.3
C CB	79.5	86.2	75.4	76.8
N HA	80.2	87.9	80.3	74.4
CA N	80.7	87.4	73.4	80.1
CB N	76.6	87.5	81.1	66.0
C HA	81.9	86.4	77.1	81.2
CA CB	81.7	86.2	76.8	81.2
CB HA	79.1	85.1	77.5	76.0
CA HA	83.0	86.9	75.2	84.2
C CB N	81.2	87.8	82.4	75.0
C N HA	83.4	87.8	82.5	80.2
C CA N	82.6	87.5	77.6	82.0
CA CB N	83.1	88.3	83.5	79.0
C CA HA	83.3	87.4	77.3	83.6
C CB HA	82.2	86.5	79.1	80.8
CB N HA	81.2	87.8	83.1	75.1
C CA CB	82.4	87.0	78.3	81.3
CA N HA	84.0	88.6	80.7	82.3
CA CB HA	83.0	86.7	77.6	83.3
C CB N HA	83.7	88.1	84.2	79.7
C CA CB N	83.7	88.4	83.9	79.7
C CA N HA	84.7	88.7	82.6	82.6
C CA CB HA	83.8	87.3	79.4	83.2
CA CB N HA	84.2	88.5	83.0	81.8
C CA CB N HA	84.8	88.6	84.2	82.1

Appendix B

Derivations

B.1 Divergence of the test functions B_k in Boltzmann inversions

In Chapter 6 we showed how derive the divergence of $B_k = \nabla f_k$, but left the derivation of the divergence of $B_k = \frac{\nabla f_k}{\|\nabla f_k\|}$ and $B_k = \frac{\nabla f_k}{\|\nabla f_k\|^2}$ unanswered. Here, we quickly demonstrate how to derive these divergences starting starting with $B_k = \frac{\nabla f_k}{\|\nabla f_k\|^2}$.

$$
\begin{aligned}
(\nabla f)^T \nabla ||\nabla f||^{-2} &= (\nabla f)^T \nabla \frac{1}{(\nabla f)^T \nabla f} \\
&= -\frac{(\nabla f)^T}{[(\nabla f)^T \nabla f]^2} \nabla [(\nabla f)^T \nabla f] \\
&= -2\frac{(\nabla f)^T}{[(\nabla f)^T \nabla f]^2} (\nabla \nabla^T f) \nabla f \\
&= -2\frac{(\nabla f)^T (\nabla \nabla^T f) \nabla f}{\|\nabla f\|^4} \\
&= -2\frac{\sum_{ij} (\partial_i f)(\partial_i \partial_j f)(\partial_j f)}{\|\nabla f\|^4}
\end{aligned}
$$

where $\nabla \nabla^T f$ is the Hessian of f and $\partial_i f = \frac{\partial f}{\partial x_i}$. In the third step, we use the fact that

$$\nabla [(\nabla f)^T \nabla f] = 2(\nabla \nabla^T f) \nabla f$$

which follows from the fact that for two vector valued functions F and G (i.e. $F(x) = (F_1(x), \ldots, F_n(x))^T$), we have:

$$\nabla(F^T G) = (\nabla F^T)G + (\nabla G^T)F$$

where ∇F^T is the Jacobian matrix with entries $\partial_i F_j$.

Yet another feature would be $B_k = \frac{\nabla f_k}{\|\nabla f_k\|}$ with divergence:

$$\begin{aligned}
\nabla \frac{\nabla f_k}{\|\nabla f_k\|} &= \frac{\Delta f_k}{\|\nabla f_k\|} + (\nabla f_k)^T \nabla \|\nabla f_k\|^{-1} \\
&= \frac{\Delta f_k}{\|\nabla f_k\|} + (\nabla f_k)^T \nabla [(\nabla f_k)^T (\nabla f_k)]^{-1/2} \\
&= \frac{\Delta f_k}{\|\nabla f_k\|} - \frac{1}{2} (\nabla f_k)^T [(\nabla f_k)^T (\nabla f_k)]^{-3/2} \nabla [(\nabla f_k)^T (\nabla f_k)] \\
&= \frac{\Delta f_k}{\|\nabla f_k\|} - \frac{(\nabla f_k)^T (\nabla \nabla^T f_k)(\nabla f_k)}{\|\nabla f_k\|^3}
\end{aligned}$$

We now consider expansion:

$$v(\boldsymbol{x}) = \boldsymbol{\lambda} \cdot \boldsymbol{f}(\boldsymbol{x}) + g(\boldsymbol{x})$$

and a series of vector fields

$$\boldsymbol{B}_k(\boldsymbol{x}) = \boldsymbol{\nabla} f_k(\boldsymbol{x}), \quad k = 1, \ldots, K.$$

We obtain the following system of equations:

$$\langle \boldsymbol{\nabla} f_k \cdot \boldsymbol{\nabla} (\boldsymbol{\lambda} \cdot \boldsymbol{f} + g) \rangle_X = \langle \Delta f_k \rangle_X, \quad k = 1, \ldots, K. \tag{B.1}$$

where $\Delta = \boldsymbol{\nabla} \cdot \boldsymbol{\nabla}$ is the Laplace operator. This is a linear system of equations determining the expansion coefficients λ_k:

$$\boldsymbol{A} \cdot \boldsymbol{\lambda} = \boldsymbol{b}, \quad A_{kl} = \langle \boldsymbol{\nabla} f_k \cdot \boldsymbol{\nabla} f_l \rangle_X, \; b_k = \langle \Delta f_k - \boldsymbol{\nabla} f_k \cdot \boldsymbol{\nabla} g \rangle_X. \tag{B.2}$$

Clearly, $\boldsymbol{A}$ is a positive semi-definite $K \times K$ matrix.

Appendix C

Publications

Ideas and figures have previously appeared in the following publications:

- *M. Mechelke* and M. Habeck. *Robust probabilistic superposition and comparison of protein structures.* BMC Bioinformatics. 2010 Jul;11:363.
  URL `http://www.ncbi.nlm.nih.gov/pubmed/20594332`.

- I. Kalev, *M. Mechelke*, K.O. Kopec , T. Holder, S. Carstens, and M. Habeck. *CSB: a Python framework for structural bioinformatics.* Bioinformatics. 2012 Nov;28(22):2996.
  URL `http://www.ncbi.nlm.nih.gov/pubmed/22942023`.

- *M. Mechelke* and M. Habeck. *Calibration of Boltzmann distribution priors in Bayesian data analysis.* Phys Rev E. 2012 Dec;86(6 Pt 2):066705.
  URL `http://www.ncbi.nlm.nih.gov/pubmed/23368076`.

  – Text and figures from this manuscript appear in Chapter 3 of this thesis.

- *M. Mechelke* and M. Habeck. *A probabilistic model for secondary structure prediction from protein chemical shifts.* Proteins. 2013 Jun;81(6):984-93.
  URL `http://www.ncbi.nlm.nih.gov/pubmed/23368076`.

  – Text and figures from this manuscript appear in Chapter 2 of this thesis.

- *M. Mechelke* and M. Habeck. *Estimation of interaction potentials through the configurational temperature formalism.* J. Chem. Theory Comput., 2013 Dec; Epub ahead of print.
  URL `http://pubs.acs.org/doi/abs/10.1021/ct400580p`.

  – Text and figures from this manuscript appear in Chapter 5 of this thesis.

Appendix D

Contributions

Chapter 2 - Predicting secondary structure from chemical shifts

This work is part of a manuscript that has been published Mechelke and Habeck (2013b). Martin Mechelke and Michael Habeck conceived the project. Martin Mechelke implemented the algorithms and analysed the results.

Chapter 3 - Weighting priors in Bayesian data analysis

This work is part of a manuscript that has been published Mechelke and Habeck (2012). Michael Habeck and Martin Mechelke conceived the project. Martin Mechelke performed the simulations and analysed the results; Michael Habeck contributed to the Ising model and to the discussion.

Chapter 4 - Optimal combination of statistical potentials in NMR structure calculation

Martin Mechelke and Michael Habeck conceived the project. Martin Mechelke implemented the Ramachandran potential, performed the simulations and analysed the results; Michael Habeck contributed to the discussion.

Chapter 5 - Estimating energy functions from Boltzmann ensembles

This work is part of a manuscript that has been published Mechelke and Habeck (2013a). Martin Mechelke and Michael Habeck conceived the project, Martin Mechelke implemented the algorithms and analysed the results; Michael Habeck contributed to the discussion.

Bibliography

H. Akaike. *A new look at the statistical model identification*. Automatic Control, IEEE Transactions on, 19(6):(1974) 716 – 723. ISSN 0018-9286.

A. Baumgärtner. *Statics and dynamics of the freely jointed polymer chain with lennard-jones interaction*. The Journal of Chemical Physics, 72:(1980) 871.

T. Bayes. *An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price*. Philosophical Transactions (1683-1775), pages 370–418.

M. Bayrhuber, T. Meins, M. Habeck, S. Becker, K. Giller, S. Villinger, C. Vonrhein, C. Griesinger, M. Zweckstetter, and K. Zeth. *Structure of the human voltage-dependent anion channel*. Proc. Natl. Acad. Sci. USA, 105:(2008) 15370–15375.

P. D. Beale. *Exact Distribution of Energies in the Two-Dimensional Ising Model*. Phys. Rev. Lett., 76:(1996) 78–81.

A. Ben-Naim. *Statistical potentials extracted from protein structures: Are these meaningful potentials?* The Journal of Chemical Physics, 107:(1997) 3698.

H. J. Berendsen, D. van der Spoel, and R. van Drunen. *Gromacs: A message-passing parallel molecular dynamics implementation*. Computer Physics Communications, 91(1):(1995) 43–56.

J. Besag. *On the statistical analysis of dirty pictures*. Journal of the Royal Statistical Society. Series B (Methodological), pages 259–302.

R. B. Best and G. Hummer. *Optimized molecular dynamics force fields applied to the helix- coil transition of polypeptides*. The Journal of Physical Chemistry B, 113(26):(2009) 9004–9015.

C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

F. Bloch, W. W. Hansen, and M. Packard. *The nuclear induction experiment*. Phys. Rev., 70:(1946) 474–485.

W. Boomsma, K. Mardia, C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. *A generative, probabilistic model of local protein structure*. Proc. Natl. Acad. Sci. USA, 105:(2008) 8932–8937.

M. Born and H. Green. *A general kinetic theory of liquids. i. the molecular distribution functions*. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 188(1012):(1946) 10–18.

S. Brown, N. J. Fawzi, and T. Head-Gordon. *Coarse-grained sequences for protein folding and design*. Proceedings of the National Academy of Sciences, 100(19):(2003) 10712–10717.

T. A. Brown. *Genomes 2*. Bios Scientific Publishers, 2002.

A. T. Brünger and M. Nilges. *Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy.* Q. Reviews of Biophys., 26(1):(1993) 49–125.

N.-V. Buchete, J. E. Straub, and D. Thirumalai. *Orientational potentials extracted from protein structures improve native fold recognition*. Protein Science, 13(4):(2004) 862–874.

M. Buck, S. Bouguet-Bonnet, R. W. Pastor, and A. D. MacKerell Jr. *Importance of the cmap correction to the charmm22 protein force field: dynamics of hen lysozyme*. Biophysical journal, 90(4):(2006) L36–L38.

C. Camilloni, A. D. Simone, W. F. Vranken, and M. Vendruscolo. *Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts.* Biochemistry, 51(11):(2012) 2224–2231.

D. A. Case. *The use of chemical shifts and their anisotropies in biomolecular structure determination*. Curr. Opin. Struct. Biol., 8:(1998) 624–630.

F. Castellani, B. van Rossum, A. Diehl, M. Schubert, K. Rehbein, and H. Oschkinat. *Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy.* Nature, 420:(2002) 98–102.

A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo. *Protein structure determination from NMR chemical shifts*. Proc. Natl. Acad. Sci. USA, 104:(2007) 9615–9620.

J. Cavanagh, W. J. Fairbrother, A. G. Palmer III, and N. J. Skelton. *Protein NMR spectroscopy, principles and practice*. Academic Press, 1996.

D. Chandler. *Introduction to modern statistical mechanics*, volume 1. Oxford University Press, 1987.

M. S. Cheung, M. L. Maguire, T. J. Stevens, and R. W. Broadhurst. *DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure*. J. Magn. Reson., 202:(2010) 223–233.

G. Cornilescu, J.-S. Hu, and A. Bax. *Identification of the hydrogen bonding network in a protein by scalar couplings.* J. Am. Chem. Soc., 121:(1999) 2949–2950.

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

R. T. Cox. *Probability, frequency and reasonable expectation.* Am. J. Phys., 14:(1946) 1–13.

R. Das. *Four small puzzles that rosetta doesn't solve*. PLoS One, 6(5):(2011) e20044.

I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, et al. *Molprobity: all-atom contacts and structure validation for proteins and nucleic acids*. Nucleic acids research, 35(suppl 2):(2007) W375–W383.

B. De Finetti. *Theory of probability*. CUP Archive, 1970.

K. A. Dill. *Dominant forces in protein folding*. Biochemistry, 29:(1990) 7133–7155.

B. Donald. *Algorithms in Structural Molecular Biology*. Computational Molecular Biology. MIT Press, 2011.

S. Duane, A. D. Kennedy, B. Pendleton, and D. Roweth. *Hybrid Monte Carlo*. Phys. Lett. B, 195:(1987) 216–222.

H. R. Eghbalnia, L. Wang, A. Bahrami, A. Assadi, and J. L. Markley. *Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements*. J. Biomol. NMR, 32:(2005) 71–81.

R. R. Ernst, G. Bodenhausen, and A. Wokaun. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford University Press, New York, 1990.

F. Fabiola, R. Bertram, A. Korostelev, and M. S. Chapman. *An improved hydrogen bond potential: impact on medium resolution protein structures*. Protein science, 11(6):(2002) 1415–1423.

M. Feig. *Is alanine dipeptide a good model for representing the torsional preferences of protein backbones?* Journal of Chemical Theory and Computation, 4(9):(2008) 1555–1564.

P. Ferrara, J. Apostolakis, and A. Caflisch. *Evaluation of a fast implicit solvent model for molecular dynamics simulations*. Proteins, 46:(2002) 24–33.

A. M. Ferrenberg and R. H. Swendsen. *Optimized Monte Carlo Data Analysis*. Phys. Rev. Lett., 63:(1989) 1195–1198.

P. L. Freddolino, S. Park, B. Roux, and K. Schulten. *Force field bias in protein folding simulations*. Biophysical journal, 96(9):(2009) 3772.

A. Gelman and C. P. Robert. *"not only defended but also applied": The perceived absurdity of bayesian inference*, 2011.

S. Geman and D. Geman. *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*. IEEE Trans. PAMI, 6(6):(1984) 721–741.

S. Geman and D. E. McClure. *Statistical methods for tomographic image reconstruction*. Bul. Int. Stat. Inst., LII-4:(1987) 5–21.

C. J. Geyer. *Markov chain Monte Carlo maximum likelihood*. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. 1991.

H. Gong, Y. Shen, and G. D. Rose. *Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly*. Protein Sci., 16:(2007) 1515–1521.

I. J. Good. *Probability and the Weighing of Evidence*. C. Griffin London, 1950.

S. Griep and U. Hobohm. *PDBselect 1992-2009 and PDBfilter-select*. Nucleic Acids Res., 38:(2010) D318–319.

M. Habeck. *Statistical mechanics analysis of sparse data*. J. Struct. Biol., 173:(2011) 541–548.

M. Habeck. *Evaluation of marginal likelihoods using the density of states*. In N. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, pages 486–494. JMLR: W&CP 22, 2012.

M. Habeck, M. Nilges, and W. Rieping. *Replica-Exchange Monte Carlo scheme for Bayesian data analysis*. Phys. Rev. Lett., 94:(2005a) 0181051–0181054.

M. Habeck, W. Rieping, and M. Nilges. *Bayesian estimation of Karplus parameters and torsion angles from three-bond scalar coupling constants*. J. Magn. Reson., 177:(2005b) 160–165.

M. Habeck, W. Rieping, and M. Nilges. *Weighting of experimental evidence in macromolecular structure determination*. Proc. Natl. Acad. Sci. USA, 103:(2006) 1756–1761.

T. Hamelryck. *Probabilistic models and machine learning in structural bioinformatics*. Statistical methods in medical research, 18(5):(2009a) 505–526.

T. Hamelryck. *Probabilistic models and machine learning in structural bioinformatics*. Statistical methods in medical research, 18(5):(2009b) 505–526.

K. He, J. Sun, and X. Tang. *Single image haze removal using dark channel prior*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33(12):(2011) 2341–2353.

B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. *Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation*. Journal of Chemical Theory and Computation, 4(3):(2008) 435–447.

K. Hinsen. *The molecular modeling toolkit: a new approach to molecular simulations*. Journal of Computational Chemistry, 21(2):(2000) 79–85.

U. Hobohm, R. Scharf, M. Schneider, and C. Sander. *Selection of a representative set of structures from the Brookhaven protein data bank*. Protein Sci., 1:(1992) 409–417.

J. D. Honeycutt and D. Thirumalai. *Metastability of the folded states of globular proteins*. Proc. Natl. Acad. Sci. U.S.A., 87(9):(1990) 3526–3529.

V. S. Honndorf, N. Coudevylle, S. Laufer, S. Becker, C. Griesinger, and M. Habeck. *Inferential NMR/X-ray-based structure determination of a dibenzo[a,d]cycloheptenone inhibitor-p38 MAP kinase complex in solution*. Angew. Chem. Int. Ed. Engl., 51(10):(2012) 2359–2362.

R. W. Hooft, C. Sander, and G. Vriend. *Objectively judging the quality of a protein structure from a Ramachandran plot.* Comput. Appl. Biosci., 13:(1997) 425–430.

R. W. Hooft, G. Vriend, C. Sander, and E. E. Abola. *Errors in protein structures*. Nature, 381:(1996) 272.

S. Hovmöller, T. Zhou, and T. Ohlson. *Conformations of amino acids in proteins*. Acta Cryst. sect. D, 58:(2002) 768–776.

J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback. *Bayesian inference of phylogeny and its impact on evolutionary biology*. science, 294(5550):(2001) 2310–2314.

L. H. Hung and R. Samudrala. *Accurate and automated classification of protein secondary structure with PsiCSI*. Protein Sci., 12:(2003) 288–295.

A. Hyvärinen. *Estimation of non-normalized statistical models using score matching*. Journal of Machine Learning Research, 6:(2005) 695–709.

J. Inoue and K. Tanaka. *Dynamics of the maximum marginal likelihood hyperparameter estimation in image restoration: gradient descent versus expectation and maximization algorithm*. Phys. Rev. E, 65(1 Pt 2):(2002) 016125.

S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth. *Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching*. J. Chem. Phys., 120:(2004a) 10896.

S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth. *Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching*. The Journal of chemical physics, 120:(2004b) 10896.

E. T. Jaynes. *Information Theory and Statistical Mechanics*. Phys. Rev., 106:(1957) 620–630.

H. Jeffreys. *Theory of Probability.* Oxford University Press, 1939.

M. R. Jensen, K. Houben, E. Lescop, L. Blanchard, R. W. H. Ruigrok, and M. Blackledge. *Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of sendai virus nucleoprotein.* J Am Chem Soc, 130(25):(2008) 8055–8061.

O. G. Jepps, G. Ayton, and D. J. Evans. *Microscopic expressions for the thermodynamic temperature*. Phys. Rev. E, 62:(2000) 4757–4763.

V. Johnson, W. Wong, X. Hu, and C. Chen. *Image restoration using gibbs priors: Boundary modeling, treatment of blurring, and selection of hyperparameter*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13:(1991) 413–425.

D. T. Jones. *Protein secondary structure prediction based on position-specific scoring matrices*. J. Mol. Biol., 292:(1999) 195–202.

M. I. Jordan and A. Y. Ng. *On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes*. Advances in neural information processing systems, 14:(2002) 841.

W. Kabsch and C. Sander. *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. Biopolymers, 22:(1983) 2577–2637.

M. Karplus. *Vicinal proton coupling in nuclear magnetic resonance.* J. Am. Chem. Soc., 85:(1963) 2870–2871.

R. Kass and A. Raftery. *Bayes factors*. American Statistical Association, 90:(1995) 773–775.

R. Kindermann, J. L. Snell, et al. *Markov random fields and their applications*, volume 1. American Mathematical Society Providence, RI, 1980.

S. Kirkpatrick, C. D. J. Gelatt, and M. P. Vecchi. *Optimization by simulated annealing.* Science, 220:(1983) 377–385.

H. Kiwata. *Physical consideration of an image in image restoration using bayes' formula*. Physica A, 391(6):(2012) 2215 – 2224.

T. Kortemme, A. Morozov, and D. Baker. *An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes*. J. Mol. Biol., 326:(2003) 1239–1259.

B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. *Design of a novel globular protein fold with atomic-level accuracy*. Science, 302:(2003) 1364–1368.

S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. *The weighted histogram analysis method for free-energy calculations on biomolecules*. J. Comp. Chem., 13:(1992) 1011–1021.

J. Kuszewski and G. M. Clore. *Sources of and solutions to problems in the refinement of protein NMR structures against torsion angle potentials of mean force*. J. Magn. Reson., 146:(2000) 249–254.

D. Labudde, D. Leitner, M. Kruger, and H. Oschkinat. *Prediction algorithm for amino acid types with their secondary structure in proteins (PLATON) using chemical shifts*. J. Biomol. NMR, 25:(2003) 41–53.

T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande. *To milliseconds and beyond: challenges in the simulation of protein folding*. Curr. Opin. Struct. Biol, 23(1):(2012) 58–65.

O. F. Lange, D. Van der Spoel, and B. L. De Groot. *Scrutinizing molecular mechanics force fields on the submicrosecond timescale with nmr data*. Biophysical journal, 99(2):(2010) 647–655.

P. S. Laplace. *Memoir on the probability of the causes of events*. Statistical Science, 1(3):(1774) 364–378.

R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. *PROCHECK: a program to check the stereochemical quality of protein structures*. J. Appl. Cryst., 26:(1993) 283–291.

C. W. Lawrence, A. Bonny, and S. A. Showalter. *The disordered c-terminus of the rna polymerase ii phosphatase fcp1 is partially helical in the unbound state*. Biochem Biophys Res Commun, 410(3):(2011) 461–465.

T. Lazaridis and M. Karplus. *Discrimination of the native from misfolded protein models with an energy function including implicit solvation*. J. Mol. Biol., 288:(1999) 477–487.

M. H. Levitt. *Spin dynamics*. Wiley New York, 2001.

S. Z. Li. *Markov random field modeling in image analysis*. Springer, 2009.

K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. *Systematic validation of protein force fields against experimental data*. PLoS ONE, 7(2):(2012) e32131.

J. P. Linge and M. Nilges. *Influence of non-bonded parameters on the quality of NMR structures: a new force-field for NMR structure calculation*. J. Biomol. NMR, 13:(1999) 51–59.

H. Lu and J. Skolnick. *A distance-dependent atomic knowledge-based potential for improved protein structure selection*. Proteins: Structure, Function, and Bioinformatics, 44(3):(2001) 223–232.

A. P. Lyubartsev and A. Laaksonen. *Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach*. Phys. Rev. E, 52(4):(1995) 3730–3737.

D. J. C. MacKay. *Bayesian interpolation*. Neural Computation, 4:(1992) 415–447.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge UK, 2003.

A. D. Mackerell. *Empirical force fields for biological macromolecules: overview and issues*. Journal of computational chemistry, 25(13):(2004) 1584–1604.

T. K. Mal, S. J. Matthews, H. Kovacs, I. D. Campbell, and J. Boyd. *Some NMR experiments and a structure determination employing a {^{15}N, ^{2}H} enriched protein*. J. Biomol. NMR, 12:(1998) 259–276.

K. V. Mardia, C. C. Taylor, and G. K. Subramaniam. *Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data*. Biometrics, 63:(2007) 505–512.

J. L. Markley, D. H. Meadows, and O. Jardetzky. *Nuclear magnetic resonance studies of helix-coil transitions in polyamino acids*. J. Mol. Biol., 27:(1967) 25–40.

R. McGreevy and L. Pusztai. *Reverse monte carlo simulation: a new technique for the determination of disordered structures*. Molecular Simulation, 1(6):(1988) 359–367.

M. Mechelke and M. Habeck. *Calibration of boltzmann distribution priors in bayesian data analysis*. Physical Review E, 86(6):(2012) 066705.

M. Mechelke and M. Habeck. *Estimation of interaction potentials through the configurational temperature formalism*. Journal of Chemical Theory and Computation, 0(ja):(2013a) null.

M. Mechelke and M. Habeck. *A probabilistic model for secondary structure prediction from protein chemical shifts*. Proteins: Structure, Function, and Bioinformatics.

N. Metropolis, M. Rosenbluth, A. Rosenbluth, A. Teller, and E. Teller. *Equation of state calculations by fast computing machines*. J. Chem. Phys., 21:(1957) 1087–1092.

S. P. Mielke and V. Krishnan. *Characterization of protein secondary structure from nmr chemical shifts*. Progress in Nuclear Magnetic Resonance Spectroscopy, 54(3-4):(2009) 141 – 165.

T. Minka. *Estimating a dirichlet distribution*. `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`, 2000.

S. Miyazawa and R. L. Jernigan. *Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation*. Macromolecules, 18(3):(1985) 534–552.

L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink. *The martini coarse-grained force field: extension to proteins*. Journal of Chemical Theory and Computation, 4(5):(2008) 819–834.

J. Mullinax and W. Noid. *Generalized yvon-born-green theory for molecular systems*. Physical review letters, 103(19):(2009) 198104.

J. W. Mullinax and W. G. Noid. *Recovering physical potentials from a model protein databank*. Proc. Natl. Acad. Sci. U. S. A., 107:(2010) 19867–19872.

V. Muñoz and L. Serrano. *Elucidating the folding problem of helical peptides using empirical parameters*. Nature Structural & Molecular Biology, 1(6):(1994) 399–409.

D. Neuhaus and M. P. Williamson. *The nuclear Overhauser effect in structural and conformational analysis, 2nd ed.* Wiley-VCH Inc., New York, 2000.

V. S. Pande. *(Compressed) sensing and sensibility*. Proc. Natl. Acad. Sci. USA, 108:(2011) 14713–14714.

A. Pastore and V. Saudek. *The relationship between chemical shift and secondary structure in proteins*. J. Magn. Reson., 90:(1990) 165–176.

A. Pertsemlidis, J. Zelinka, J. W. Fondon, R. K. Henderson, and Z. Otwinowski. *Bayesian statistical studies of the Ramachandran distribution*. Stat Appl Genet Mol Biol, 4:(2005) Article35.

J. W. Ponder and D. A. Case. *Force fields for protein simulations*. Adv. Protein Chem., 66:(2003) 27–85.

L. L. Porter and G. D. Rose. *Redrawing the Ramachandran plot after inclusion of hydrogen-bonding constraints*. Proc. Natl. Acad. Sci. USA, 108:(2011) 109–113.

W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge UK, 1989.

J. M. Pryce and A. D. Bruce. *Statistical mechanics of image restoration*. Journal of Physics A, 28(3):(1995) 511.

E. M. Purcell, H. C. Torrey, and R. V. Pound. *Resonance absorption by nuclear magnetic moments in a solid*. Phys. Rev., 69:(1946) 37–38.

L. Rabiner. *A tutorial on hmm and selected applications in speech recognition*. Proceedings of the IEEE, 77(2):(1989) 257–286.

G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. *Stereochemistry of polypeptide chain configurations*. J. Mol. Biol., 7:(1963) 95–99.

D. Reith, M. Pütz, and F. Müller-Plathe. *Deriving effective mesoscale potentials from atomistic simulations*. Journal of computational chemistry, 24(13):(2003) 1624–1636.

G. Rickayzen and J. G. Powles. *Temperature in the classical microcanonical ensemble*. The Journal of Chemical Physics, 114:(2001) 4333.

W. Rieping, M. Habeck, and M. Nilges. *Inferential Structure Determination*. Science, 309:(2005) 303–306.

W. Rieping and W. F. Vranken. *Validation of archived chemical shifts through atomic coordinates*. Proteins, 78:(2010) 2482–2489.

C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. *Protein structure prediction using Rosetta*. Meth. Enzymol., 383:(2004) 66–93.

B. Rost. *Protein secondary structure prediction continues to rise*. Journal of Structural Biology, 134:(2001) 204–218.

B. Rost and S. I. O'Donoghue. *Sisyphus and protein structure prediction.* Computer Applications in the Biosciences, 13:(1997) 345 – 356.

B. Rost and C. Sander. *Prediction of protein secondary structure at better than 70% accuracy*. J. Mol. Biol., 232(2):(1993) 584–599.

S. Roweis and Z. Ghahramani. *A unifying review of linear gaussian models*. Neural computation, 11(2):(1999) 305–345.

H. H. Rugh. *Dynamical approach to temperature*. Phys. Rev. Lett., 78:(1997) 772–774.

D. Rykunov and A. Fiser. *New statistical potential for quality assessment of protein models and a survey of energy functions*. BMC bioinformatics, 11(1):(2010) 128.

J. J. Sakurai, S.-F. Tuan, and E. D. Commins. *Modern quantum mechanics*. American Journal of Physics, 63:(1995) 93.

R. Samudrala and M. Levitt. *Decoys 'r' us: a database of incorrect conformations to improve protein structure prediction*. Protein Science, 9(7):(2000) 1399–1401.

R. Samudrala and J. Moult. *An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction*. Journal of molecular biology, 275(5):(1998) 895–916.

A. Savelyev and G. A. Papoian. *Chemically accurate coarse graining of double-stranded dna*. Proceedings of the National Academy of Sciences, 107(47):(2010) 20340–20345.

G. Schwarz. *Estimating the dimension of a model*. The Annals of Statistics, 6:(1978) 461–464.

S. A. Shahid, B. Bardiaux, W. T. Franks, L. Krabben, M. Habeck, B. J. van Rossum, and D. Linke. *Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals*. Nat. Methods.

M. S. Shell. *The relative entropy is fundamental to multiscale and inverse thermodynamic problems*. The Journal of chemical physics, 129:(2008) 144108.

M.-y. Shen and A. Sali. *Statistical potential for assessment and prediction of protein structures*. Protein science, 15(11):(2006) 2507–2524.

Y. Shen, F. Delaglio, G. Cornilescu, and A. Bax. *TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts*. J. Biomol. NMR, 44:(2009) 213–223.

Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker, and A. Bax. *Consistent blind protein structure generation from NMR chemical shift data*. Proc. Natl. Acad. Sci. USA, 105:(2008) 4685–4690.

D. Shortle. *Propensities, probabilities, and the boltzmann hypothesis*. Protein science, 12(6):(2003) 1298–1302.

M. J. Sippl. *Knowledge-based potentials for proteins*. Curr. Opin. Struct. Biol., 5:(1995) 229–235.

D. Sivia and J. Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, USA, 2nd edition, 2006.

J. Skilling. *Nested sampling*. AIP Conference Proceedings, 735(1):(2004) 395–405.

J. Skolnick. *In quest of an empirical potential for protein structure prediction*. Curr. Opin. Struct. Biol., 16:(2006) 166–171.

J. Sohl-Dickstein and B. J. Culpepper. *Hamiltonian annealed importance sampling for partition function estimation*. CoRR, abs/1205.1925.

L. Song, B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola. *Hilbert space embeddings of hidden markov models*. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 991–998. 2010.

A. Soper. *Empirical potential monte carlo simulation of fluid structure*. Chemical Physics, 202(2):(1996) 295–306.

J. M. Sorenson and T. Head-Gordon. *Toward minimalist models of larger proteins: A ubiquitin-like protein*. Proteins: Structure, Function, and Bioinformatics, 46(4):(2002) 368–379.

R. Sprangers, A. Velyvis, and L. E. Kay. *Solution nmr of supramolecular complexes: providing new insights into function*. Nature methods, 4(9):(2007) 697–703.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. *Gaussian process optimization in the bandit setting: No regret and experimental design*. arXiv preprint arXiv:0912.3995.

Y. Sugita, A. Kitao, and Y. Okamoto. *Multidimensional replica-exchange method for free-energy calculations*. The Journal of Chemical Physics, 113:(2000) 6042.

R. H. Swendsen and J.-S. Wang. *Replica Monte Carlo simulation of spin glasses*. Phys. Rev. Lett., 57:(1986) 2607–2609.

S. Tanaka and H. A. Scheraga. *Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins*. Macromolecules, 9(6):(1976) 945–950.

J. M. Thompson, N. G. Sgourakis, G. Liu, P. Rossi, Y. Tang, J. L. Mills, T. Szyperski, G. T. Montelione, and D. Baker. *Accurate protein structure modeling using sparse nmr data and homologous structure information*. Proceedings of the National Academy of Sciences, 109(25):(2012) 9875–9880.

D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. L. Dunbrack. *Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model*. PLoS Comput. Biol., 6(4):(2010) e1000763.

C. Tsallis. *Possible Generalization of Boltzmann-Gibbs Statistics*. J. Stat. Phys., 52:(1988) 479–487.

G. Vriend. *WHAT IF: a molecular modeling and drug design program*. J. Mol. Graph., 8:(1990) 52–56.

C. Wang, N. Komodakis, and N. Paragios. *Markov random field modeling, inference & learning in computer vision & image understanding: A survey*. Computer Vision and Image Understanding.

C. C. Wang, J. H. Chen, W. C. Lai, and W. J. Chuang. *2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts*. J. Biomol. NMR, 38:(2007a) 57–63.

F. Wang and D. P. Landau. *An efficient, multiple range random walk algorithm to calculate the density of states*. Phys. Rev. Lett., 86:(2001) 2050–2053.

L. Wang and B. R. Donald. *Analysis of a systematic search-based algorithm for determining protein backbone structure from a minimum number of residual dipolar couplings*. Proc IEEE Comput Syst Bioinform Conf, pages 319–330.

L. Wang, H. R. Eghbalnia, and J. L. Markley. *Nearest-neighbor effects on backbone alpha and beta carbon chemical shifts in proteins*. J. Biomol. NMR, 39:(2007b) 247–257.

T. Wang and R. C. Wade. *Force field effects on a β-sheet protein domain structure in thermal unfolding simulations*. Journal of Chemical Theory and Computation, 2(1):(2006) 140–148.

Y. Wang and O. Jardetzky. *Probability-based protein secondary structure identification using combined NMR chemical-shift data*. Protein Sci., 11:(2002) 852–861.

M. Wells, H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge, and A. R. Fersht. *Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain.* Proc Natl Acad Sci U S A, 105(15):(2008) 5762–5767.

M. P. Williamson, T. F. Havel, and K. Wüthrich. *Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry*. Journal of molecular biology, 182(2):(1985) 295–315.

M. P. Williamson and V. S. Madison. *Three-dimensional structure of porcine $C5_{\mathrm{desarg}}$ from ^{1}H nuclear magnetic resonance data*. Biochemistry, 29:(1990) 2895–2905.

D. S. Wishart and D. A. Case. *Use of chemical shifts in macromolecular structure determination*. Meth. Enzymol., 338:(2001) 3–34.

D. S. Wishart and B. D. Sykes. *The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data*. J. Biomol. NMR, 4:(1994) 171–180.

L. Wroblewska and J. Skolnick. *Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? i. large scale amber benchmarking*. Journal of computational chemistry, 28(12):(2007) 2059–2066.

K. Wüthrich. *NMR of Proteins and Nucleic Acids*. John Wiley, New York, 1986.

J. Yvon. *La théorie statistique des fluides et l'équation d'état*, volume 203. Hermann & cie, 1935.

A. Zemla, C. Venclovas, K. Fidelis, and B. Rost. *A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment*. Proteins: Structure, Function, and Genetics, 34:(1999) 220–223.

J. Zeng, P. Zhou, and B. Donald. *Protein side-chain resonance assignment and NOE assignment using RDC-defined backbones without TOCSY data*. Journal of Biomolecular NMR, 50:(2011) 371–395.

J. Zhang and Y. Zhang. *A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction*. PloS one, 5(10):(2010) e15386.

H. Zhou and Y. Zhou. *Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction*. Protein science, 11(11):(2002) 2714–2726.

Z. Zhou, R. N. Leahy, and J. Qi. *Approximate maximum likelihood hyperparameter estimation for Gibbs priors*. IEEE Transactions on Image Processing, 6(6):(1997) 844–861.