Fakultät für Informatik

Lehrstuhl für Bioinformatik / I12

Bayesian structure determination from Chromosome Conformation Capture data and avenues to improve conformational sampling

Simeon Carstens

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. Nils Thuerey

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Burkhard Rost
2. Priv.-Doz. Dr. Michael Nilges

Die Dissertation wurde am 16.12.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 08.02.2016 angenommen.

Abstract

Recent developments in experimental techniques have opened the door to studies of the spatial organisation of genomes at an unprecedented level of detail. For a long time, a gap in knowledge existed between genome structure at sub-kilobase resolution and at length scales accessible to optical microscopy. This gap is now filled with Chromosome Conformation Capture (3C)-based techniques, in which contact frequencies between loci are measured.

The 3D structure of the genome is of obvious biological interest, as it has been known to be closely linked to its function. This is in complete analogy to the three-dimensional structure of other biological macromolecules such as proteins and RNA. Obtaining a three-dimensional structure from experimental data is a difficult task in structural biology: experimental data is noisy and the exact physical processes which lead from structure to data are unknown. An elegant framework is Inferential Structure Determination (ISD), in which structure determination is seen as a problem of statistical inference. The central object in ISD is a complicated posterior probability distribution, which incorporates both data and prior information about the structure. Sampling structures and other parameters from it is the ultimate and computationally expensive step of structural modeling using ISD and results in statistically well-defined structural ensembles and modeling parameter estimates.

The purpose of this work is two-fold: investigating algorithms to enhance sampling from difficult probability distributions such as the ISD posterior and extending the scope of ISD to chromatin structure determination from high-throughput 3C (HiC) data.

I apply a recently proposed method for sampling from difficult probability dis-

tributions, Replica Exchange with Non-equilibirum switches (RENS), on protein models of different complexity. The goal is to assess the computational efficiency of RENS in complex, realistic applications such as ISD. Currently, ISD implementations employ a related, wide-spread technique called Replica Exchange (RE). While I am able to show that RENS indeed improves sampling compared to RE, the main result is that the large amount of additional computational time required by RENS renders it much less efficient.

RE and related algorithms simulate not only the target probability distribution, but also a sequence of copies of it, which are easier to sample and, by exchanging configurations, thus enhance sampling of the target distribution. An idea of an iterative scheme to optimize the choice of this sequence is presented. I test a preliminary implementation of this approach on a simple system and find that, at least in our example, sampling quality is not improved and even inferior to a simple, heuristic sequence of distributions.

Turning to applications of ISD, I demonstrate that ISD also is a suitable method to infer 3D structures of a single chromosome from both sparse single single-cell and, by using simulated contact frequencies, probably also from rich, population-based HiC data.

Zusammenfassung

In letzter Zeit haben Fortschritte in der Entwicklung experimenteller Techniken die Tür zur Erforschung der räumlichen Struktur von Genomen in bisher unerreichtem Detail weit aufgestoßen. Lange Zeit war wenig über die Strukur von Genomen zwischen Auflösungen von unter einigen kb und mit optischer Mikroskopie zugänglichen Längenskalen bekannt. Diese Lücke wird nun von Chromosome Conformation Capture (3C)-basierten Methoden geschlossen, mit Hilfe derer Kontaktfrequenzen zwischen Loki gemessen werden.

Die dreidimensionale Struktur eines Genoms ist von offensichtlichem Interesse für die Biologie, da bekannt ist, dass sie eng mit der Funktion des Genoms zusammenhängt. Dies ist analog zur räumlichen Struktur von Proteinen, RNA und anderen biologischen Makromolekülen. Die Bestimmung einer dreidimensionalen Struktur anhand von experimentellen Daten ist ein schwieriges Problem der Strukturbiologie: experimentelle Daten sind zwangsläufig verrauscht, meist unvollständig und die genauen physikalischen Prozesse, durch die ein Rückschluss von den Daten auf die Struktur ermöglicht wird, sind nicht bekannt. Ein eleganter Ansatz, dieses Problem zu lösen, ist das Prinzip der Inferentiellen Strukturbestimmung (ISD, Inferential Structure Determination). In dieser Methode wird Strukturbestimmung als ein Inferenzproblem betrachtet. Der wichtigste Bestandteil dieses Ansatzes ist eine komplizierte *a posteriori*-Wahrscheinlichkeitsverteilung, die Informationen aus den Daten und Vorwissen über die Struktur verbindet. Aus ihr werden sowohl Strukturen als auch weitere unbekannte Parameter, die zur Strukturbestimmung nötig sind, gezogen. Dies ist der letzte und rechentechnisch aufwändigste Schritt einer Strukturbestimmung mittels ISD und mündet in statistisch wohldefinierten

Struktur-Ensembles und Schätzungen der weiteren Modellparameter.

Die vorliegende Arbeit hat zwei Dinge zum Ziel: zum einen sollen Algorithmen zum effizienteren Ziehen von Stichproben aus komplexen Wahrscheinlichkeitsverteilungen erforscht werden und zum anderen der Anwendungsbereich von ISD auf die Bestimmung der Struktur von Chromatin auf der Basis von genomweiten 3C (HiC)-Daten erweitert werden.

Ich wende eine vor relativ kurzer Zeit vorgeschlagene Methode namens Replica Exchange with Non-equilibrium Switches (RENS) zum Ziehen von Strichproben aus Wahrscheinlichkeitsverteilungen auf Proteinmodelle verschiedener Komplexität an. Ziel ist es, die Effizienz von RENS hinsichtlich der Rechenzeit in realistischen Anwendungen wie ISD zu beurteilen. ISD-Implementationen benutzen derzeit Replica Exchange (RE), eine weitverbreitete, verwandte Methode. Zwar kann ich zeigen, dass RENS im Vergleich zu RE repräsentativere Stichproben zieht, dazu allerdings unverhältnismäßig viel Rechenzeit benötigt und somit deutlich ineffizienter als RE ist.

RE und verwandte Algorithmen simulieren nicht nur eine Zielwahrscheinlichkeitsverteilung, sondern auch eine Reihe von Kopien, aus denen das Ziehen von Stichproben einfacher ist. Das gelegentliche Austauschen von Konfigurationen beschleunigt dann das Ziehen von repräsentativen Stichproben aus der Zielverteilung. Ich stelle einen Ansatz für ein iteratives Schema zum Bestimmen einer optimalen Sequenz von interpolierenden Verteilungen vor und teste ihn an einem einfachen System. Für dieses Beispiel stellt sich allerdings heraus, dass die Qualität der gezogenen Stichproben geringer ist als wenn eine einfache, heuristische Reihe von interpolierenden Wahrscheinlichkeitsverteilungen benutzt wird.

Schließlich wende ich mich einer neuen Anwendung von ISD zu und zeige, dass ISD auch eine gut geeignete Methode zur Inferenz von 3D-Strukturen einzelner Chromosomen anhand von spärlichen HiC-Daten einzelner Zellen ist. Mittels simulierter Kontaktfrequenzen demonstriere ich weiterhin, dass ISD prinzipiell auch die Bestimmung von Strukturen anhand von wesentlich ergiebigeren populationsbasierten HiC-Daten ermöglichen sollte.

Contents

Introduction

A majority of the efforts in structure determination of biological macromolecules has, in the past decades, mainly been spent on finding native conformations of proteins: biophysical methods like nuclear magnetic resonance (NMR, see Wüthrich [2001] for a historical perspective), X-ray crystallography (starting with the structures of myoglobin [Kendrew et al., 1958] and hemoglobin [Perutz et al., 1960], see, e.g., Shi [2015] for a review) and electron microscopy (EM, Adrian et al. [1984]; Kühlbrandt [2014]; Unwin and Henderson [1975]; now reaching atomic resolution [Bartesaghi et al., 2015]), to name the three most important ones, have yielded over 100000 structures of proteins and complexes with other proteins and / or nucleic acids deposited in the Protein Data Bank [Berman et al., 2000]. These fuel our understanding of biological processes at an atomistic scale. The knowledge of how proteins function on a near-atomic level has not only brought great advances in numerous fields of biology, but also forms the basis of structure-based drug design.

Until recently, the genome, on the other hand, did not enjoy the same attention of structural biology, and the three-dimensional structure of chromosomes, let alone whole genomes, has eluded our knowledge. While coarse-grained structural information is available through a range of techniques like FISH imaging [Bauman et al., 1980; Hulspas and Bauman, 1992] or molecular biology methods to measure interaction of DNA with other cell constituents, no method offering a sufficient resolution to determine the fold of chromosomes was available. A major advance thus came from recently introduced chromosome conformation capture (3C) based techniques [Dekker et al., 2002; Dostie et al., 2006; Lieberman-Aiden et al., 2009; Simonis et al., 2006; Zhao et al., 2006], which

rely on crosslinking chromatin and subsequent next-generation sequencing of crosslinked DNA fragments to determine contact frequencies between a great number of loci. Depending on the exact method and the sequencing depth, these methods can reach resolutions of up to 1 kb [Rao et al., 2014], which allows, for example, to probe the spatial interaction of regulatory elements and thus opens the door to a microscopic view of the genome as a highly organized set of polymers for whose function (or malfunction) the three-dimensional fold and arrangement of chromosomes is of paramount importance [Hughes et al., 2014; Lupiáñez et al., 2015; Pombo and Dillon, 2015].

Recent genome-wide 3C experiments on single cells [Nagano et al., 2013] showed that, although conserved structural domains exist on larger scales, genome architecture is highly variable from cell to cell. This has important implications for the interpretation of data of 3C experiments performed on a cell population, which results in average contact frequencies not of one molecule for each chromosome, but of millions. The high cell-to-cell variability and, in the case of single cell HiC data, the sparseness of the data, make it clear that any unjustified assumption about parameters of the process leading from measurement to final 3D structures is likely to have a strong influence on the result. This is analogous to the determination of protein structures: the physics by which we describe a chain of amino acids and the theory we have about, for example, the relation between interatomic distances and the intensity of peaks in a NMR spectrum, is only approximate and also the experimental noise is not known. Robust and objective methods are thus required to find meaningful and honest estimates of the actual biological structure and to quantify its uncertainty, be it a protein or a chromosome. To this end, Rieping, Habeck, and Nilges [2005a] developed a probabilistic approach to structure determination termed Inferential Structure Determination (ISD; discussed in Sec. 1.2) to eliminate heuristics and other biases in previous approaches. In ISD, macromolecular structure determination is viewed as a problem of statistical inference from noisy and incomplete data. Prior knowledge about the structure (from, e.g., physics) and unknown modeling parameters is encoded in a *prior* probability and, after measuring the data, updated by incorporating the new information

by means of the *likelihood*, another probability. Prior and likelihood are multiplied using Bayes' theorem $P(A|B) \propto P(B|A)P(A)$ to yield a joint probability for all unknown parameters called the *posterior*.

As one is usually concerned with a continuum of possible structures and unknown modeling (*nuisance*) parameters, all probabilities turn into probability distributions and in order to not only find the most likely structure, but also to get an estimate of its uncertainty, the practicioner needs to sample from the posterior distribution. It is in general high-dimensional and the random variables it describes are highly correlated. For these reasons, sampling from it is very difficult and requires advanced Markov Chain Monte Carlo (MCMC) techniques, to which an introduction is given in Sec. 1.3.

In light of these considerations, the work presented here is centered on testing and improving methods for sampling difficult probability distributions and on the extension of the scope of the ISD framework from protein structure determination to a more objective estimation of chromosome structures from single cell and population HiC data.

To enhance MCMC sampling it is advisable to employ Replica Exchange (RE; Geyer [1991]; Swendsen and Wang [1986]) methods which not only simulate the posterior distribution of interest, but also by some transformation "flattened" versions of it, which, by exchanging configurations, prevent the simulation of the target distribution from getting stuck in modes. In Sec. 2, we test Replica Exchange with Non-equilibrium Switches (RENS; Ballard and Jarzynski [2009, 2012]), a recently proposed variant of RE. It relies on making exchange candidate states more likely in transformed distributions by means of non-equilibrium trajectories. We test RENS, which has so far only been tested on systems of low dimensionality, on complex protein systems.

Sec. 3 presents an idea and first tests of an iterative method to automatically determine optimally interpolating distributions for RE-based MCMC schemes. It relies on calculation of the density of states of the system of interest and, as a side effect, thus may prove useful for a wide range of applications relying on approximative knowledge of this important quantity such as free energy or evidence calculations.

Turning to applications of ISD to HiC data, we propose a Bayesian determination of chromatin structures from single cell data to improve on previous, minimization-based approaches and a possible extension of the ISD approach to explicit ensemble-based modeling from population HiC data in Sec. 4.
Finally, Sec. 5 summarizes the present work and discusses perspectives and open questions.

All parts of this work are, in one sense or another, intertwined. Advanced RE sampling schemes enhance the sampling of complex posterior distributions, which is of great importance, as dimensionality of the structure determination problem is likely to increase with future applications of ISD to population HiC data (Sec. 4.4) and practical applicability of ISD thus requires an efficient use of computational resources. By means of histogram reweighting techniques [Chodera et al., 2007; Ferrenberg and Swendsen, 1988; Habeck, 2012a], optimized RE(NS) simulations can prove useful for efficient Bayesian model comparision; a method allowing to compare different models in light of the data. Since RENS, as will be discussed later, is effective, but not very efficient, it is in dire need of optimized schedules in order to calculate as few non-equilibrium trajectories as possible.

1 Computational methods for structure determination

The acquisiton of three-dimensional genomic structures from 3C-based data greatly profited from computational modeling methods already known from protein structure determination. This step is crucial, as neither the biophysical experiments performed to determine protein structures nor the contact frequency maps obtained from genome-wide 3C (HiC, Lieberman-Aiden et al. [2009]) directly result in a three-dimensional model of the macromolecule under consideration. Instead, one has to resort to computational methods to find (possibly coarse-grained) structures fitting the observed data. This requires a method to judge whether a structure agrees with the data and thus, knowledge about the physical processes leading from the real structure to the data is essential. Often, though, the data does not give sufficient information to determine a structure to a reasonable degree. For this reason, one usually demands a candidate structure to not only fit the data, but also prior information already known before performing the experiment. Prior information can be gleaned from the physics or chemistry of the biomolecule. We know, for example, that bond lengths in a protein are, to a reasonable approximation, fixed [Leach, 2001] and we can estimate electrostatic interactions between residues. Information of course may also come from biology. The mere fact that the genome is contained in either the nucleus of a eukaryotic cell or in prokaryotic cell may sound trivial, but is actually valuable information for whole genome modeling, as done, for example, in [Kalhor et al., 2012]. Furthermore, we know from FISH imaging experiments that eukaryotic interphase chromosomes do

not intermingle much, but instead occupy distinct territories [Cremer and Cremer, 2010]. These are only a few examples of prior information available for computational modeling of genome structures. A candidate structure could also be asked to be reasonably similar to a homologuous molecule. The complexity and heterogeneity of both data and prior information thus make finding structures which fulfill both the restraints given by the data and comply with the prior knowledge a difficult problem of computational biology.

We want to formalize the process of structure determination and thus introduce some terminology. We denote a candidate structure by a vector $\mathbf{x}$, containing the positions of atoms or distinct units of coarse-graining in some coordinate system. For proteins, one often chooses internal coordinates, in which the polymer chain is described in terms of bond lengths, bond angles and dihedral angles (internal coordinates, Leach [2001]). This coordinate system is more adapted to the geometry of a polymer chain and allows to decouple fast-changing degrees of freedom (bond lengths) from slow ones. Furthermore, computation time can be saved by keeping bond lengths and bond angles fixed and thus reducing the number of degrees of freedom. On the other hand, it is more complicated to calculate distances between distant units, which involves transforming internal to cartesian coordinates.

A function $f(\mathbf{x};\alpha) = \hat{D}$ back-calculating mock data $\hat{D}$ from the candidate structure coordinates possibly parametrized by other modeling parameters α is called a *forward model*. Because of our limited knowledge of the physical processes leading to an experimental outcome, the forward model is usually only an approximation to physical reality.

For this reason and because of unevitable measurement errors and possibly incomplete data, we allow deviations of the mock data from the experimentally measured values. Physical or non-physical prior information about the sought-for structure is encoded in a (possibly effective) potential energy function $E_{\text{prior}}(\mathbf{x})$. This prior information scoring function will attain an extremum for a structure fitting the prior constraints best.

We now introduce two very different approaches to structure determination, namely the idea of conventional scoring function optimization and ISD [Rieping,

Habeck, and Nilges, 2005a], an alternative based on Bayesian inference.

1.1 Minimization-based approaches

One approach to finding a three-dimensional structure $\mathbf{x}$ from data D and prior information encoded in the energy $E_{\text{prior}}(\mathbf{x})$ is to set up a data scoring function and to combine it with the prior scoring function in a total score given by

$$E_{tot}(\mathbf{x}; w_{\text{data}}, \alpha) = w_{\text{data}} E_{\text{data}}(\mathbf{x}; \alpha) + E_{\text{prior}}(\mathbf{x}) \tag{1.1}$$

with E_{data} the data scoring function attaining its minimum at the back-calculating data matching the experimental data best, but allowing some degree of deviation by, e.g. a harmonic restraint. w_{data} is a weighting factor which weighs the data term against the prior information. If we believe the data to be of bad quality, we should pick a low value for w_{data}.

It is common practice to set w_{data} and α to values determined by heuristics or cross-validation [Brunger et al., 1993; Brunger, 1992]. We can then numerically minimize this function using optimization algorithms. Simulated annealing (SA, Kirkpatrick et al. [1983]; Černý [1985]) is especially popular in protein structure determination software like ARIA [Linge et al., 2003; Rieping et al., 2007], CNS [Brünger, 2007; Brünger et al., 1998] and CYANA [Güntert et al., 1997; Güntert, 2004; Güntert et al., 1991; Herrmann et al., 2002; López-Méndez and Güntert, 2006], but is also often employed in chromatin structure determination from chromosome conformation capture data, for example in [Kalhor et al., 2012; Nagano et al., 2013]. Inspired by the slow annealing of a melt to form crystals with as few defects as possible, the system consisting of the structural coordinates changing under the influence of a potential energy given by total scoring function is first optimized at a high (non-physical) "temperature" and then slowly cooled until the system "freezes" in the global minimum. Optimization at a certain temperature is done either by Molecular Dynamics (MD) or MCMC algorithms (Sec. 1.3), which makes sure the system does not get stuck in local minima by also allowing it to move to a certain

extent in regions with unfavorable scoring function values.

Minimization by SA, given sufficiently slow cooling, is guaranteed to locate the global minima of the scoring function [Robert and Casella, 2011] and thus results in structures fitting both the data and the prior constraints. But the inverse problem we solved numerically is underdetermined: the data will in general be not sufficient to uniquely determine an optimal structure. While the prior energy adds additional constraints, the relative weight between the prior energy and the data energy is unknown. There might be several, equally optimal structures, but the setup of the minimization algorithm might cause it to find not all of them. Repeating the optmiziation several times from different initial values for the coordinates will, depending on the amount of data and prior information, give different structures with comparably good scores. But the resulting set of structures is not statistically well-defined and will in general depend on parameters of the minimization procedure. By setting forward model parameters and the weight *a priori* and minimizing the resulting scoring function, we thus obtain biased structures without any meaningful measure of uncertainty.

1.2 Inferential Structure Determination

Both the problem of determining uncertainty and the choice of weights and forward model parameters, which is cumbersome at best, and at worst, when cross-validation is too time-consuming or instable, highly subjective, can be avoided elegantly.

In the Inferential Structure Determination (ISD; Rieping, Habeck, and Nilges [2005a]) framework, determination of macromolecular structures is instead viewed as a problem of statistical inference. We would like to consistently quantify our knowledge about a structure $\mathbf{x}$ given in general incomplete and noisy data D and any information I we already have about the structure in question. Cox [1946] proved that probability theory is the only way to consistently quantify uncertainty. Before having measured the data D, the belief in

the proposition "in the experimental conditions, the molecule's structure was $\mathbf{x}$" is given by *prior* knowledge I and thus quantified by the prior probability $P(\mathbf{x}|I)$. After having measured D, we possess more information about $\mathbf{x}$ and we have to update our belief using the new information. The updated belief is then quantified through the probability $P(\mathbf{x}|D, I)$. Bayes' theorem now allows us to decompose this *posterior* (to measuring the data) probability by writing

$$P(\mathbf{x}|D, I) = \frac{P(D|\mathbf{x}, I)P(\mathbf{x}|I)}{P(D)} ,$$
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} ,$$

where we recognize the prior probability $P(\mathbf{x}|I)$ and introduce two new probabilities. $P(D|\mathbf{x}, I)$ is the likelihood of the data D given the structure $\mathbf{x}$, that is, the quantification of our belief in the proposition "if, at the time of measurement, the structure had been $\mathbf{x}$, we would have obtained the data D". The probability $P(D|I)$ normalizes the posterior distribution and, as the expectation value of the likelihood under the prior distribution, quantifies the belief in being able to obtain the data D given the prior information I and our choice of the likelihood. This is called *evidence*.

The likelihood $P(D|\mathbf{x}, I)$ is usually a composition of the aforementioned forward model, which back-calculates "mock" data $\hat{D}$ from a candidate structure $\mathbf{x}$, and a probability $g(\hat{D}, \sigma|D)$, which turns the discrepancy between $\hat{D}$ and D into a number $\in [0, 1]$. The latter is called *error model* and is parameterized by errors σ that specify the total error, which includes both experimental noise and the error we neccessarily introduce by imperfectly back-calculating data from a structure by means of an approximate forward model.

But, as mentioned before, the structure $\mathbf{x}$ is not the only unknown in a structure determination problem. The forward model possibly depends on unknown parameters α and we are not sure about the errors σ. This means that, for a fully probabilistic treatment of the structure determination problem, we have to expand the hypothesis space and, although they are of secondary interest and thus termed *nuisance parameters*, also regard α, σ as unknown modeling parameters; giving them the same importance as the structure $\mathbf{x}$ and estimating them from the data and prior information. Taken together; the likelihood

thus has to be written as

$$L(D|\mathbf{x}, \alpha, \sigma, I) = g\left[f(\mathbf{x}; \alpha); \sigma\right]$$

and is for our purposes often regarded as a function $L(\mathbf{x}, \alpha, \sigma)$ of the model parameters, in which case it is called the *likelihood function*.
Even the prior probability distribution might not be completely specified and depend on *hyper parameters* γ. We take this fact into account by adding γ to the set of parameters which need to be estimated along with the structure, so that finally we are quantifying not only our uncertainty about the structure, but about *all* unknown parameters in a joint posterior probability

$$P(\mathbf{x}, \alpha, \sigma, \gamma|D, I) = \frac{P(D|\mathbf{x}, \alpha, \sigma, I)P(\mathbf{x}, \alpha, \sigma, \gamma|I)}{P(D|I)} \ .$$

Under the assumption that structure and nuisance parameters are independent from each other, the prior factorizes. It is important to note that this is not always the case in structure determination. If we were, for example, to infer a protein structure from X-ray data, the unknown phase information might play into the forward model and, while being a nuisance parameter, depends on the structure. In the following, we will also be concerned not with a discrete set of structures and nuisance parameters our belief in we test, but a continuum and thus introduce probability densities instead of probabilities. The posterior distribution then becomes

$$p(\mathbf{x}, \alpha, \sigma, \gamma|D, I) = \frac{p(D|\mathbf{x}, \alpha, \sigma, I)p(\mathbf{x}|I)p(\alpha|I)p(\sigma|I)p(\gamma|I)}{P(D|I)} \ . \tag{1.2}$$

While specifying prior distributions and error models, in order to remain objective, we have to watch out not to make unjustified assumptions. For this reason, we assume minimally informative prior distributions which reflect only the information we actually possess. This can be rigorously implemented by

following the Maximum Entropy (ME) principle [Jaynes, 1957]. In this framework, the information a probability distribution $p(x)$ contains about a random variable x is measured by the Shannon entropy $I = -\int \mathrm{d}x\ p(x) \log p(x)$ and we seek the probability distribution maximizing the Shannon entropy under the constraint of being true to the prior knowledge I. One trivial constraint is that any probability distribution should be normalized, $\int \mathrm{d}x\ p(x) = 1$. An example for this construction is the prior distribution for the structure $\mathbf{x}$. Suppose we have only physical prior information. If we knew the exact potential energy function $E(\mathbf{x})$ of the molecule under consideration and neglect other interactions with, for example, other cell constituents or a solvent, then, for a measurement performed at a fixed temperature T, ME yields the canoncial ensemble

$$p(\mathbf{x}|I) = \frac{1}{Z(\beta)} \mathrm{e}^{-\beta E(\mathbf{x})} \tag{1.3}$$

with the inverse temperature $\beta = {}^{1}\!/_{k_\mathrm{B}T}$ as the minimally informative prior distribution under the constraints $\langle E(\mathbf{x}) \rangle = \int \mathrm{d}\mathbf{x}\ p(\mathbf{x}) E(\mathbf{x})$ and $\int \mathrm{d}\mathbf{x}\ p(\mathbf{x}) = 1$. In reality, though, molecular force fields are always approximative and the inverse temperature β should rather be regarded as a weight for the information encoded in the potential energy. Thus β will not exactly correspond to the experimental inverse temperature, but can be estimated from the data by model comparison [Mechelke and Habeck, 2012].

We also need unbiased nuisance parameter prior distributions. If we limit ourselves to error models of the form

$$g(\hat{D}, \sigma|D) = \frac{1}{Z(\sigma)^N} \mathrm{e}^{-\frac{1}{2\sigma^2}\chi^2(\hat{D};D)}\ , \tag{1.4}$$

with N denoting the total number of data points and χ^2 the total deviation of the back-calculated from the experimental data, we do not need to invoke the Maximum Entropy principle if we notice that the error σ does not have an absolute meaning: its value can only be interpreted in conjunction with χ^2, as any rescaling of the former can be compensated by rescaling the latter. The

error σ and other parameters with the same behaviour are thus called *scale parameters*. If we know that a random variable x is a scale parameter, its probability distribution $p(x)$ must be invariant under a scale transformation; that is, we ask

$$p(x)\,\mathrm{d}x = p(\tau x)\,\mathrm{d}(\tau x)\ .$$

This condition is, by the substitution rule for probability densities, sufficient to find

$$p(x) \propto \frac{1}{x}\ . \qquad \text{Jeffreys prior (1.5)}$$

Note that the Jeffreys prior is an improper prior, that is, it is not normalizable. But this normally does not pose a problem, as the posterior distribution will usually nevertheless be well-behaved. In ab-initio structure determination the prior information comes from physics. If we are looking for a protein structure, we already know the sequence of amino acids and thus have prior information about van der Waals- and electrostatic interactions. For any polymer, by definition the distance between one monomer and its neighbor is limited and it is also reasonable to assume that monomers do not overlap. But other kinds of prior information can be imagined. When looking for a protein structure, one might already have information about the structure of a homologuous protein and homology modeling could complement the measured data. Other prior information could, for example, come from evolutionary contacts (see, e.g., Hopf et al. [2014] for evolutionary contacts applied to protein complexes). In any case, it is not obvious how to weight prior information. Mechelke and Habeck [2014] have developed a method to include statistical knowledge-based potentials as prior information in NMR protein structure determination and determine their weight from the experimental data.

After specifying the prior distributions and the likelihood, the problem of inferring unknown structures and nuisance parameters is formally solved. The maximum of the posterior distribution (Eq. 1.2) are the most likely structure and nuisance parameters given the data and the prior information, so a maximum a-posteriori (MAP) estimate would give the sought answer to the

structure determination problem, but no information about the uncertainty of the estimate. A "structural error bar" can only be obtained by sampling from the posterior distribution, which is in general difficult. By drawing samples from the posterior distribution, we can obtain an ensemble of structures and nuisance parameters in which each ensemble member has its own associated posterior probability weight. In practice, though, it is difficult to sample from this distribution and one has to resort to Markov Chain Monte Carlo (MCMC) methods discussed below, which allow to approximate the posterior distribution by samples drawn using a random process.

It should be noted that it is not neccessary to stick with the full posterior distribution. Nuisance parameters can be integrated out (*marginalized*), reducing the dimensionality of the posterior distribution. Given a joint distribution $p(x, y)$ for two random variables, the the marginal distribution for x is given by

$$p(x) = \int \mathrm{d}y \; p(x, y) = \int \mathrm{d}y \; p(x|y)p(y) \; ,$$

and contains all information about y encoded in $p(x, y)$. This is very different from setting y to a fixed value y_0, which would correspond to assigning $p(y) = \delta(y - y_0)$ and thus $p'(x) = \int \mathrm{d}y \; p(x|y)\delta(y - y_0)$. Marginalization can be done analytically for the errors σ of several error models, such as a Gaussian or log-normal distribution.

As already mentioned, the MAP estimate of the ISD posterior distribution yields an objective estimate of the unknown structure. This fact can be exploited to replace the heuristic, biased scoring function in conventional minimization-based approaches by a Bayesian one given by the negative logarithm of the posterior distribution [Nilges et al., 2008]. If we assume a posterior with an error model of the form described in Eq. 1.4, and, other than the error σ, no further nuisance parameters, the negative logarithm of the posterior is (neglecting constant terms) given by

$$\begin{aligned} -\log p(\mathbf{x}, \sigma|D, I) &= \frac{1}{\sigma^2}\frac{\chi^2(\mathbf{x}; D)}{2} + \beta E_{\mathrm{prior}}(\mathbf{x}) + N \log Z(\sigma) - \log p(\sigma) \\ &= w_{\mathrm{data}} E_{\mathrm{data}}(\mathbf{x}) + w_{\mathrm{prior}} E_{\mathrm{prior}}(\mathbf{x}) + N \log Z(\sigma) - \log p(\sigma) \; , \end{aligned}$$

with the correspondences

$$
\begin{aligned}
w_{\text{data}} &= \frac{1}{\sigma^2} \ , \\
w_{\text{prior}} &= \beta \ , \\
E_{\text{data}}(\mathbf{x}) &= \frac{\chi^2(\mathbf{x}; D)}{2} \ .
\end{aligned}
$$

In practice, one can absorb w_{prior} in w_{data} and the other factors. While these terms have equivalents in scoring functions for conventional approaches (Eq. 1.1), the factors involving σ do not, which demonstrates that scoring functions lacking these terms are not complete if one aims for an unbiased estimate.

1.3 Markov Chain Monte Carlo sampling

Sampling from joint posterior distributions is a difficult task and can, in general, not be done using standard random number generators available in many programming languages. But approximative, iterative techniques are available to solve this problem. A major class of methods is called Markov Chain Monte Carlo, which we use extensively in this work and thus give a thorough introduction to.

Often, we want to calculate quantities which can be expressed as an average with respect to a probability distribution $p(x)$. Take, for example, the integral of a function $f(\mathbf{x})$ over a domain $\Omega \subseteq \mathbb{R}^n$;

$$
I = \int_{\Omega} \mathrm{d}\mathbf{x} \ f(\mathbf{x}) \ . \tag{1.6}
$$

Introducing an arbitrary probability distribution with support Ω, we can also write

$$
I = \int_{\Omega} \mathrm{d}\mathbf{x} \ \frac{f(\mathbf{x})}{p(\mathbf{x})} p(\mathbf{x}) = \left\langle \frac{f}{p} \right\rangle_{p(\mathbf{x})} \tag{1.7}
$$

and have thus interpreted I as an average of a different function with respect to the probability density p. In the most naive approach to numerical interpretation we would discretize Ω in uniform hypercubes as supporting points

and approximate I by a finite sum, but this not only becomes quickly infeasible in higher dimensions but is also a waste of resources, as, depending on f, many of the supporting points may not contribute significantly to the total, approximative value of I.

The main idea of Monte Carlo methods now is to approximate an integral given by Eq. 1.7 by drawing N representative samples $x^1, \ldots, x^N$ from the distribution $p(x)$ and thus avoid evaluating unneccessarily many supporting points. But the problem lies in actually obtaining samples from $p(x)$ with significant statistical weight, which is in general a challenging task due to high dimensionality and correlation between variables. Techniques like rejection sampling (which was already used by von Neumann [1951]) or importance sampling (Kahn and Harris [1951], Andrieu et al. [2003] or Robert and Casella [2004]) use an approximating distribution $q(x)$, in order not to have to draw samples directly from $p(x)$, but finding an appropiate $q(x)$ that is easy to sample from is basically impossible for complex and high-dimensional target distributions $p(x)$ with unknown modes. Just drawing N samples from a uniform distribution over Ω will not be efficient, because for a sufficiently complex $p(x)$, most of the x^i will have very little statistical weight and thus contribute little to the average. In an extreme case, one might hit not even a single mode of $p(x)$.

In their seminal paper, Metropolis et al. [1953] achieved sampling $p(x)$ by simulating a Markov chain which prefentially explores regions of high probability and whose samples approach the target distribution $p(x)$. For details on the theory of Markov chains, see Robert and Casella [2004]. Here we sketch only the most important points, sacrificing technical and mathematical detail. A Markov chain is a random process over a (discrete or continuous) state space, which, in the case of biomolecular simulation, is given by all possible conformations. This process needs to fulfill

$$p(x^t|x^{t-1} \cdots x^0) = p(x^t|x^{t-1}) \ . \qquad \text{(Markov condition)}$$

In other words, it is memory-less, and the t-th state x^t of the chain only depends on the state x^{t-1} at "time" $t-1$. The Markov chain is then completely specified by the probability $p(x^0)$ for the first state x^0 and transition kernel

$T(x^{t+1}|x^t)$. In algorithms discussed later, the transition kernel is not constant, but depends on the time t and the Markov chain is thus *time-heterogeneous*. For now, we only discuss time-homogeneous chains. The transition kernel gives the probability of obtaining x^{t+1} given that the current state of the chain is x^t. In order to have a unique limiting distribution $f(x)$, the Markov chain needs to be irreducible, aperiodic and positive recurrent. Irreducibility means that every state is accessible from every other state and aperiodicity guarantees that the chain does not get stuck in cycles. A Markov chain is *recurrent* if the chain returns to every state infinitely often, and *positive recurrent*, if the expected recurrence times are finite. A chain fulfilling these three conditions is called *ergodic* and will, after a burn-in period, eventually loose the memory of its initial state x^0. In practice, though, one enforces a sufficient but not neccessary condition for the limiting distribution to be the target distribution we wish to sample from, called *detailed balance*:

$$p(x)T(y|x) = p(y)T(x|y) \ \forall x, y, \forall t \ . \tag{1.9}$$

Most MCMC algorithms are constructed in a way that this condition holds, although we stress that algorithms obeying the weaker condition of *global balance*, $p(x) \int \mathrm{d}y \ T(y|x) = \int \mathrm{d}y \ p(y)T(x|y)$, are also correct, e.g. Convective Replica Exchange [Spill et al., 2013].
In the following, we shortly review the algorithms we use for sampling from the ISD posterior distribution.

1.3.1 Metropolis-Hastings algorithm

Different MCMC algorithms only differ in the transition kernel $T(x^{t+1}|x^t)$ and in the construction of the target distribution $p(x)$. In all MCMC algorithms relevant in the context of ISD, the transition kernel $T(x^{t+1}|x^t)$ is decomposed into a proposal density q, from which a new proposal x^{t*} state is drawn given the current state x^t, and a probability p_{acc} to accept the proposal x^{t*} as the next state in the Markov chain;

$$T(x^{t+1}|x^t) = q(x^{t*}|x^t)p_{\mathrm{acc}}(x^{t+1} = x^{t*}|x^t, x^{t*}) \ .$$

In practice, one designs q under an application-dependent rationale to give good proposal states and then finds a p_{acc} such that $T(x^{t+1}|x^t)$ obeys at least the balance condition.

If one chooses a symmetric proposal distribution, that is, $q(x^{t*}|x^t) = q(x^t|x^{t*})$, a convenient and valid choice for p_{acc} is

$$p^t_{\text{acc}}(x^{t+1} = x^{t*}|x^t, x^{t*}) = \min\left\{1, \frac{p(x^{t*})}{p(x^t)}\right\}. \tag{1.10}$$

This acceptance probability and the symmetric proposal distribution $q(x^{t*}|x^t) \propto \theta(x^t + \epsilon)\theta(x^t + \epsilon)$, with $\theta(x)$ being the Heaviside function equaling 1 for $x \geq 0$ and 0 otherwise, was chosen by Metropolis et al. [1953]. They used the very first MCMC algorithm to calculate expectation values with respect to the Boltzmann distribution of a system of N particles interacting via a potential depending on pairwise distances only. Their algorithm was later generalized to its current form by Hastings [1970]. Fig. 1.1 illustrates the process of transitioning from one state to the next in the Markov chain constructed by the Metropolis-Hastings algorithm. All subsequently described MCMC algorithms are derived from the Metropolis-Hastings algorithm and differ only in the proposal distribution and the acceptance rule.

1.3.2 Gibbs sampling

The ISD posterior distribution is a probability distribution for a set of quite different variables containing the structure $\mathbf{x}$ and the nuisance parameters α and σ.

In theory, we could apply the previously described Metropolis-Hastings sampling method with a simple (e.g., uniform) proposal density, thus changing all three variables by some stepsize and then accepting / rejecting the move. But very small stepsizes or, more generally, proposal distributions q peaked close to the current state would be necessary to attain a reasonably acceptance probability. The reason for this is that all variables are highly correlated and thus the variable on which the posterior distribution at a current step in the Markov chain most strongly depends would set an upper limit on the width

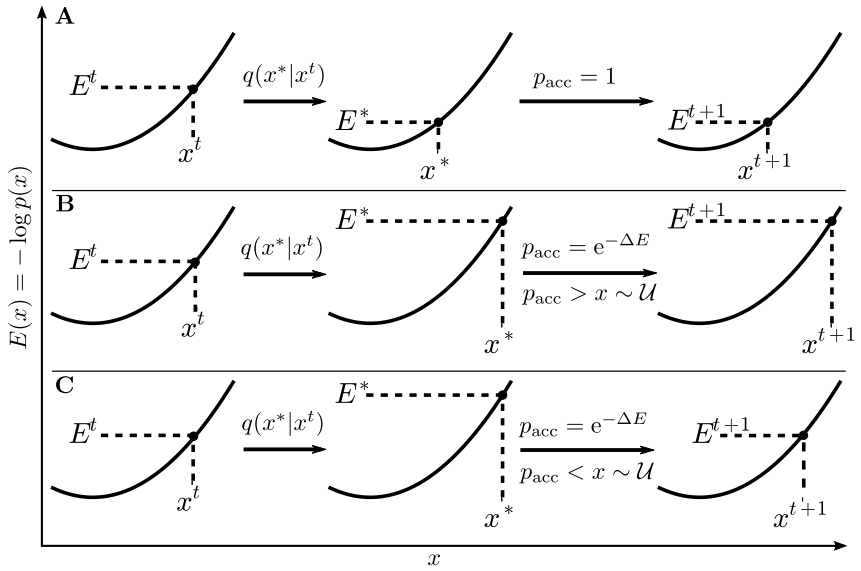

Figure 1.1: Illustration of possible outcomes of a Metropolis-Hastings move $x^t \to x^{t+1}$. The current state is x^t (*left*), then a proposal state x^* is drawn from the proposal distribution $q(x^*|x^t)$ (*middle*) and finally accepted or rejected according the Metropolis acceptance rule (*right*) to obtain the next state x^{t+1} in the Markov chain.

A: proposing a state with a lower energy than the current state leads to certain acceptance.

B, C: a state with an increased energy is accepted (*B*) with probability $\exp\{-[E(x^*) - E(x^t)]\} =: \exp(-\Delta E)$, else rejected (*C*).

of the proposal distribution. This leads to insignificant changes in the other variables and thus produces a highly correlated Markov chain, whose states will take an unfeasible amount of time to converge to the posterior distribution. It is thus useful to decouple the variables and instead sample from the posterior distribution by "taking turns". Furthermore, for some variables, the conditional distributions might be of standard form and decoupling the variables for sampling would allow to take advantage of random number generators already implemented in popular programming languages or, more general, to use appropriate and easier-to-implement samplers for each variable.

Gibbs sampling [Geman and Geman, 1984] achieves exactly this decoupling of variables. The idea is to sample from the conditional probability distribution of one variable while keeping all others fixed, then sample from the distribution of the next variable conditioned on the recently sampled value for the previous variable and the previous values for the other variables. This is repeated until samples have been drawn from all conditional posterior distributions. The set of the freshly-drawn samples then is a new sample from the joint distribution. Thus, its transition kernel $T(x^{t+1}|x^t)$ is given by the iteration

$$
\begin{aligned}
x_0^{t+1} &\sim p(x_0|x_1^t, \ldots, x_N^t) \ , \\
x_1^{t+1} &\sim p(x_1|x_0^{t+1}, x_2^t, \ldots, x_N^t) \ , \\
&\vdots \\
x_N^{t+1} &\sim p(x_N|x_0^{t+1}, \ldots, x_{N-1}^{t+1}) \ .
\end{aligned}
$$

It is interesting to note that, formally, the Gibbs sampler is a special case of the Metropolis-Hastings algorithm and it is astonishing that the conditional distributions contain enough information to recover the joint distribution [Robert and Casella, 2004]. With Gibbs sampling, we thus can dissect the challenging task of sampling from the full posterior distribution into sampling from conditional posterior distributions. For each of the variables, we employ an appropiate sampling method. While for the nuisance parameters the conditional distributions are often standard distributions like the normal-, Gamma- or log-normal distribution for which samplers are readily available in many programming languages, the structural coordinates usually do not follow any

standard distribution.

1.3.3 Hamiltonian Monte Carlo

By far the most challenging conditional posterior distribution to sample from is the one for the structural model parameters. The coordinates describing the three-dimensional fold of the polymer are still highly correlated: a displacement of one monomer not only affects other monomers restrained to it through the likelihood, but all other monomers due to the prior information. In a dense system, a single displacement could lead to overlap with several other monomers, but independent of the density, each monomer is coupled to all others through the fact that the polymer is a connected chain. For this reason, again, a Metropolis-Hasting scheme with a naive proposal distribution is inefficient. We thus need a method giving distant proposals with high acceptance rate. This can be achieved by taking into account the gradient of the negative log-probability, which is the central idea of Hybrid (or Hamiltonian) Monte Carlo (HMC, [Duane et al., 1987]).
First, an auxiliary variable v is introduced, whose sampling distribution must be symmetric and is usually taken to be the Normal distribution; $v \sim \mathcal{N}(0,1)$. Thus, HMC samples from the joint distribution

$$g(x,v) \propto p(x) \times \mathcal{N}(0,1) = \exp\left\{-\left[-\log p(x) + \frac{v^2}{2}\right]\right\} \ .$$

From a physics point of view, this is nothing but the Boltzmann distribution of a system at inverse temperature $\beta = 1$ with Hamiltonian

$$H(x,v) = -\log p(x) + \frac{v^2}{2} \ , \tag{1.11}$$

x taking the role of particle positions and v being their momenta. To obtain samples from $p(x)$, one then has to simulate from $p(x,v)$ and marginalize over v, which is trivial, because the two variables are independent.
In practice, HMC is implemented by making use of Gibbs sampling (Sec. 1.3.2): first, momenta v are drawn, then Hamilton's equations of motion are solved

with (x^t, v^t) as initial conditions, resulting in a state (q^{t*}, v^{t*}) at some pseudo-time τ. As the Hamiltonian flow $\phi_\tau : (x^t, v^t) \mapsto (x^{t*}, v^{t*})$ is volume-preserving and reversible, this procedure constitutes a valid transition kernel. In general, though, Hamilton's equations cannot be solved exactly and one has to resort to a volume-perserving and reversible numerical integration scheme. The most common choice is the leap frog integrator (see, e.g., [Hairer et al., 2003]) or related algorithms, as they require only one gradient evaluation per time step while maintaining second order accuracy. Other symplectic and reversible integrators can be chosen, such as RESPA [Tuckerman et al., 1992], which relies on a Trotter factorization of the Liouville propagator to decouple the motion of degrees of freedom with different timescales. But all methods for numerical integration have in common that they neccessarily introduce an error, which makes the trajectory deviate from the actual ensemble. In HMC, it is accounted for by an acceptance / rejection step. The most common choice is

$$p_{\text{acc}}^{\text{HMC}}(q^{t+1} = q^{t*}, v^{t+1} = v^{t*} | q^t, v^t) = \min\left\{1, \mathrm{e}^{-\Delta H}\right\} \tag{1.12}$$

with $\Delta H = H(q^{t*}, v^{t*}) - H(q^t, v^t)$ being the difference in total energy.

The transition kernel is thus a product of three components: the probability to draw a certain momentum, a Dirac delta distribution reflecting the fact that evolution under the flow defined by the integration scheme is deterministic, and the acceptance probability. It can be shown to fulfill detailed balance, guaranteeing that HMC indeed samples from the target distribution $p(x)$.
Recent work by Sohl-Dickstein et al. [2014] interprets a HMC step (up to momentum randomization) as a sequence of operators acting on a discrete state space. This view of HMC is very intuitive and also leads them to propose a detailed balance-violating variant of HMC, termed "Look Ahead HMC", which is able to reduce random-walk behaviour and results in performance gains.
The performance of HMC critically depends on choosing appropiate timesteps Δt and number of integration steps $\tau/\Delta t$. Too large timesteps give distant and, as such, less correlated proposals, but the numerical error increases and proposals are most likely rejected. Similarily, a small number of integration steps

with appropiate timestep gives high acceptance rate, but subsequent states in the Markov chain are closer to each other. While in the above description the mass matrix of the artificial system was set to unity, this is another important parameter to tune. Recently, the fact that probability distributions are a Riemannian manifold led to the exploitation of differential geometric concepts to automatically tune the mass matrix [Girolami and Calderhead, 2011; Lan et al., 2012].

In the context of ISD, the part of the log-posterior gradient stemming from the likelihood can, depending on the number of data points and the forward model, be quite cheap to compute. The gradient of the force field $E(x)$, on the other hand, is usually way more complicated and expensive to calculate. The reason for this is the usually employed volume exclusion, which, for N atoms, in a naive implementation, requires N^2 distance calculations. These necessarily involve square roots and are thus responsible for a significant part of computation time spent calculating the intermolecular forces. Luckily, Verlet- [Verlet, 1967] and cell [Boris, 1986; Mattson and Rice, 1999] lists reduce the amount of distances to be calculated.

It is important to note that the Hamiltonian (Eq. 1.11) generating the dynamics is a convenient, but not neccessary choice. Any Hamiltonian dynamics can be chosen, as long as it gives reasonable proposal states and thus good acceptance rates. This is a possibility to save computation time for gradient evaluations by fixing some degrees of freedom, or, in large data sets, disregarding some data points. It is important, though, to perform the acceptance criterion using the "real" Hamiltonian.

1.3.4 Replica Exchange

While for small systems with simple likelihoods and prior distributions, Gibbs sampling in conjunction with appropriate subsamplers is sufficient, in larger system sizes and for highly multimodal posterior distributions, the respective samplers can easily get stuck in local minima, thus increasing correlation times and slowing down convergence. This problem can be alleviated by so-called

extended-ensemble algorithms, in which the (pseudo-physical) system is not only simulated at the target (pseudo-) temperature, but also at higher temperatures, which effectively flatten out its energy landscape and by some means give the low-temperature Markov subchain access to distant states from the high-temperature ensemble. While several methods exist (see Iba [2001] for a review), we focus on Replica Exchange (RE, Geyer [1991]; Swendsen and Wang [1986]).

In RE and modifications of it, one simulates not only the target distribution, but also variants of it which are easier to sample. The typical choice to generate the family of distributions is to simulate Boltzmann distributions at different temperatures, but one can as well, as proposed by [Sugita et al., 2000], modify the energy function, which, in the following, we often do, or vary other thermodynamic parameters.

For simplicity, we assume that we are interested in simulating the Boltzmann distribution $p(x) = \exp[-\beta E(x)]/Z(\beta)$ of a system with potential energy $E(x)$ and that the only replica parameter is the inverse temperature β. Then, at randomly chosen or fixed intervals, one tries to exchange configurations between different simulations (i.e., temperatures) and after that continues with normal equilibrium MCMC or MD sampling. These exchanges are constructed in a way that at least the balance condition is fulfilled. This scheme allows the low-temperature simulations to get access to conformations sampled by the high-temperature simulations through a random walk in temperature space. Fig. 1.2 illustrates the method and shows the trace of two states in a simulation with 5 replicas as they traverse the temperature ladder. Describing RE in a more formal way, one samples not from the target density $p(x)$, but from the joint distribution

$$p(x) = p_0(x_0) \dots p_N(x_N). \tag{1.13}$$

If one is interested in a specific distribution $p_k(x_k)$, one can just use all sampled x_k and ignore the remaining components of x, because all components of x are mutually independent. RE and related algorithms in fact simulate a time-heterogenous Markov chain: the regular sampling using regular MCMC

algorithms or MD constitute one transition kernel and the exchange moves another one. Not all exchange moves are accepted, since this would disturb the respective equilibrium distributions of the single chains. Instead, an exchange move $(x_i^t, x_j^t) \to (x_i^{t+1}, x_j^{t+1}) = (x_j^t, x_i^t)$ between simulations i and j is accepted with a probability

$$p_{\text{acc}}^{RE}(x_i^{t+1} = x_j, x_j^{t+1} = x_i) = \min\left\{1, \frac{p_i(x_j^t)p_j(x_i^t)}{p_i(x_i^t)p_j(x_j^t)}\right\} \ . \tag{1.14}$$

This acceptance probability is again designed to make the exchange transition kernel obey the detailed balance condition, but, as always, it is not the only possible acceptance probability to do so. If we again take a physics perspective and introduce pseudo-energies $E_i(x_i) \coloneqq -\log p_i(x_i)$ and further assume that $E_i(x_i) = \beta_i E(x_i)$, we are simulating the Boltzmann ensembles of a system with potential energy $E(x)$ at N different inverse temperatures β_i. The acceptance probability then takes the form

$$p_{\text{acc}}^{\text{RE}}(x_i^{t+1} = x_j^t, x_j^{t+1} = x_i^t) = \min\left\{1, \mathrm{e}^{\Delta\beta\Delta E}\right\} \tag{1.15}$$

with $\Delta\beta = \beta_i - \beta_j$ and $\Delta E = E(x_i^t) - E(x_j^t)$. From Eq. 1.15 it is clear that in order to get reasonable exchange acceptance rates, the inverse temperatures β_i, β_j must not be too different. This is why usually only exchanges between neighbouring simulations are attempted. Nevertheless, one might need many intermediate distributions to allow efficient sampling. This obviously increases computational cost and it is difficult to choose an efficient schedule. Additionally, by construction, RE performs a random walk in temperature space and too many intermediate distributions slow down the diffusion of "high-temperature states" to the target distribution; the average time a state needs to cross the whole temperature ladder scales with $\sqrt{N}$ [Hukushima and Nemoto, 1996]. Much effort has been put into finding systematic methods to optimize RE schedules; a subject we revisit in Sec. 3. In structure calculations using ISD we make massive use of RE to sample from conditional posterior distributions of structures, which often vary in not only one, but two temperature-like parameters [Habeck et al., 2005]. A great advantage of RE is that the samples drawn in the high-temperature simulations can be used to estimate very

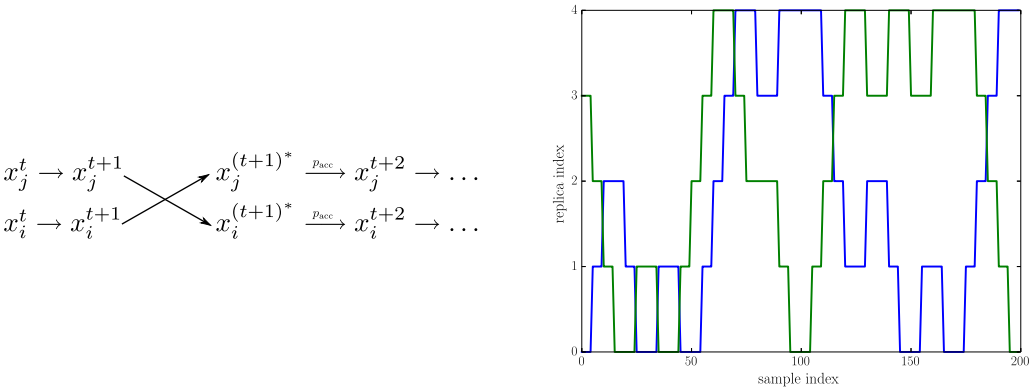

Figure 1.2: *Left*: schematic depiction of Replica Exchange (RE): two chains i, j sample different distributions $p_i(x), p_j(x)$. The step $t \to t+1$ is performed in both chains with some MCMC or MD method. The transition $t+1 \to t+2$ then consists in proposing state x_i^{t+1} to chain j and, vice-versa, x_j^{t+1} to chain i. The swap is then accepted with a probability $p_{\rm acc}$ and local sampling continued.

Right: trace of two states in a RE simulation with five different ensembles. Exchanges were attempted every five steps. Longer horizontal lines mean rejected swap attempts.

accurate averages with respect to the target distribution $p_0(x_0)$ by means of reweighting techniques like the Multistate Bennett Acceptance Ratio method (MBAR, [Shirts and Chodera, 2008]) and the Weighted Histogram Analysis Method (WHAM, Sec. 1.4). In the context of Bayesian analyses like ISD, this is very convenient to estimate model evidences, as done in, e.g., [Mechelke and Habeck, 2012].

1.4 Weighted Histogram Analysis Method (WHAM)

When running multicanonical algorithms like RE, the primary goal is to enhance sampling of the target distribution. But instead of discarding them as an (expensive) by-product, we would like to use samples from *all* interpolating distributions to improve the estimate of an average quantity or the target distribution it self. This is non-trivial, as samples from one ensemble will not have the same weight in another ensemble and it is not obvious how to reweigh them to accomodate for this fact. The solution is found in the multiple histogram reweighting method (WHAM, Ferrenberg and Swendsen [1989]; Kumar et al. [1992]), the idea of which we will briefly sketch.

Assume, for simplicity, that we are dealing with a system in a canonical ensemble at different inverse temperatures β. The probability to find the system in the infinitesimal configuration space interval $[x, x + \mathrm{d}x]$ is given by the Boltzmann ensemble $p(x|\beta) = \exp[-\beta E(x)]/Z(\beta)$. A central quantity characterizing a physical system is the density of states (DOS)

$$\Omega(E) = \int \mathrm{d}x \; \delta[E - E(x)] \; , \tag{1.16}$$

which can also be multi-dimensional,

$$\Omega(E_0, \ldots, E_n) = \int \mathrm{d}x \; \prod_i \delta \left[E - E_i(x)\right] , \tag{1.17}$$

if the total system energy is split up in several components $E_0, \ldots, E_n$. The DOS allows the calculation of ensemble averages such as the partition function

$Z(\beta_k)$ or, more generally, the average over any function of the energy E, at arbitrary β. Using Ω, we can write

$$\langle f(E)\rangle_\beta = \frac{\int \mathrm{d}x\ f[E(x)]\mathrm{e}^{-\beta E(x)}}{\int \mathrm{d}x\ \mathrm{e}^{-\beta E(x)}} = \frac{\int \mathrm{d}E\ \Omega(E) f(E)\mathrm{e}^{-\beta E}}{\int \mathrm{d}E\ \Omega(E)\mathrm{e}^{-\beta E}}$$

and thus transform the possibly high-dimensional integral over x in a one-dimensional integral over E, which can be easily approximated numerically.

We now follow closely [Chodera et al., 2007] to establish the connection to simulations. Assume we are given discrete, independent samples from all ensembles. Noting that $\Omega(E_m)$ is the multiplicity of energy E_m, using the Boltzmann ensemble, we can now estimate the probability to find the system with energy E_m at temperature β_k as

$$p(E_m|\beta_k) = \frac{\Omega_{mk}\mathrm{e}^{-\beta_k E_m}}{Z(\beta_k)} \tag{1.18}$$

with Ω_{mk} an approximation of the DOS $\Omega(E_m)$ calculated from samples at temperature β_k. Furthermore, we approximate the partition function as

$$Z(\beta_k) = \sum_m \Delta E \Omega_m \mathrm{e}^{-\beta_k E_m}\ .$$

Here, we assume that we discretized the energies in a histogram H_{mk} with bin width ΔE, which counts the number of states x with energy $E(x) \in [E_m - {}^{\Delta E}/_2, E_m + {}^{\Delta E}/_2]$ sampled at temperature β_k.

An alternative estimation of $p(E_m|\beta_k)$, based on the histogram H_{mk}, is

$$p(E_m|\beta_k) = \frac{H_{mk}}{N_k \Delta E}\ ,$$

where we introduced the total numbers of samples N_k sampled at temperature β_k. If we equate this expression with Eq. 1.18, we can solve for Ω_{mk} and find

$$\Omega_{mk} = \frac{H_{mk} Z(\beta_k)}{N_k \Delta E \mathrm{e}^{-\beta_k E_m}}\ .$$

We realize that $Z(\beta_k)$ depends on Ω_{mk} and this equation thus has to be solved self-consistently.

WHAM in this form relies on the assumption of independently drawn samples. To account for correlation between samples, we can introduce a statistical inefficiency factor g_{mk}, which is the number of configurations after which we

obtain an uncorrelated sample. In general, it will depend both on the energy bin index m as well as on the temperature β_k. The different estimates for Ω_{mk} now have to be combined in a statistically optimal way to yield a final estimate Ω_m, which will be given by a weighted sum of all Ω_{mk} with the condition that the weights sum up to one. The statistically optimal weights can be found either by minimizing the variance of the estimator $\hat{\Omega}_{mk}$, as done by Ferrenberg and Swendsen [1989]; Kumar et al. [1992], or by maximizing the likelihood of Ω_k given the single estimates [Bartels and Karplus, 1997]. Either way, the result is a statistically optimal estimator given by the iteration

$$\hat{\Omega}_m = \frac{\sum_k g_{mk}^{-1} H_{mk}}{\sum_k g_{mk}^{-1} N_k \Delta E \mathrm{e}^{f_k - \beta_k E_m}} \ , \tag{1.19}$$

where the free energies

$$f_k = -\log \sum_m \hat{\Omega}_m \Delta E \mathrm{e}^{-\beta_k E_m} \tag{1.20}$$

have been introduced.

Chodera et al. [2007] perform a careful analysis of the WHAM equations (1.19, 1.20) for samples drawn using multicanonical algorithms by taking into account the correlation between different ensembles. Furthermore, it is important to note that WHAM is applicable not only to samples from Boltzmann ensembles at different temperature, but directly generalizes to arbitrary ensembles $p(E)$, which map the system energy to a probability, and to several temperature-like replica parameters. In the latter case, the DOS is multi-dimensional.

In fact, the ensemble under consideration does not have to be physical at all: for any (not necessarily normalized) probability distribution $p(x)$ we can, as mentioned before in the context of MCMC sampling methods (Sec. 1.3), introduce a pseudo-energy $E(x) = -\log p(x)$. Interpreting the probability distribution $q(x) = \exp[-E(x)]$ as a Boltzmann distribution at $\beta = 1$, we can now proceed as described before and calculate the DOS $\Omega(E)$ and the normalization constant.

In a Bayesian context, we consider the posterior distribution $p(x)$. The just described procedure then allows the calculation of the evidence, because it is the normalization constant of $p(x)$. WHAM is thus a powerful tool for Bayesian

model comparison of which we take advantage in Sec. 4.3.6.

It is also interesting to see that WHAM is very closely related to the Multistate Benett Acceptance Ratio (MBAR; Shirts and Chodera [2008]) method, as shown by a Bayesian analysis of Habeck [2012b]. Further work by Habeck [2012a] shows that the need for discrete energy histograms H_{km} can be avoided and thus a truly non-parametric estimate of the density of states can be obtained.

2 Testing Replica Exchange with Non-equilibrium Switches (RENS) on complex protein systems

Despite its usefulness and popularity, RE suffers from several drawbacks. Because RE is only effective when the overlap of energy distributions of neighboring replicas is sufficiently high [Kofke, 2002], many replicas are typically needed to bridge between the target and high-temperature ensemble. Furthermore, given the high-temperature ensemble and the destination ensemble, the number of intermediate replicas required for a state to be able to traverse the whole temperature range scales with the dimensionality or system size d as $\sqrt{d}$ [Hukushima and Nemoto, 1996].

To alleviate these problems, Ballard and Jarzynski [2009] have proposed replica exchange with non-equilibrium switches (RENS), which uses non-equilibrium trajectories to increase the phase space overlap between neighboring replicas. This is achieved by "dragging" the states to be exchanged into the other ensemble by virtue of a time-dependent, interpolating Hamiltonian. During the switching dynamics the system thus adapts to the changes in the Hamiltonian. Therefore the resulting proposal states tend to have a higher statistical weight in the respective ensembles. For a toy model (a one-dimensional particle in a rugged potential energy landscape), this method was shown to increase the swap rate and computational efficiency significantly, provided the switching

process was performed slowly. In a more recent article [Ballard and Jarzynski, 2012], the authors apply their algorithm to sample conformations of dialanine in implicit solvent and find not only large efficiency gains at low temperature, but also less correlated samples.

Because both the toy model as well as the dialanine are fairly small systems, it remains unclear whether RENS also holds its promises when simulating more complex, high-dimensional systems. In particular, due to the increased number of replicas needed to simulate large systems with rugged energy landscapes, it is important to assess the efficiency of RENS in problems with higher dimensionality and complexity. In this section, we study the use of RENS for a protein network model and we show an application of RENS in ISD applied to protein structure determination [Habeck et al., 2005]. We investigate various ways of generating non-equilibrium trajectories. We indeed see an increase in the swap rate but find that it is not clear if the additional computational costs introduced by the non-equilibrium switches really pay off.

2.1 An introduction to recent results in non-equilibrium statistical mechanics

The topic of non-equilibrium statistical mechanics is usually not found in undergraduate textbooks, although our world is inherently not in thermodynamic equilibrium. Closed systems only exist in theory and any thermodynamical or statistical treatment of a realistic system as such can only be an approximation. Systems which we can reasonably assume as closed might nevertheless not be in equilibrium. An example is the packing of DNA in eukaryotic nuclei: while the duration of a cell cycle is in the order of days, the time for not-intermingling chains of the length of human chromosomes to relax to an equilibrium state with maximal entropy is much longer, in the order of 500 yr [Rosa and Everaers, 2008]. While this clearly shows the need for high topoisomerase activity, it also demonstrates that non-equilibrium effects certainly play a large role in nuclear organization.

RENS depends strongly on recent results in non-equilibrium statistical mechanics and while perhaps not necessary for a qualitative understanding of the method, in light of the previous paragraph we feel that a short introduction to recent results of this exciting field is in order.

Until the late 90's, non-equilibrium statistical mechanics consisted mainly in the treatment of systems not too far from equilibrium by linear response theory (for an introduction see, e.g., the excellent lecture notes by Tong [2012]). Few relations are valid arbitrarily far from equilibrium were known, among them Evans and Searles' transient fluctuation theorem [Crooks, 1999a; Evans and Searles, 1994], until Jarzynski [1997a] discovered an equality between the work exerted in repeated non-equilibrium experiments and the free energy difference ΔF between initial and final states valid for systems driven arbitrarily far away from equilibrium. More exactly, the relation

$$\langle \mathrm{e}^{-\beta W[\gamma_{A\to B}]} \rangle_{\gamma_{A\to B}} = \mathrm{e}^{-\beta \Delta F} \qquad \text{(Jarzynski equality)}$$

equates the average of the exponential of the negative work W over infinitely many repetitions of a non-equilibrium trajectory $\gamma_{A\to B}$ starting from canonically distributed initial conditions $x_0 = (\mathrm{q}_0, \mathrm{p}_0)$ with the free energy ΔF, that is, the minimum amount of work required to drive a system from state A to state B. This is obviously a stronger statement than the theorem of maximum work, $\langle W \rangle \geq \Delta F$, where the equality holds for a quasistatic process. This result was proved in Jarzynski [1997a] for a system weakly coupled to a heat bath by considering the fully deterministic evolution of the extended system including the heat bath. Not long after, in [Jarzynski, 1997b], the Jarzynski equality was re-derived considering a thermostatted system evolving according to stochastic, Markovian dynamics satisfying detailed balance. We are mainly interested in yet another derivation found by Crooks [1998, 1999a,b], where Crooks proves a more general result implying Eq. Jarzynski equality, several other results in non-equilibrium statistical mechanics and also a result important to RENS. In the following we briefly sketch the so-called fluctuation theorem and its derivation. We will limit ourselves to the case of dynamics

in discrete time and space and follow closely Crooks [1998, 1999b]. The same reference gives the generalization to continuous time and space dynamics.

A system is called *microscopically reversible*, if the probability of a particular trajectory γ of the system in phase space is related to the probability of the time-reversed trajectory $\hat{\gamma}$ by

$$\frac{P[\gamma|\gamma(0),T]}{\hat{P}[\hat{\gamma}|\hat{\gamma}(0),\hat{T}]} = \mathrm{e}^{-\beta Q[\gamma]} \ . \qquad \text{(microscopic reversibility)}$$

The probabilities on the l.h.s. depend on a trajectory (or path) γ, not only on its start and end points. Here and in the following we denote functions with this property *path functions* and put their arguments in square brackets to distinguish them from state functions. A hat over a quantitiy denotes their time-reversed twins, for example, for a trajectory $\gamma = (x_0, \ldots, x_N)$ the time-reversed trajectory is $\hat{\gamma} = (x_N, \ldots, x_1)$. T denotes the transition kernel (see Sec. 1.3) for γ and $Q[\gamma]$ the heat produced during the trajectory γ.

A sufficient condition for a system in contact with a heat bath to be microscopically reversible is that the dynamics are Markovian, that is, memory-less in the sense of Markov chains (Sec. 1.3) and that they preserve the equilibrium distribution of the unperturbed system. This means that if a system is in equilibrium at a certain value of an external "switching parameter" λ and it is not perturbed, that is, λ is kept fixed, the system samples a canonical distribution specified by λ. Furthermore, the energy of the system always has to stay finite. The proof that microscopic reversibility indeed follows from those assumptions is very insightful and we quickly sketch it.

It is always possible to split up one transition in a time-inhomogenous Markov chain into two distinct substeps. First, a stochastic transition $x^t \to x^{t+1}$ occurs with the probability $T^t(x^{t+1}|x^t)$ and then we update the transition kernel $T^t \to T^{t+1}$, with which the next transition $x^{t+1} \to x^{t+2}$ is performed. In our physical setting of a non-equilibrium trajectory γ with substeps $x = (x^0, x^1)$, the stochastic transition corresponds to an exchange of heat with the heat bath and the transition kernel update represents a change in the switching parameter λ. In the time-reversed non-equilibrium trajectory, it is important to note that the order of stochastic transition and perturbation is flipped. Thus we

start with x^{t+1}, then perform the kernel update $T^{t+1} \to T^t$ and the system performs a stochastic transition $x^{t+1} \to x^t$. The path probability ratio for this one-step trajectory is thus

$$\begin{aligned}\frac{P[x|x^0,T]}{\hat{P}[\hat{x}|\hat{x}^0,\hat{T}]} &= \frac{\pi^1\left(x^1\right)\pi^2\left(x^2\right)}{\pi^2\left(x^1\right)\pi^1\left(x^0\right)} \\ &= \frac{\pi^1\left(x^1\right)}{\pi^1\left(x^0\right)}\frac{\pi^2\left(x^2\right)}{\pi^2\left(x^1\right)} \\ &= \exp\{-\beta\left[H^1\left(x^1\right) - H^1\left(x^0\right)\right]\} \\ &= \mathrm{e}^{-\beta Q[x]}\ ,\end{aligned}$$

where we made use of the conditions that at a given timestep t, the system samples the canonical distribution $\pi^t(x^t) \propto \exp[-\beta H^t(x^t)]$. The expression in the exponent can be identified with the heat $Q[\gamma]$ produced during the one-step trajectory because it is an energy difference not caused by an external change, otherwise it would have to be associated with work performed on the system. Generalization to trajectories with multiple steps is trivial and just adds more energy difference terms to the expression for the heat Q. The work has a similar microscopic definition. Both will be relevant in Sec. 2.2, which is why we give their definitions explicitly:

$$Q[\gamma] = \sum_t \left[H^t\left(x^t\right) - H^t\left(x^{t-1}\right)\right] \qquad \text{(microscopic definition of heat)}$$

$$W[\gamma] = \sum_t \left[H^t\left(x^{t-1}\right) - H^{t-1}\left(x^{t-1}\right)\right] \quad \text{(microscopic definition of work)}$$

By using the chain rule when calculating

$$\begin{aligned}\Delta H &= \int_0^\tau \mathrm{d}t\ \frac{\mathrm{d}H(x;t)}{\mathrm{d}t} \\ &= \underbrace{\int_0^\tau \mathrm{d}t\ \dot{x}\cdot\nabla_{\mathbf{x}}H}_{\text{heat}} + \underbrace{\int_0^\tau \mathrm{d}t\ \frac{\partial H}{\partial t}}_{\text{work}} \\ &= Q + W\end{aligned}$$

for the total energy difference ΔH after driving a system during the time interval $[0,\tau]$, we can also give a definition for the microscopic heat for continous

time and space [Crooks, 1999b; Schöll-Paschinger and Dellago, 2006].
The just-proven macroscopic reversibility property is essential in the proof Ballard and Jarzynski [2012] give for RENS fulfilling detailed balance.
Given microscopic reversibility, it is easy to prove Crooks' main result, namely,

$$\langle F \rangle_{\mathrm{F}} = \langle \hat{F} \mathrm{e}^{-\beta W_{\mathrm{d}}} \rangle_{\mathrm{R}} \ . \qquad \text{(Crooks' theorem)}$$

$F[\gamma]$ denotes any path function of a trajectory γ and $\hat{F}$ its time-reversed twin. $W_{\mathrm{d}}[\gamma] = W[\gamma] - \Delta F$ is the dissipative work defined as the difference between the total work done during the trajectory γ and the free energy difference, the latter being the minimum amount of work needed to drive a system from state A to B. The averages on the l.h.s. are taken over a set of forward trajectories F starting from canonically distributed initial states and the average on the r.h.s. is over the corresponding set of time-reversed trajectories.
Setting $F = 1$ and substituting the definition of the dissipative work, one immediately recovers the Jarzynski equality. By setting F to different functions, one can easily derive [Crooks, 2000] an entropy fluctuation theorem related to Evans and Searles' transient fluctuation theorem [Crooks, 1999a; Evans and Searles, 1994], a generalization of the Kawasaki response [Yamada and Kawasaki, 1967] and the probability distribution of a non-equilibrium ensemble. It is important to realize that all these results hold arbitrarily far from equilibrium.

2.2 The RENS method: increasing replica phase-space overlap by non-equilibrium simulations

We first describe RENS as introduced in Ballard and Jarzynski [2009, 2012]. Let A, B denote a pair of replicas with Hamiltonians $H_{\mathrm{A}}, H_{\mathrm{B}}$ and states $\boldsymbol{x}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{B}}$. While RE tries to directly exchange states by proposing new states $\boldsymbol{x}^*_{\mathrm{A}} = \boldsymbol{x}_{\mathrm{B}}$ and $\boldsymbol{x}^*_{\mathrm{B}} = \boldsymbol{x}_{\mathrm{A}}$ for replica A and replica B, RENS generates the proposal states

by calculating non-equilibrium trajectories $\gamma_{\mathrm{A}\to\mathrm{B}}, \gamma_{\mathrm{B}\to\mathrm{A}}$ of length τ starting from $\boldsymbol{x}_{\mathrm{A}}$ and $\boldsymbol{x}_{\mathrm{B}}$, respectively, and ending in states $\boldsymbol{x}^*_{\mathrm{B}}, \boldsymbol{x}^*_{\mathrm{A}}$. During these trajectories, the dynamics are governed by a Hamiltonian H_λ that depends on time-dependent "switching protocols" $\lambda, \overline{\lambda} : [0, \tau] \to [0, 1]$, $\overline{\lambda}(t) = \lambda(\tau - t)$ such that

$$H_{\lambda(0)=0} = H_{\mathrm{A}};\ H_{\lambda(\tau)=1} \ = H_{\mathrm{B}} \tag{2.1}$$

$$H_{\overline{\lambda}(0)=1} = H_{\mathrm{B}};\ H_{\overline{\lambda}(\tau)=0} = H_{\mathrm{A}}\ . \tag{2.2}$$

The initial state $\gamma_{\mathrm{A}\to\mathrm{B}}(0) = \boldsymbol{x}_{\mathrm{A}}$ is highly probable in ensemble A and "dragged" into ensemble B by virtue of the time-dependent Hamiltonian. The final state $\boldsymbol{x}^*_{\mathrm{B}}$ tends to be more likely in ensemble B than $\boldsymbol{x}_{\mathrm{A}}$ itself. The same reasoning applies to $\gamma_{\mathrm{B}\to\mathrm{A}}$.

The probability of accepting this exchange depends on the total work generated during the switching processes:

$$p_{\mathrm{acc}} = \Pr\{(\boldsymbol{x}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{B}}) \mapsto (\boldsymbol{x}^*_{\mathrm{B}}, \boldsymbol{x}^*_{\mathrm{A}})\} = \min\{1, \mathrm{e}^{-W_{\mathrm{A}\to\mathrm{B}} - W_{\mathrm{B}\to\mathrm{A}}}\} \tag{2.3}$$

where $W_{\mathrm{A}\to\mathrm{B}}$, $W_{\mathrm{B}\to\mathrm{A}}$ is the work required during the switching. For a properly thermostatted system, $W_{\mathrm{A}\to\mathrm{B}} \to \Delta F$ and $W_{\mathrm{B}\to\mathrm{A}} \to -\Delta F$ in the limit of $\tau \to \infty$ and thus $p_{\mathrm{acc}} \to 1$. Thus, the longer the switching time, the less work is expended during the switching trajectories and the higher is the probability of accepting the proposal states. RE can be viewed as the other extreme where states are swapped instantaneously such that $W_{\mathrm{A}\to\mathrm{B}} = H_{\mathrm{B}}(\boldsymbol{x}_{\mathrm{A}}) - H_{\mathrm{A}}(\boldsymbol{x}_{\mathrm{A}})$. Fig. 2.1 demonstrates the favorable effect of RENS on proposed swap states.

2.2.1 An illustrative, analytical example

We illustrate the switching time dependence of the acceptance rate for a non-thermostatted system in an analytically tractable example.

We consider a single particle with unit mass in a one-dimensional harmonic potential; the Hamiltonian is then given by $H(q, p) = {}^{p^2}\!/_2 + {}^{q^2}\!/_2$. Let two thermodynamic states of the system be defined by different temperatures T_{A},

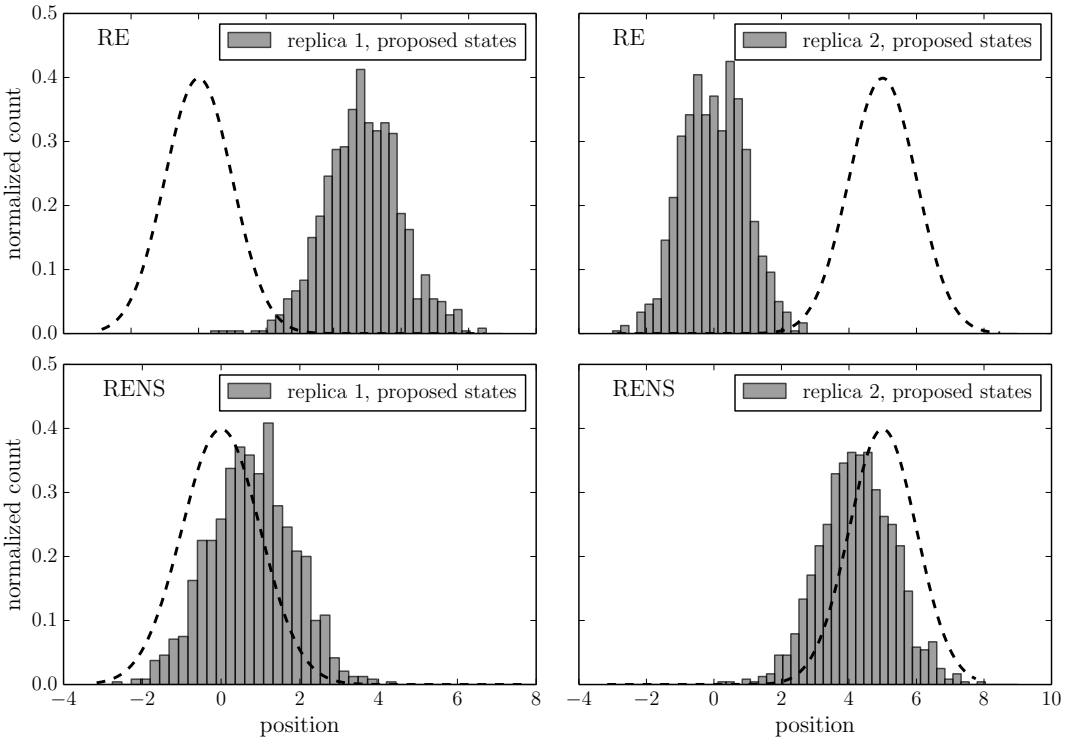

Figure 2.1: Configuration space overlap of swap proposal states for RE (*top*) and RENS (*bottom*). Two normal distributions with means $x_0 = 0$ (*left*) and $x_0 = 5$ (*right*) were simulated with HMC and swaps performed with both RE ($p_{\text{acc}} \approx 0.003$)and Langevin-thermostatted RENS (LMDRENS, $p_{\text{acc}} \approx 0.04$). Histograms show the positions proposed for a swap in each replica and dotted lines the probability distribution the states were proposed to. It seems counter-intuitive that the RENS acceptance rate is pretty low in spite of the large configuration space overlap, but acceptance rates critically depend on the work performed and considerable phase space overlap thus is not anymore an indicator for high acceptance rates.

T_B. We want to evolve the system according to the dynamics

$$\begin{aligned} \dot{q} &= p \\ \dot{p} &= -\partial_q H + \frac{1}{2T_\lambda}\dot{\lambda}\frac{\mathrm{d}T}{\mathrm{d}\lambda}p \end{aligned} \tag{2.4}$$

with $T_\lambda = T(\lambda(t))$ interpolating between T_A and T_B and a switching protocol $\lambda : [0, \tau] \to [0, 1]$, $\lambda(0) = 0$ and $\lambda(\tau) = 1$. The momentum scaling term was proposed in Ballard and Jarzynski [2009] in order to heat up the system under consideration while T_A is being switched parametrically to T_B and to cool the system down for the reverse direction.

Taking $\lambda(t) = {}^{t}\!/_{\tau}$, we look for a temperature parametrization T_λ such that ${}^{1}\!/_{2T_\lambda}\dot{\lambda}{}^{\mathrm{d}T}\!/_{\mathrm{d}\lambda} = \text{const}$ and the above equation system corresponds to the (analytically solvable) equations of motion of an time-independently damped harmonic oscillator.

This can be achieved by

$$T_\lambda = T_\mathrm{B}^{\lambda} T_\mathrm{A}^{1-\lambda} \ .$$

By this choice,

$$\frac{1}{2T_\lambda}\dot{\lambda}\frac{\mathrm{d}T}{\mathrm{d}\lambda} = \frac{1}{2\tau}\ln\left(\frac{T_\mathrm{B}}{T_\mathrm{A}}\right) =: \beta(\tau) \ .$$

We consider the case of a harmonic oscillation. This requires $\omega := \sqrt{1 - {}^{\beta^2}\!/_4}$ to be real and not zero, thus $4 > \beta(\tau)^2 > 0$. Fixing the temperatures $T_\mathrm{A}, T_\mathrm{B}$, this yields a minimum value for the switching time:

$$\tau > \frac{1}{4}\ln\left(\frac{T_\mathrm{B}}{T_\mathrm{A}}\right) \ .$$

The solution of (2.4) given β is

$$\begin{aligned} q(t) &= \left[A\cos(\omega t) + B\sin(\omega t)\right] \mathrm{e}^{\frac{1}{2}\beta t} \\ p(t) &= \left[-\omega A\sin(\omega t) + \omega B\cos(\omega t)\right] \mathrm{e}^{\frac{1}{2}\beta t} + \frac{1}{2}\beta q(t) \ . \end{aligned}$$

We want to calculate the work necessary for evolving the system under the given dynamics. For deterministic, time-reversible dynamics, Ballard and Jarzynski [2009] define the reduced work switching a system parametrically from T_A to T_B as

$$w_{AB}(\tau) = h_\mathrm{B}\left[q(\tau), p(\tau)\right] - h_\mathrm{A}(q_0, p_0) - \ln J_{AB}$$

with the reduced Hamiltonian $h_i(q,p) = {}^1\!/_{T_i} H(q,p)$, initial state q_0, p_0 and J_{AB} the Jacobi determinant of the underlying dynamics. In our case, $J_{AB} = {}^1\!/_2 \ln({}^{T_\mathrm{B}}\!/_{T_\mathrm{A}})$.

As our dynamics (2.4) without the momentum scaling term are deterministic and time-reversible, we use this definition and calculate

$$w_{AB}(\tau) = \frac{1}{2T_\mathrm{B}}\sqrt{\frac{T_\mathrm{B}}{T_\mathrm{A}}}\left[\left(\tilde{p}(\tau) - \frac{1}{2}\beta(\tau)\tilde{q}(\tau)\right)^2 + \tilde{q}(\tau)^2\right] - \frac{1}{2T_\mathrm{A}}\left(p_0^2 + q_0^2\right) - \frac{1}{2}\ln\left(\frac{T_\mathrm{B}}{T_\mathrm{A}}\right) \tag{2.5}$$

with $\tilde{q}, \tilde{p}$ the position and momentum of the undamped, unit-mass free oscillator with angular frequency ω.

Considering that $\beta(\tau) \propto \tau^{-1}$ and the fact that both $\tilde{q}(\tau)$ and $\tilde{p}(\tau)$ have an upper bound, we find

$$\begin{aligned} w_{AB}(\tau \to \infty) &= \frac{1}{2}\left(p_0^2 + q_0^2\right)\left(\frac{1}{T_\mathrm{B}}\sqrt{\frac{T_\mathrm{B}}{T_\mathrm{A}}} - \frac{1}{T_\mathrm{A}}\right) - \frac{1}{2}\ln\left(\frac{T_\mathrm{B}}{T_\mathrm{A}}\right) \\ &= \sqrt{\frac{T_\mathrm{B}}{T_\mathrm{A}}}h_\mathrm{B}(q_0,p_0) - h_\mathrm{A}(q_0,p_0) - \frac{1}{2}\ln\left(\frac{T_\mathrm{B}}{T_\mathrm{A}}\right) \ . \end{aligned} \tag{2.6}$$

This result shows that for our setup the reduced work does not drop to zero but to a constant depending on initial values and the temperatures as the switching time is increased.

Now we calculate the minimum total work, which is $w_\infty = w_{AB}(\tau \to \infty) + w_{BA}(\tau \to \infty)$. Provided that for both trajectories γ_{AB} and γ_{BA} the initial values q_0, p_0 are drawn from the equilibrium distributions corresponding to T_A and T_B, respectively, we can take the ensemble averages of Eq. (2.6) and find

$$\begin{aligned} w_\infty &= \langle w_{AB}\rangle_\mathrm{A} + \langle w_{BA}\rangle_\mathrm{B} \\ &= \sqrt{\frac{T_\mathrm{A}}{T_\mathrm{B}}} + \sqrt{\frac{T_\mathrm{B}}{T_\mathrm{A}}} - 2 \ . \end{aligned} \tag{2.7}$$

This corresponds to a maximum acceptance rate of

$$p_\infty^\mathrm{acc} = \mathrm{e}^{-w_\infty} < 1 \ . \tag{2.8}$$

2.2.2 Three different kinds of non-equilibrium dynamics

RENS crucially depends on the dynamics used in the switching protocol λ. The optimal design of switching protocols for non-equilibrium processes has recently received much attention, for example, by Nilmeier et al. [2011], Then and Engel [2008] and Sivak and Crooks [2012]. We study a continuous protocol proposed by Ballard and Jarzynski [2009] (note that we do not use an additional momentum scaling term in the equations of motion) as well as a scheme with a piecewise constant protocol as described by Opps and Schofield [2001] and Nilmeier et al. [2011]. We set $k_{\mathrm{B}}T = 1$ and parametrize the Hamiltonian by

$$H_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \frac{\boldsymbol{p}^2}{2} + U_\lambda(\boldsymbol{q}) \ . \tag{2.9}$$

The Hamiltonian is switched from $H_{\lambda(t=0)}$ to $H_{\lambda(t=\tau)}$ using three different schemes for the non-equilibrium dynamics for generating the non-equilibrium trajectories $\gamma_{\mathrm{A}\to\mathrm{B}}$ and $\gamma_{\mathrm{B}\to\mathrm{A}}$. In AMDRENS and LMDRENS, we calculate molecular dynamics (MD) trajectories employing the Andersen thermostat and Langevin dynamics, respectively. Thermostatting is neccessary to obtain the desirable scaling behaviour of RENS, that is, $p_{\mathrm{acc}} \to 1$ for $\tau \to \infty$. In HMCRENS, we employ a Markov chain Monte Carlo (MCMC) algorithm to relax the system after each perturbation. For this kind of dynamics, the thermostat is already implied in the acceptance step of the MCMC algorithm. In the following, we describe these three dynamics in greater detail.

AMDRENS: Andersen-thermostatted molecular dynamics

To generate a continously varying Hamiltonian, we choose a protocol that depends linearly on the time t, $\lambda(t) = {}^{t}/_{\tau}$. We use the velocity Verlet scheme to integrate Hamilton's equations of motion:

$$\begin{aligned} \dot{\boldsymbol{q}} &= \nabla_{\boldsymbol{p}} H = \boldsymbol{p} \\ \dot{\boldsymbol{p}} &= -\nabla_{\boldsymbol{q}} H = -\nabla_{\boldsymbol{q}} U(\boldsymbol{q}; \lambda(t)) \ , \end{aligned}$$

from $t = 0$ to $t = \tau$. We thermostat the system using the Andersen thermostat [Andersen, 1980]. In our implementation, new momenta are drawn from the Maxwell-Boltzmann distribution, $\boldsymbol{p} \sim \exp\{-\boldsymbol{p}^2/2\}$, with a certain probability p_{update} at every timestep. For Hamiltonian dynamics in combination with the Anderson thermostat, Ballard and Jarzynski [2009] define the work as

$$W = H(\boldsymbol{x}_\tau; \lambda_\tau) - H(\boldsymbol{x}_0; \lambda_0) - \sum_i Q_i \tag{2.10}$$

where $Q_i = H(\boldsymbol{x}'_{t_i}; \lambda_i) - H(\boldsymbol{x}_{t_i}; \lambda_i) = \boldsymbol{p}'^2_{t_i}/2 - \boldsymbol{p}^2_{t_i}/2$ is the heat generated by the Andersen update at time t_i; here, $\boldsymbol{x}'$ and $\boldsymbol{p}'$ denote the state $\boldsymbol{x}$ and its momentum $\boldsymbol{p}$ after the momentum update. The potential energy U does not appear in this expression, as only the momentum is changed, while the positions remain the same. For this reason, the potential energy after and before the momentum update are identical and cancel out.

LMDRENS: Langevin-thermostatted molecular dynamics

In LMDRENS, we use Langevin dynamics [Langevin, 1908] to thermostat the system during the non-equilibrium trajectories. The equations of motion for Langevin dynamics are [Sivak et al., 2013]

$$\mathrm{d}\boldsymbol{q} = \boldsymbol{p}\ \mathrm{d}t \tag{2.11}$$

$$\mathrm{d}\boldsymbol{p} = -\nabla_{\boldsymbol{q}} U(\boldsymbol{q}; \lambda(t))\ \mathrm{d}t - \gamma \boldsymbol{p}\ \mathrm{d}t + \sqrt{2\gamma}\ \mathrm{d}W(t) \tag{2.12}$$

with the friction coefficient γ and $W(t)$ a delta-correlated Gaussian process with zero mean, where we chose temperature units such that $\beta = 1/k_{\mathrm{B}}T = 1$ and unit masses for each degree of freedom.

To integrate these equations of motion, we use the integration scheme proposed by Bussi and Parrinello [2007] in the form described in Sivak et al. [2013]. This method employs the velocity Verlet integration scheme for the deterministic substeps and allows the clear separation of heat, protocol work and shadow work, that is, numerical error in the energy. We obtain the work as introduced by Sivak et al. [2013]:

$$W_{\mathrm{A}\to\mathrm{B}} = H_{\mathrm{B}}(\boldsymbol{x}_\tau) - H_{\mathrm{A}}(\boldsymbol{x}_0) - Q\ , \tag{2.13}$$

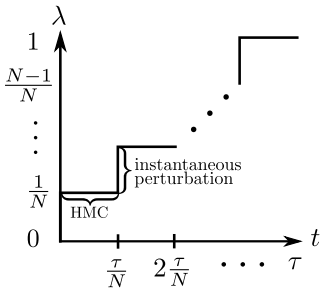

Figure 2.2: Switching protocol $\lambda(t)$ for HMCRENS. The protocol are $N = \tau/L\Delta t$ successive steps of an instantaneous perturbation, followed by a HMC propagation consisting of L MD steps with timestep ΔT and a Metropolis acceptance step

taking advantage of the (integrator-specific) easy on-the-fly calculation of the dissipated heat Q during integration. Note that the dynamics is not Hamiltonian and that H simply measures the instantaneous total energy of the system.

HMCRENS: stepwise perturbations with HMC relaxation

Following Nilmeier et al. [2011], we choose a non-equilibrium protocol that alternates between thermodynamic perturbation and MCMC relaxation steps. The Hamiltonian is the same as for AMDRENS, but the protocol varies stepwise according to

$$\lambda(t) = \sum_{n=1}^{N} \frac{n}{N} \Theta\left(t - \frac{n-1}{N}\tau\right) \Theta\left(\frac{n}{N}\tau - t\right) \ . \tag{2.14}$$

Here, $\Theta(x)$ denotes the Heaviside function and N the number of equally sized steps in the protocol. A single step consists of increasing λ instantaneously by a constant value $\Delta\lambda = 1/N$ (*perturbation*), followed by Hamiltonian Monte Carlo (HMC) [Duane et al., 1987] with a trajectory of length $L = \tau/N\Delta t$ (*propagation*). Fig. 2.2 illustrates this protocol. The work for such a trajectory is [Nilmeier et al., 2011]:

$$W = \sum_{n=1}^{N} [U_n(\boldsymbol{q}_{n-1}) - U_{n-1}(\boldsymbol{q}_{n-1})] \ . \tag{2.15}$$

Note that contributions from the kinetic energy do not appear in this expression. They cancel out because we are perturbing the system by instantaneously

changing the potential energy while leaving both temperature and the actual system state unchanged.

An early version of HMCRENS, called annealed swapping, has been introduced by Opps and Schofield [2001]. A sequential variant similar to simulated annealing as well as the estimator of the work (2.15) also form the basis of annealed importance sampling (AIS) introduced by Neal [2001].

2.3 Efficacy and scaling behavior of RENS for a coarse-grained protein model

The difficulty of conformational sampling depends on the dimensionality of the system as well as on the degree of multimodality and correlation between conformational degrees of freedom. We test RENS on various systems arising in computational biology which differ in these properties and compare it to RE.

To study the performance of RENS and RE in problems with increasing dimensionality, we consider the Gaussian network model (GNM) (see, e.g., the review by Rader et al. [2006]) for 16 proteins whose length varies between 25 and 500 amino acids. The conformational degrees of freedom are the Cartesian coordinates of the C_α atoms. Therefore, the dimensionality d of configuration space ranges between 75 and 1500. The resulting probability distribution is a multivariate Gaussian distribution and thus unimodal. In general, the degrees of freedom are coupled, because the covariance matrix of this distribution (the connectivity matrix of the GNM) is non-diagonal.

For every protein, we perform simulations with two replicas at two different force constants $k_0 = 0.3$ and $k_1 = 1.0$ as well as $k_0 = 0.85$ and $k_1 = 1.0$, respectively. The λ-dependent potential U_λ is parametrized as

$$U_\lambda(\boldsymbol{q}) = k_\lambda U_{\mathrm{GNM}}(\boldsymbol{q}); \quad k_\lambda = \lambda\, k_2 + (1-\lambda)\, k_0 \ . \tag{2.16}$$

where

$$U_\lambda(\boldsymbol{q}) = \sum_{ij} [\boldsymbol{\Gamma}^{-1}]_{ij} (\boldsymbol{q}_i - \boldsymbol{q}_i^0) \cdot (\boldsymbol{q}_j - \boldsymbol{q}_j^0)$$

and Γ_{ij} is the Kirchhoff matrix based on the native structure $\boldsymbol{q}_i^0$ and a cutoff distance of 7.5 Å.

To assess the performance of the RENS variants, it is necessary to first minimize the impact of insufficient equilibrium sampling between the exchange attempts. This is achieved by realizing that under the GNM the Cartesian coordinates $\boldsymbol{q}_i$ follow a multivariate Normal distribution with covariance matrix $\boldsymbol{\Gamma}$. Uncorrelated equilibrium samples can therefore be generated between exchange attempts by using random number generators for the multivariate Normal distribution.

To compare the three switching protocols, we set up the length of the non-equilibrium trajectories such that it amounts to a fixed number of integration steps. In AMDRENS, we run 150/3000 integration steps for $k_0 = 0.85/k_0 = 0.3$ using the Andersen thermostat and a linear switching protocol $\lambda(t) = t/\tau$. In case of LMDRENS, we chose the same linear switching protocol and integrate the stochastic Langevin equation by running the integrator described in [Bussi and Parrinello, 2007; Sivak et al., 2013] for the same number of steps. For HMCRENS, we have to choose the number of intermediate steps during the switching process. Test simulations show that the highest acceptance rate is reached when we divide the non-equilibrium trajectories into $N_{\mathrm{k_0=0.3}} = L_{\mathrm{k_0=0.3}} = 3000$ and $N_{\mathrm{k_0=0.85}} = L_{\mathrm{k_0=0.85}} = 150$ steps, respectively. That is, after each perturbation step we relax the state by running HMC with a single leapfrog step. That using as many intermediate steps as possible is advantageous has also been empirically confirmed in the context of NCMC [Chodera, 2012; Nilmeier et al., 2011]. We should note that HMC requires energy evaluations to calculate the acceptance probability and the leap-frog scheme employed actually takes two force evaluations when only taking one step. These additional energy / force evaluations were minimized in our implementation by re-using already calculated values whenever possible but nonetheless, a N-steps HMCRENS

trajectory is computationally more expensive than a N-steps AMDRENS or LMDRENS trajectory.
Care has to be taken when choosing the integration timesteps used in the non-equilibrium trajectories. While a small timestep results in little numerical error, states also do not change much and thus don't adapt to the ensemble change, which leads to low acceptance rates. But also large timesteps are problematic, because numerical error increases, and in the case of HMCRENS additionally results in low acceptance rates of the relaxation steps. For this reason, timesteps for all RENS variants were determined by running trial simulations for each protein spanning a range of values and choosing the timestep yielding the highest acceptance rates. In general, this optimal timestep is different for each protein.

It is interesting to examine the work which determines the acceptance probability for RENS. By averaging the total work performed on the system during the non-equilibrium trajectory and dividing this value by the number of residues, we can calculate a work per residue, which gives us a direct insight into the scaling behavior of RENS. As RE can be regarded as the $\tau \to 0$ limit of RENS, we also plot the RE "work" $W_{\mathrm{RE}} = U_{\mathrm{B}}(\boldsymbol{x}_{\mathrm{A}}) - U_{\mathrm{A}}(\boldsymbol{x}_{\mathrm{A}}) + U_{\mathrm{A}}(\boldsymbol{x}_{\mathrm{B}}) - U_{\mathrm{B}}(\boldsymbol{x}_{\mathrm{B}})$ of the RE simulations. In Fig. 2.3, RE and both RENS variants show within standard deviation an approximately constant work per residue consistent with the above analysis of the acceptance rates. The work per residue for HMCRENS and AMDRENS are very similar and much lower than the work per residue for RE. This is the reason for the generally higher HMCRENS / AMDRENS swap acceptance rates in Fig. 2.3

The approximately constant work per residue for both all RENS variants and RE reflects the fact that the work is an extensive quantity and that in both RE and RENS we indeed have to expect acceptance rates to decrease exponentially with increasing number of degrees of freedom.

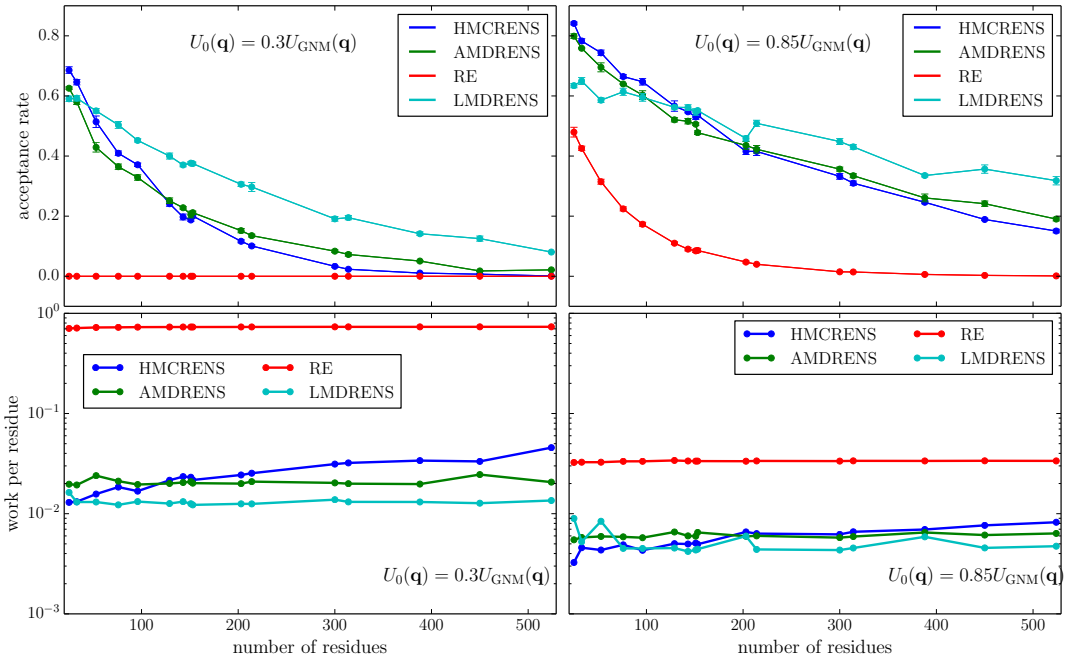

Figure 2.3: RENS behaviour for different system sizes and different phase space overlaps: Gaussian Network Model (GNM) simulation of sixteen proteins with different number of residues. Two replicas defined by $U_\lambda(q) = k_\lambda U_{\mathrm{GNM}}(q)$ have a different phase space overlap defined by global force constants $k_0 = 0.3$ (low phase space overlap, *left*), $k_0 = 0.85$ (higher phase space overlap, *right*) and $k_1 = 1.0$.

Top: average acceptance rates. *Bottom*: average work per residue. Due to the log-scale, error bars are not shown. Standard deviations decrease with higher residue number and range from $\sigma_{\mathrm{min}} \approx 5 \times 10^{-3}$ to $\sigma_{\mathrm{max}} \approx 0.175$, the latter attained for RE at $k_0 = 0.3$ and $n_{\mathrm{res}} = 25$. Approximately constant values show that the work as defined in RENS is an extensive quantity.

2.4 RENS in the context of a complex ISD posterior distribution

The last test system is a posterior probability distribution arising in ISD (Sec. 1.2. Currently conformational samples are generated from the posterior distribution using replica exchange Monte Carlo with two control parameters [Habeck et al., 2005]. The conformational degrees of freedom q_i are the torsion angles that parameterize rotations about covalent bonds.

The posterior probability distribution over all torsion angles is:

$$p(\boldsymbol{q}|D, I) \propto p(D|\boldsymbol{q}, I)\, p(\boldsymbol{q}|I)$$

where D denotes the experimental data and I the prior knowledge. The posterior probability is proportional to the product of the likelihood function $(D|\boldsymbol{q}, I)$ (the probability of observing the actual data if the conformation is $\boldsymbol{q}$) and the prior probability $p(\boldsymbol{q}|I)$. The prior probability is the Boltzmann distribution $p(\boldsymbol{q}|I) \propto \exp\{-E(\boldsymbol{q})\}$ based on a simple non-bonded force field $E(\boldsymbol{q})$. Because we use ISD only to benchmark the RENS algorithm, we do not take full advantage of its capability to determine statistically well-defined and unbiased structure ensembles. That is, we fix additional model parameters such as data weights to values obtained in previous simulations and sample only the conformational degrees of freedom.

The posterior distribution $p(\boldsymbol{q}|D, I)$ is high-dimensional and multimodal; moreover, the torsion angles are highly coupled. Therefore $p(\boldsymbol{q}|D, I)$ is a challenging real-world application of RENS and RE. In contrast to the GNM for which one could directly draw equilibrium samples, the ISD posterior can only be sampled using Monte Carlo or MD simulations.

To implement RE and RENS for this problem, we use a replica schedule in which two parameters, ν and α, control the weight of the data and of the force field, respectively [Habeck et al., 2005].[1] To calculate the trajectories neccessary for a RENS swap attempt between two replicas A and B, defined

[1] Note that in Habeck et al. [2005] the control parameters ν and α are denoted λ and

by parameters $(\nu_\mathrm{A}, \alpha_\mathrm{A})$ and $(\nu_\mathrm{B}, \alpha_\mathrm{B})$, the λ-dependent potential is then parameterized by

$$U_\lambda(\boldsymbol{q}) = \lambda\, U(\boldsymbol{q}; \nu_\mathrm{B}, \alpha_\mathrm{B}) + (1-\lambda)\, U(\boldsymbol{q}; \nu_\mathrm{A}, \alpha_\mathrm{A}) \tag{2.17}$$

where

$$U(\boldsymbol{q}; \nu, \alpha) = -\nu \log p(\boldsymbol{q}|D, I) + \frac{\alpha}{\alpha - 1} \log[1 + (\alpha - 1)\, E(\boldsymbol{q})]$$

is the modified log posterior distribution based on the Tsallis ensemble [Tsallis, 1988] as prior probability.

In HMCRENS, we use a protocol that is comprised of 200 switching steps, each of which consist of a perturbation step and a relaxation step. In the perturbation step, the replica parameters ν, α are increased / decreased linearly. The relaxation step consists of an HMC step using 100 leapfrog steps. AMDRENS was implemented similarly as for the GNM simulations, that is, the replica parameters were switched continuously during a thermostatted MD trajectory comprising 2×10^4 steps. In both HMCRENS and AMDRENS, 2×10^4 integration steps are performed in total. Again, note that HMCRENS needs more force / energy evaluations than AMDRENS.

To make the problem harder, we use only 7 replicas instead of the aforementioned 10–20 replicas. We expect the RE simulations to have difficulties relaxing to the native state of the protein reasonably fast due to low exchange rates, but our results show that exchange rates of only a few percents are sufficient and thus RE still performs reasonably well. Nevertheless we expect RENS to achieve higher exchange rates because of increased phase space overlap.
In Fig. 2.4, we first look at the acceptance rates. All RENS variants beat RE by accepting more swaps, with LMDRENS performing best, albeit with large variations within the five equivalent simulations.

Fig. 2.6 shows the evolution of the total ensemble pseudo-energy $E(\boldsymbol{x}) = \sum_{i=1}^{7} E_i(\boldsymbol{x}_i) = -\sum_{i=1}^{7} \log p_i(\boldsymbol{x}_i)$ averaged over five independent simulation

q, respectively. However, to avoid confusion with the switching parameter $\lambda(t)$ and the conformational degrees of freedom $\boldsymbol{q}$, we denote the data weight ν and the Tsallis parameters α.

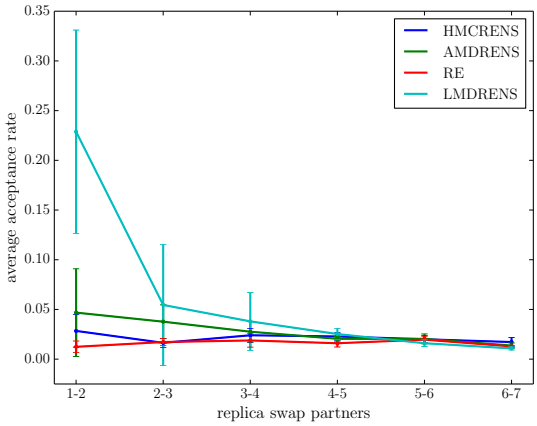

Figure 2.4: Acceptance rates averaged over five simulations in which 1.5×10^{-3} samples are drawn from the ISD posterior distribution for 1UBQ conditioned on all nuisance parameters. Errorbars show the standard deviation.

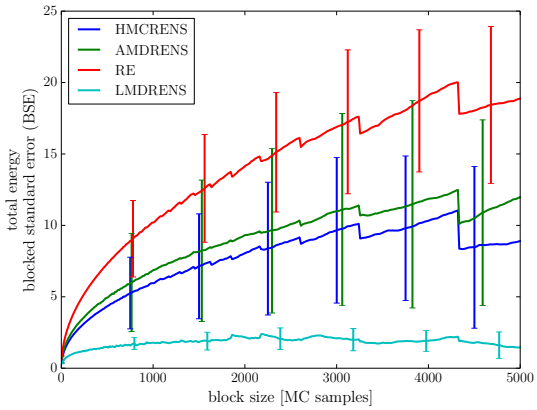

Figure 2.5: Blocked standard error of the total ensemble energy. Errors are averaged over five simulations in which 1.5×10^{-3} samples are drawn from the ISD posterior distribution for 1UBQ conditioned on all nuisance parameters. The subset of errorbars shows the standard deviation.

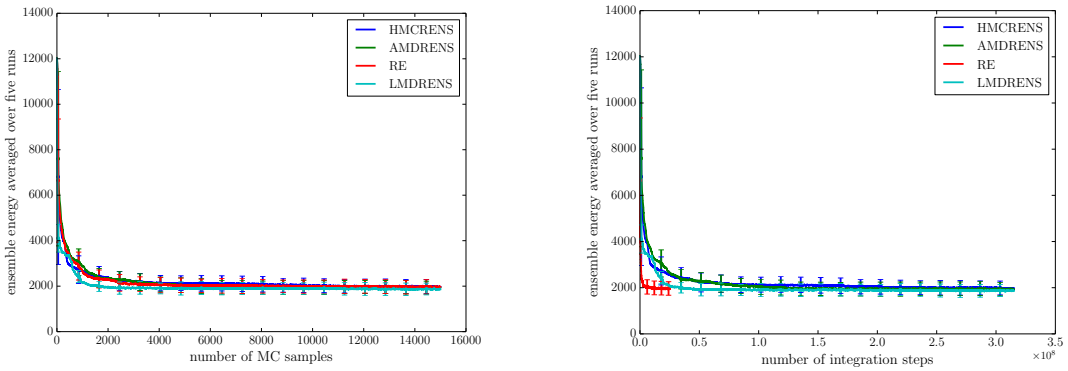

Figure 2.6: Total ensemble pseudo-energy averaged over five independent runs. Ensembles consist of 1.5×10^{-3} samples from the ISD posterior for 1UBQ. In each RE(NS) simulation, 7 replicas were simulated. Energies were plotted against the number of MC samples (*Left*) and the (approximate) number of integration steps performed (*Right*). The subset of errorbars shows the standard deviation between five independent simulations.

runs. LMDRENS is able to relax significantly faster to the global minimum with respect to the total number of MC samples. As the error bars show, this effect is reproducible over several independent runs converging to the same mean total energy, which gives us the confidence that the simulations are indeed converged. TMDRENS, HMCRENS and AMDRENS perform similar to RE. This is in agreement with LMDRENS having the clearly higher acceptance rates.

To assess the sampling quality achieved when sampling from a probabiltiy distribution $p(x)$, we employ block averaging [Flyvbjerg and Petersen, 1989; Grossfield and Zuckerman, 2009]. This method makes use of the standard error

$$SE(A) = \frac{\sigma(A)}{\sqrt{(N)}} \; ;$$

it is the standard deviation between estimates of $A(x)$ based on independent samples x divided by the total number on independent samples from which the standard deviation is calculated. As such it is a measure of the accuracy of method employed to obtain the samples.

Block averaging then works by calculating a running average of the standard

error over blocks of different sizes. A trajectory is divided in equally-sized blocks, within which the standard error $SE(A)$ is calculated. This process is repeated for systematically increasing block sizes. Once the block sizes are large enough that samples within consecutive blocks are not correlated anymore, the standard error will plateau at its real value. Block averaging thus gives an estimate of correlation times. It is important to note that this procedure is valid only for a dynamical simulation, that is, for samples which are indeed time-correlated. For this reason, it does not make sense to consider, for example, the energy of the target ensemble in a replica exchange simulation. But we can use the total ensemble energy as a quantity which we use to measure convergence, as the super-Markov chain mentioned in Sec. 1.3.4 does indeed consist of sequentially correlated samples.

Calculating the corresponding block standard errors for the total ensemble energy shows that convergence is faster and correlation times smaller for all RENS variants, with LMDRENS again leading the field. It is interesting to see that according to this analysis, the other methods might actually fail to produce completely decorrelated samples within blocks of 5×10^3 MCMC samples.

Yet, faster convergence and better sampling quality come at a high cost: we choose the length of the non-equilibrium trajectories to be 200 steps consisting of a perturbation and a 100 step HMC relaxation for HMCRENS, which makes a total of 2×10^4 integration steps per trajectory. The same number of integration steps was performed for each of the TMDRENS and LMDRENS trajectories. This means that for all RENS variants, the computational expense per MC sample is much higher than for RE. We choose the (approximate) number of leapfrog integration steps performed as a measure for computational effort and plot the target ensemble energy as a function of the performed integration steps in the right diagram of Fig. 2.6. RE is clearly more efficient than all of the RENS variants, followed by LMDRENS, which performs much worse, but much better than the other RENS variants. We suggest therefore LMDRENS as the most promising candidate for trying to beat RE in larger systems by optimizing the switching protocol $\lambda(t)$ or by using other benefits of RENS like

a possible "waste recycling" of the non-equilibrium trajectories to estimate free energies by statistical reweighting [Hummer and Szabo, 2001].

3 Outlook on an adaptive Replica Exchange scheme

While the idea of Replica Exchange (RE; Sec. 1.3.4) is relatively simply, its practical application often is not trivial. While a high-temperature distributions is easy to pick, it is neither clear how many intermediate distributions have to be simulated nor how they should be related to the high- and low-temperature distributions. Both factors affect the efficiency: more replicas necessarily require more computation time and suboptimal schedules may will exhibit drops in acceptance rate, hindering the quick diffusion of states from the high-temperature replica to the target replica.
In this chapter, we borrow a result, which applies to optimal estimation of normalization constants, to outline a fully automatic scheme to determine intermediate distributions, which possibly result in better sampling quality as compared to simple schedules.

3.1 Previous approaches on optimizing Replica Exchange schedules

Considering the wide-spread applications of RE, it is not surprising that many efforts have been spent on finding general recipes for optimal schedules. Two principal approaches exist: we could aim for uniform acceptance rates between all replicas, making sure that ascending and descending on the temperature ladder spend an equal amount of time in each replica. Kofke [2002] show that,

wehen simulating Boltzmann ensembles, this can be achieved using a geometric progression of temperatures. Their result rests on the strong assumption that the constant volume heat capacity is constant across the whole temperature range, which excludes system exhibiting phase transitions in the temperature range of interest. There seems to be a consensus that an acceptance rate of around 20 % yields optimal performance [Kone and Kofke, 2005; Rathore et al., 2005]. General, iterative approaches to obtain a specific uniform acceptance rate have been developed [Rathore et al., 2005; Schug et al., 2004]. A recently proposed, general framework [Habeck, 2015] is capable of constructing optimal schedules based on minimizing the relative entropy (Kullback-Leibler divergence) between neighbouring replicas and, at the same time, produces samples and an estimate of the density of states. These schedules also result in approximately uniform acceptance rates.

The second approach is to focus on the mixing between the lowest and the highest-temperature replica and thus to maximize the number of round trips between between them [Katzgraber et al., 2006]. While not an adaptive scheme, a method described by Spill et al. [2013] tries to enforce round trips by violating the detailed balance condition.

In the approach we describe in this chapter, we do not assume any knowledge about the system / probabiltiy distribution under consideration and thus believe it may be suitable candidate for a general, adaptive scheme, which also yields the density of states as a useful by-product.

3.2 Optimal interpolating distributions for free energy estimation by thermodynamic integration

In Bayesian analysis, the evidence of data D is the normalization constant of the posterior distribution. Its sibling in statistical physics (in the canonical ensemble) is the partition function $Z(\beta) = \int \mathrm{d}x \exp\left[-\beta E(x)\right]$ of a Boltz-

mann distribution $p(x) = {}^{1}/_{Z(\beta)} \exp\left[-\beta E(x)\right]$ at inverse temperature β. As the knowledge of the partition function allows us to calculate every thermodynamic quantity of interest such as free energies, heat capacities and so on, it is of primary interest in computational physics. As such it comes as no surprise that several methods have been devised to estimate partition functions or rather their negative logarithms, namely, free energies, from MD or MCMC simulations. In practice, though, one is only interested in the difference between free energies. Perhaps the most popular among these methods is thermodynamic integration (TI, Smit and Frenkel [2002])), which is based on the identity

$$\Delta F = \int \mathrm{d}\lambda \, \left\langle \frac{\partial}{\partial \lambda} E_\lambda(\mathbf{q}) \right\rangle \ . \qquad \text{(TI identity)}$$

Here, F denotes the free energy difference between two different canonical ensembles defined at the same inverse temperature β and two different values $\lambda = 0, 1$. This is reminiscent of the non-equilibrium processes discussed in Sec. 2.1 and indeed, the TI identity is the special case of the Jarzynski equality for infinitely slow switching. Another popular method to approximate free energy differences is thermodynamic perturbation, which can be seen to be the opposite limit of instantaneous switching of the Jarzynski equality. In practice, TI consists of simulating the system of interest at closely-spaced values of λ and calculating the expectation value for all λ's using the samples obtained. Then, the integral in the TI identity can be approximated using quadrature or the trapezoidal rule. TI has been known among physicists since the 70's [Gelman and Meng, 1998], but seems to have been independently rediscovered by Ogata [1989] and was not known in the statistics community. Gelman and Meng [1998] point out this lack of communication between different fields and analyse TI from a statistical point of view. Of particular interest to us is their discussion of the problem of optimal interpolating distributions $q_\lambda(x) = p_\lambda(x) Z_\lambda$, which immediately reminds us of the quest for optimal schedules in RE(NS) methods. We now follow closely Gelman and Meng while adapting their notation. Gelman and Meng's interpolating distributions are optimal in

the sense that they minimize the variance of the Monte Carlo estimator

$$\widehat{\Delta F} = \frac{1}{n} \sum_i \partial_\lambda E_{\lambda_i}(\mathbf{q}_i) \ ,$$

where a uniform distribution of λ_i is assumed. Any other distribution $p(\lambda_i)$ could be absorbed in the interpolating distributions q_{λ_i}.

The variance of this unbiased estimator is

$$\text{var} \ \left(\widehat{\Delta F}\right) = \frac{1}{n} \left[\int_0^1 \int \mathrm{d}\lambda \mathrm{d}\mathbf{q} \ (\partial_\lambda E_\lambda(\mathbf{q}) \ )^2 - \Delta F^2 \right] \ . \tag{3.1}$$

In the calculation to obtain this result covariances have to equal zero, so the assumption of independent samples is indeed key to the following results. We are now looking for the family of interpolating distributions $q_\lambda(\mathbf{q}) = \exp\left[-E_\lambda(\mathbf{q})\right]$ which minimizes var $\left(\widehat{\Delta F}\right)$. This is equivalent to minimizing the integral in Eq. 3.1, which we rewrite using the product rule as

$$\int_0^1 \int \mathrm{d}\lambda \mathrm{d}\mathbf{q} \ (\partial_\lambda E_\lambda(\mathbf{q}) \ )^2 = \int_0^1 \mathrm{d}\lambda \ \left[\frac{\mathrm{d}}{\mathrm{d}\lambda} \log Z(\lambda)\right]^2 + \int_0^1 E_\lambda \left[\frac{\mathrm{d}}{\mathrm{d}\lambda} \log p_\lambda(\mathbf{q})\right]^2 \ . \tag{3.2}$$

We can now minimize each term on the l.h.s seperately. For the first term, the Cauchy-Schwarz inequality yields

$$\int_0^1 \mathrm{d}\lambda \left[\frac{\mathrm{d}}{\mathrm{d}\lambda} \log Z_\lambda\right]^2 \geq \Delta F^2$$

with the equality for

$$Z^*_\lambda \propto \exp(\lambda \Delta F) \ . \tag{3.3}$$

The integrand in the second term on the r.h.s. of Eq. 3.2 equals the Fisher information $I(\lambda)$ of the probability density $p_\lambda(\mathbf{q})$. Defining

$$\alpha_H = \arctan\left[\frac{H(p_0,p_1)}{\sqrt{4 - H^2(p_0,p_1)}}\right]$$

with the Hellinger distance $H(p_0,p_1) = \left[\int \mathrm{d}\mathbf{q} \left(\sqrt{p_1(\mathbf{q})} - \sqrt{p_0(\mathbf{q})}\right)^2\right]^{1/2}$ between the probability distributions $p_{0/1}(\mathbf{q})$, one can prove the inequality

$$\int_0^1 \mathrm{d}\lambda \ I(\lambda) \geq 16 \alpha_H^2 \ . \tag{3.4}$$

The equality is given by a lengthy expression, which, combined with Eq. 3.3, yields the optimal unnormalized interpolating densities

$$q_\lambda(\mathbf{q}) \propto \mathrm{e}^{\lambda \Delta F} \left\{ \sqrt{\frac{q_0(\mathbf{q})}{Z_0}} \left[a\left(Z_{0/1}\right) - b\left(Z_{0/1}\right) \right] + \sqrt{\frac{q_1(\mathbf{q})}{Z_1}} \left[a\left(Z_{0/1}\right) + b\left(Z_{0/1}\right) \right] \right\}^2 \tag{3.5}$$

with

$$a\left(Z_{0/1}\right) = \frac{\cos\left[\alpha_H\left(Z_{0/1}\right)(2\lambda - 1)\right]}{2\cos\left[\alpha_H\left(Z_{0/1}\right)\right]} ,$$
$$b\left(Z_{0/1}\right) = \frac{\sin\left[\alpha_H\left(Z_{0/1}\right)(2\lambda - 1)\right]}{2\sin\left[\alpha_H\left(Z_{0/1}\right)\right]} .$$

In these equations the dependence of a, b, α_H on the evidences Z_0, Z_1 is made explicit, because it is important to see that this expression for the interpolating distributions yielding a minimal variance of $\widehat{\Delta F}$ not only involves a difficult integral

$$G = \int \mathrm{d}\mathbf{q} \exp\left\{ -\frac{1}{2} \left[E_0(\mathbf{q}) + E_1(\mathbf{q}) \right] \right\} \tag{3.6}$$

to calculate the Hellinger distance, but also already requires the knowledge of the evidences, whose determination was the original goal of Gelman and Meng's analysis. They stop at this point and use the lower bound for the variance as a reference to compare other estimators to.

3.3 Illustration of interpolating distributions for an analytically tractable system

An easy toy system, for which free energies and Hellinger distances can be calculated analytically, is a pair of normal distributions with different means $\mu_{0/1}$ and standard deviations $\sigma_{0/1}$. The partition functions are given by

$$Z_{0/1} = \sqrt{2\pi\sigma_{0/1}^2}$$

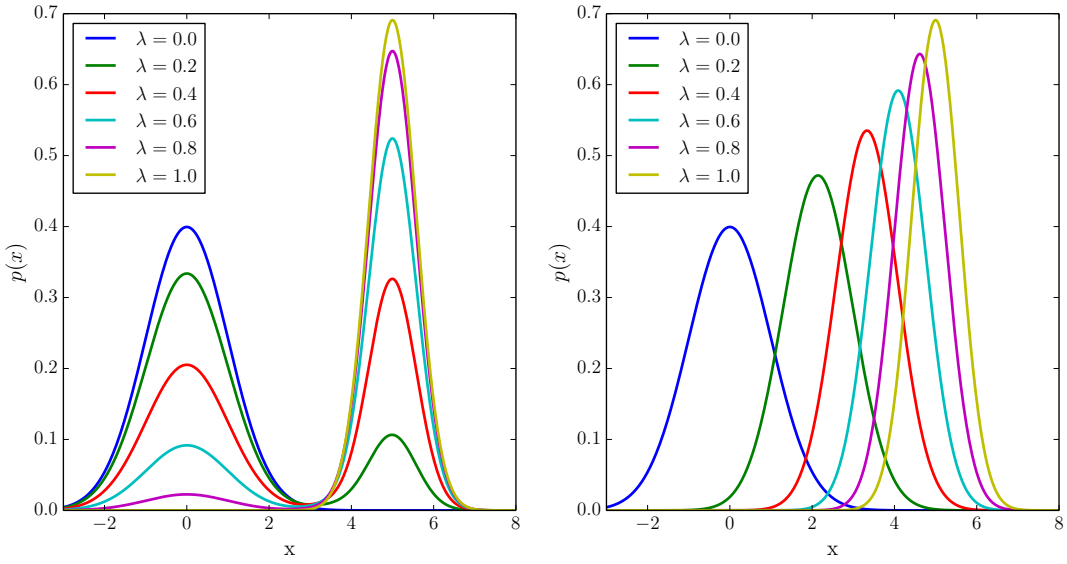

Figure 3.1: Interpolating distributions between two normal distributions with means $\mu_{0/1} = 0, 5$ and standard deviations $\sigma_{0/1} = 1, 1/\sqrt{3}$.
Left: distributions minimizing the variance of the free energy estimator. *Right*: linear interpolation.

and the integral defined in Eq. 3.6 is Gaussian;

$$\begin{aligned} G &= c \int \mathrm{d}q \ \exp\left[-\frac{1}{2\sigma_{\mathrm{tot}}^2}(x-\mu_{\mathrm{tot}})^2\right] \\ &= c\sqrt{2\pi\sigma_{\mathrm{tot}}^2} \ . \end{aligned}$$

The mean, standard deviation and constant are given by

$$\begin{aligned} \mu_{\mathrm{tot}} &= \frac{\mu_0\sigma_1^2 + \mu_1\sigma_0^2}{\sigma_0^2 + \sigma_1^2} \ , \\ \sigma_{\mathrm{tot}} &= \sigma_0\sigma_1\sqrt{\frac{2}{\sigma_0^2 + \sigma_1^2}} \ , \\ \log c &= -\frac{1}{4}\left[\frac{\mu_0^2}{\sigma_0^2} + \frac{\mu_1^2}{\sigma_1^2}\right] + \frac{1}{2\sigma_{\mathrm{tot}}^2}\mu_{\mathrm{tot}}^2 \ . \end{aligned}$$

The resulting intermediate distributions are plotted in Fig. 3.1. On the one hand, compared to a linear interpolation of the log-probabilities, the statistically optimal distributions have a consistently higher overlap, which would lead to higher acceptance rates in a RE. On the other hand, the interpolating distributions are bimodal and thus harder to sample from. This is due to the mixture nature of the interpolating distributions: if the two target distributions have very pronounced modes, we can expect them to persist in the

interpolations. One possibility to efficiently sample from such a mixture is using a Gibbs sampling scheme (Sec. 1.3.2): first, we draw a component of the three-component mixture with a probability given by the weights in Eq. 3.5 and then sample from the chosen component using any suitable MCMC sampling scheme (Sec. 1.3), or, in this case, built-in random number generators. Another way to sample from this mixture might be the use of non-equilibrium MCMC methods such as NCMC [Nilmeier et al., 2011], in which a distant proposal could be obtained by flattening the energy landscape to escape from the current mode.

3.4 Exact free energy estimates as a criterion for sampling quality?

While interpolating distributions derived above are proven to give, at least in theory, the best estimates of free energy differences, it is not immediately obvious that they constitute also an efficient RE schedule. But in our opinion, there are good arguments that this is indeed the case.

Free energies are intimately connected with state populations. Only well populated states have enough probability weight to contribute significantly to the free energy of a system. So if we are able to estimate free energies correctly, we can be optimistic to also get the state populations right.

Daniel Zuckerman and co-workers are looking for methods to reliably quantify sampling quality, which is not an easy task, but nonetheless obviously extremely important to measure progress in sampling methods. The usual approch would be to calculate an effective sample sizes (ESS) by estimation of a decorrelation time. But of what variable? This is a fundamental problem, as the variable needs to capture the slowest time scales in the system lest it underestimate the correlation time. Furthermore, correlation time estimation usually requires a dynamical trajectory, that is, a trajectory, in which subsequent states are in fact time-correlated, as already pointed out in the discussion of block averaging in Sec. 2.4. To alleviate these issues, they suggest

in [Zhang et al., 2010] to quantify the degree of sampling by analyzing the variances of state populations, from which an objective ESS can be estimated. This is justified as physical states are by definition representative of the slowest timescales in a system. In their work, Zhang et al. discretize configuration space in bins and, by a hierarchical method relying on transition rates, find bins representing states. The ESS is then, following earlier work Lyman and Zuckerman [2007], estimated by realizing that a variance of a state population can be estimated from a known number of independent samples by means of binomial statistics, but also the other way round. The success of this approach described in [Zhang et al., 2010] gives us confidence that a scheme, which aims to determine free energies, that is, log-ratios of partition functions to which state populations contribute most, will also propose interpolating distributions which assure good sampling. If then the number of intermediate distributions is chosen such that exchange acceptance rates stay above a reasonable minimum to ensure good mixing, we should be safe. But finally, the utility of such a scheme can only be judged in an actual application.

3.5 Iterative determination of optimal mixture weights

In many applications, such as ISD, we can neither easily normalize the probability distribution we wish to sample from nor can we calculate the Hellinger distance to a different distribution analytically. To be able apply the possibly optimal RE schedules in full generality, we thus need a way to approximate both the normalization constants and the Hellinger distance. We now observe that both quantities involve integrals, whose integrand only depends on the random variable through the energy $E(x) = -\log p(x)$ and thus can be reformulated as energy integrals using a (multidimensional) density of states (Eq.

1.16):

$$Z_{0/1} = -\log \int \mathrm{d}\mathbf{q}\ \mathrm{e}^{-E_0(x)} = -\log \int \mathrm{d}E_0\ \mathrm{e}^{-E_0}\ , \tag{3.7}$$

$$G = \int \mathrm{d}\mathbf{q}\ \mathrm{e}^{-\frac{1}{2}[E_0(\mathbf{q})+E_1(\mathbf{q})]} = \int\int \mathrm{d}E_0\ \mathrm{d}E_1\ \Omega(E_0, E_1)\mathrm{e}^{-\frac{1}{2}(E_0+E_1)}\ . \tag{3.8}$$

By using WHAM (Sec. 1.4) we should then be able to obtain estimates for both quantities based on samples from all replicas.

These considerations immediately suggest to iteratively solve Eq. 3.5 by starting a RE simulation with a fixed number of replicas and initial interpolating distributions q_λ^0, drawing samples from them and using these samples to get a first estimate q_λ^1 of the interpolating distributions. The simulation is then continued with the freshly calculated schedule to yield the next approximation q_λ^2 using all previously drawn samples to estimate an updated, more accurates DOS. We would then expect this iterative scheme to converge to the optimal schedule, at which point the iterative procedure can be halted and the productive RE run can be started with the now optimized schedule. The advantages of this scheme are obvious: as WHAM allows the inclusion of samples from all iterations, we expect the weights to converge quickly and furthermore, if the purpose of the RE simulation is a Bayesian analysis, a good DOS estimate and thus the evidences and other high-dimensional integrals over the posterior density come for free.

Unfortunately, at the moment, it is not clear if the WHAM equations (Eq. 1.19 and 1.20) can be modified to accomodate the fact that the mixed component in Eq. 3.5 is only known up to a constant (the multiplication of two probability densities will in general not again be a normalized density). This will be key for an implementation of our method working under completely general conditions. An important caveat is that the derivation just presented relies on the strong assumption of having drawn independent samples. As samples obtained by MCMC methods are, to a certain extent, always correlated, it will not be possible to actually reach the lower bound given in Eq. 3.4. But we can still hope for our iterative scheme to prove superior to heuristic methods.

3.6 A first test on a one-dimensional toy system

For a proof-of-concept that the discussed optimal interpolations are indeed advantageous replica schedules, we consider a simple one-dimensional system. We bypass the described technical hurdles in applying WHAM to unnormalized densities and instead determine normalizations and the Hellinger distance using quadrature.

We take as our target distribution a one-dimensional energy landscape $E_0(x)$ adapted from Smit and Frenkel [2002]. Due to the four minima, this distribution is a challenge for simple MCMC algorithms like a Random Walk Metropolis-Hastings scheme (Sec. 1.3.1). Its minima are at $[0, \pi, 2\pi, 3\pi]$. Using RE, we try to enhance sampling by interpolating between $p_0(x) = \exp[-E_0(x)]$ and a normal distribution with mean $\mu = {}^{3\pi}/_{2}$ and standard deviation $\sigma = 5$, which is easy to sample. Local sampling is performed by a random walk Metropolis-Hastings sampler; for each replica, stepsizes are adjusted to give an acceptance rate of 50 % in a preliminary run. Fig. 3.2 shows the statistically optimal interpolations and a linear log-probability interpolation $E_\lambda(x) = \lambda E_1(x) + (1-\lambda)E_0(x)$, which we note to be quite similar.

To assess sampling quality, we first use a block-averaging analysis (Fig. 3.3) as discussed in Sec. 2.4. Unfortunately, the results disappoint our hopes of better sampling quality by using optimized interpolations. We next run ten independent simulations and calculate the average standard error for the mean log-probability of the target distribution $p_0(x)$, that is, rough energy landscape. As independent samples we take the mean log-probabilities calculated from the single, independent simulations. Comparison of standard errors then should tell us which interpolating distributions produce the more accurate estimate of $\langle \log p_0(x) \rangle$. While the difference in standard errors is small ($\mathrm{SE}(\log p_0) \approx 2.49 \times 10^{-3}$ for the linearily interpolating schedule and $\mathrm{SE}(\log p_0) \approx 2.53 \times 10^{-3}$ for the statistically optimal interpolation), this test, too, indicates a slightly

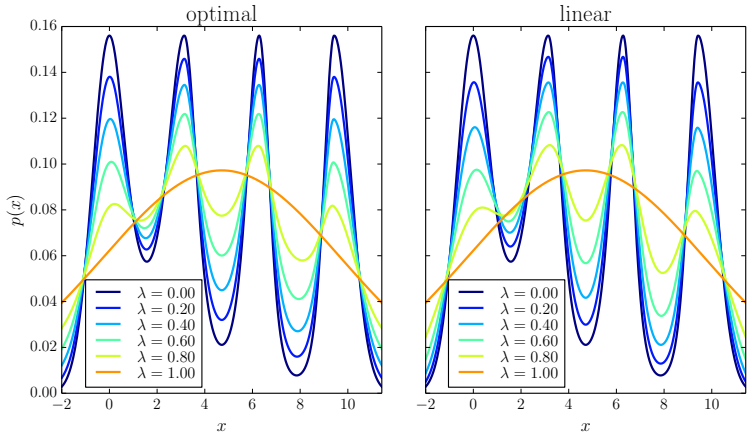

Figure 3.2: Optimally (*left*) and linearly (*right*) interpolating distributions between a simple four-minima energy landscape and a normal distribution for $N = 6$ replicas.

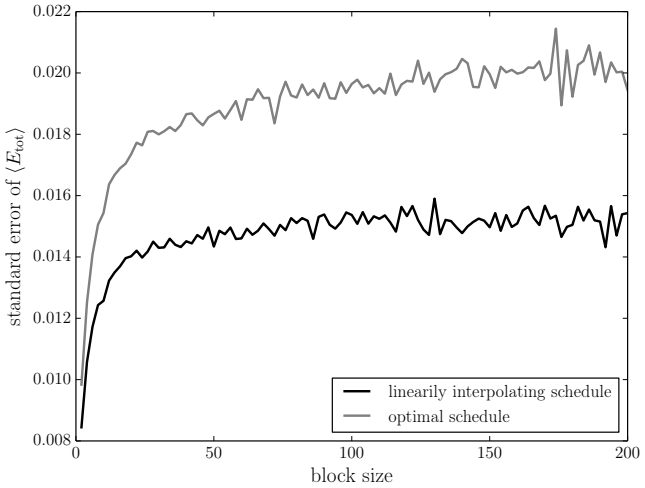

Figure 3.3: Comparing sampling quality when using optimal and linearly interpolating schedules by block averaging analysis. Standard errors are over the total ensemble energy.

better sampling quality for the linear log-probability interpolation.

This result is surprising; Gelman and Meng [1998] specifically state that the linear log-probability interpolation cannot be optimal, but nevertheless seems to be superior in sampling quality, at least in our simple example. It is also interesting to consider the RE acceptance rates: they are consistently and comparably high (between 87% and 93 %) for both interpolation schedules, but the calculations with the statistically optimal schedule exhibit a somewhat lower acceptance rate (77 %) between $p_{\lambda=0.8}$ and p_1. It could thus be that this drop is mainly responsible for the measured difference in sampling quality.

4 Bayesian structure determination from HiC data

After having investigated methods to improve the sampling of difficult posterior distributions, we now turn our attention to an application and extension of the ISD framework on the problem of inferring probable structures of interphase chromatin. While our approach does not neccessitate the advanced RE methods discussed above because of very sparse data and limitation to a maximum number of two chromosomes, its application to data spanning the whole genome would significantly increase the complexity of the structure determination problem and then likely profit from an optimized, efficient sampling scheme.

Although the focus of this work is methodology, we briefly sketch a few relevant facts about genome architecture in mammalian cells to set the scene and put the following work in a biological context.

4.1 Genome architecture in the mammalian nucleus

The total size of the human genome is around 3 Gbp. One basepair has a spatial extension of roughly 3.4 Å, thus the total length of the double-stranded DNA in a human genome, accounting for diploidicity, is approximately two meters. Yet, it is packed in a nucleus with a diameter of a few µm. Thus, DNA in the nucleus is extremely compacted. Nature has adopted several mechanisms to

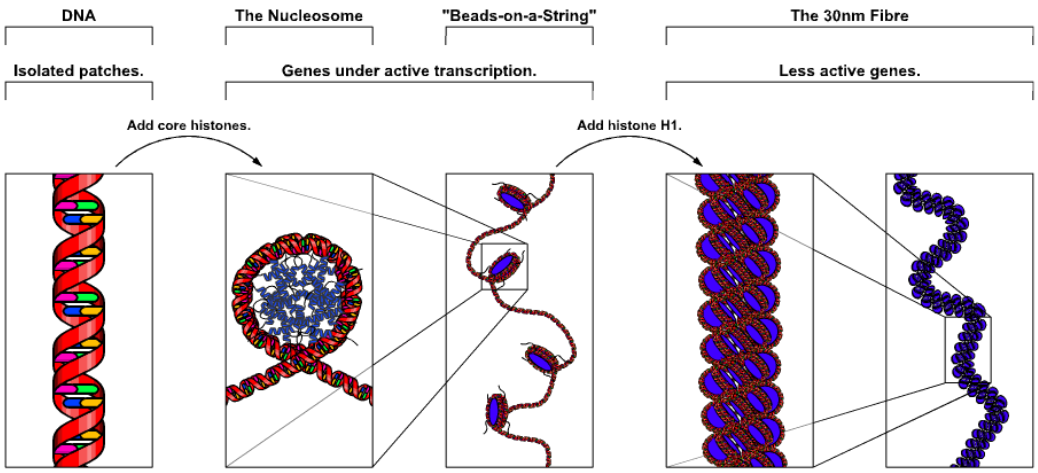

Figure 4.1: Different levels of DNA compaction in the interphase nucleus. Adapted from an image by Wikipedia user Zephyris (Richard Wheeler); https://en.wikipedia.org/wiki/User:Zephyris

achieve this which are summarized in Fig. 4.1 and which we describe following the corresponding section in the book by Hames and Hooper [2011]. The first level of DNA compaction is the wrapping of double-stranded DNA around an octamer of basic, evolutionarily very conserved proteins called *histones*. 140–150 bp of DNA are wrapped around one histone complex and form a *nucleosome*. Nucleosomes are connected by stretches of linker DNA with lengths between 50–70 bp and have a diameter of 11 nm. DNA at this level of compaction resembles very much beads (nucleosomes) on a string (linker DNA), is relatively accessible for transcription and replication and is thus believed to be the functional form of active *euchromatin*. The overall packing ratio is about seven. These beads on a string then organize themselves in a 30 nm *fiber*, whose exact structure is not clear. Mainly two models are discussed, namely the assembly of a solenoid and a zig-zag structure. It is currently under debate whether the formation of a 30 nm fiber is only possible due to the highly artificial environment in *in vitro* experiments. The packing ratio of the 30 nm fiber is approximately six. Including the compaction by nucleosome formation, we end up with a DNA compaction of around 40 in the 30 nm fiber. DNA in this highly compacted form is less accessible for transcription and is thus associated with inactive chromatin (*heterochromatin*). For transcription,

DNA can be made more accessible by histone tail modifications, which weaken chemical attractions between histone tails and DNA.

In the following, we will be concerned with data from cells fixed in interphase, in which chromosomes are mainly present as a 30 nm fiber. The next level of compactification consists of folds of the 30 nm fiber, such that the final compaction ratio in interphase is in the range of 10^2 – 10^3. In metaphase, chromosomes become even more compacted, although the exact mechanism is not clear.

The 30 nm fiber is positioned non-randomly in the nucleus. While this work is mainly concerned with intra- and interchromosomal contacts, also other organizational elements exist.

Already before the advent of genome-wide 3C techniques, other large-scale features have been known. We refer the reader to a review by Cremer and Cremer [2010], from which we reproduce the most important facts.

Already in 70s experiments showed [Stack et al., 1977; Zorn et al., 1979] that interphase chromosomes do not intermingle, but instead form distinct domains, nowadays called *chromosome territories*, which existence was ultimately confirmed by fluorescence in-situ hybridization (FISH) chromosome paint [Hulspas and Bauman, 1992]. Boyle et al. [2001] showed that gene content is an important factor for the positioning of a chromosome territory with respect to the nuclear periphery; gene-poor chromosomes are usually found situated close to it, while gene-rich chromosomes are located more towards the center of the nucleus. There is more evidence that the three-dimensional structure of the genome is intimately related with gene regulation. Guelen et al. [2008] showed that, in human interphase chromosomes, interactions between the genome and the nuclear lamina occur through lamina-associated domains (LADs) of 0.1 Mbp – 10 Mbp in size, which typically show low gene expression levels.

Another obvious question is whether homologuous chromosomes associate or not. This seems to be organism-specific; in human lymphoblast and fibroblast cells, for example, spatial association between two copies of chromosomes seems to be rather infrequent or, for gene-rich chromosomes, more an effect of being located in the nuclear interior. Furthermore, the distribution of chromo-

somes in the nucleus is not homogenous, but there are chromatin-free regions called interchromatin departments (ICD, Visser et al. [2000]) which appear to have the functional role of accumulating nuclear components involved in transcription.

Taken together, we can draw a picture of a highly organized nucleus, in which the spatial arrangement of genomic domains or chromosomes is an important factor for gene regulation. Investigation of the connection between transcription control mechanisms and three-dimensional architecture of chromatin on a genome-wide scale in a high resolution has recently become possible with methods described in the following.

4.2 Chromosome Conformation Capture techniques

A revolution allowing a more detailed view on nuclear architecture came with a range of experiments, which give information about contact frequencies between different loci and which mainly differ in coverage. We briefly trace these recent developments, but refer to, e.g., de Wit and de Laat [2012] for a thorough review.

The first of these techniques was chromosome conformation capture (3C; Dekker et al. [2002]), in which nuclei are isolated and fixed with formaldehyde, a chemical that cross-links proteins to proteins and DNA. The crosslinked DNA is digested by a restriction enzyme, which cuts DNA at specific, usually 6 bp-long sequences. Examples for such 6-cutters are HindIII used by Dekker et al. or BgIII [Tolhuis et al., 2002]. Then, the ends are ligated under diluted conditions such that intra-molecular ligation events are more likely. Up to this step, the protocol is, with the exception of HiC, identical for 3C and the methods derived from it.

In the 3C experiment, cross-linking is reversed and the ligation product can then be detected by (originally semi-)quantitative PCR using primers specific to both parts of the ligation product. The amplification efficency of different

primers then contains information about ligation frequencies. Using 3C, it is possible to probe a few selected loci against the same set of loci.

In chromosome conformation capture on-chip / circular chromosome conformation capture (4C; Simonis et al. [2006]; Zhao et al. [2006]), 3C is combined with micro-array technology (or, nowadays, next-generation sequencing (NGS)) and thus allows probing the interactions of one locus with many fragments. 4C works by subjecting the 3C ligation products to a second round of digestion with a different restriction enzyme and ligation. The result are circular, chimeric molecules, on which inverse PCR with primers specific to the outer restriction sites is applied. This way, only the sequences specific to one contact partner (the "bait") have to be known.

Coverage was increased even further with the development of carbon-copy chromosome conformation capture (5C; Dostie et al. [2006]), which uses a mixture of designed forward and reverse primers with universal PCR sequences on their ends. The primers which anneal next to each other are then ligated and amplified. NGS or micro-array readout then results in an interaction frequency map for the regions to which the 5C primers were designed. While 3C effectively is a "one-vs-one" and 4C a "one-vs-many" method, 5C can be described as a "many-vs-many" assay.

The final step in increasing coverage to "all-vs-all" was done by Lieberman-Aiden et al. [2009] with the development of HiC, the protocol of which is summarized in Fig. 4.2. In this method, the 3C protocol is slightly modified by filling sticky DNA ends resulting from restriction with biotin-labeled nucleotides. These blunt ends are then ligated and after shearing a biotin pull-down makes sure that only fragments containing a ligation junction undergo further analysis. Amplification and sequencing then yields a list of all detected chimeric sequences. Alignment to the reference genome finally results in a genome-wide interaction frequency matrix.

Several high-throughput variants of the 3C protocol have been proposed since then, among them Tethered Conformation Capture (TCC; Kalhor et al. [2012]), which is essentially a HiC experiment in which ligation is performed on a solid substrate, leading to fewer random ligations and thus increased signal-to-

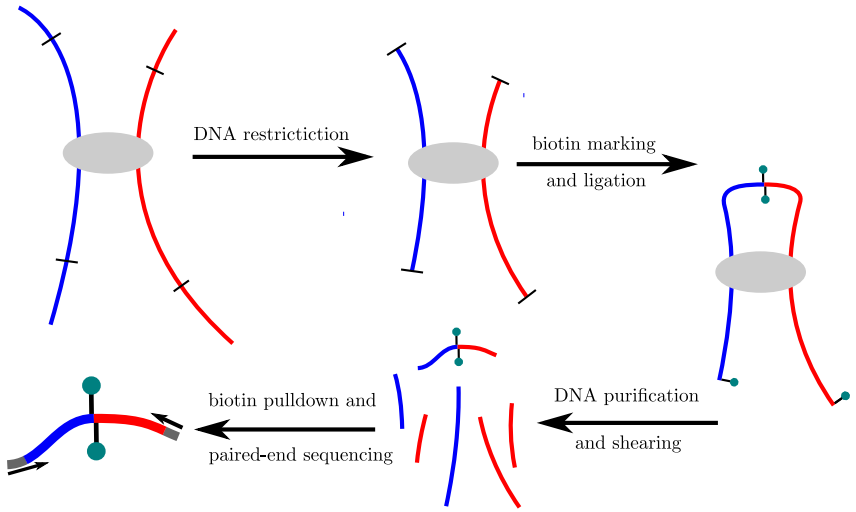

Figure 4.2: Schematic overview of the HiC experiment with experimental details omitted. Crosslinked chromatin is digested with a restriction enzyme, which cuts at specific sites. Overhangs are filled, marked with biotin and ligated. After ligation, crosslinking is reversed, DNA purified and biotin is removed from unligated ends. DNA is then sheared, the biotin-marked fragments are pulled down, subsequently amplified and sequenced.

noise ratio, and methods to map interactions between promoters and enhancers [Hughes et al., 2014; Jäger et al., 2015]; genomic elements, which are required to initiate transcription of a gene. In DNase HiC [Ma et al., 2015], the HiC protocol is heavily modified to use DNase to digest crosslinked chromatin and thus to circumvent a problem common to all previous HiC approaches, namely the dependence on the local restrictiction site distribution. Resolution is thus increased and such is the genome coverage, owing to little DNA loss. Most relevant to this work, though, is the development of Single Cell HiC [Nagano et al., 2013], in which genome-wide chromosome conformation capture is performed in single nuclei. This method results in data which is very sparse but contains information about the spatial organization of the genome in a single cell. The main result is that genome architecture is highly variable from cell to cell.

After sequencing readout, several steps of preprocessing and statistical analysis have to be performed to account for various sources of experimental biases identified by Yaffe and Tanay [2011].

A main source of noise is the ligation step. Several kinds of non-informative read pairs will be present in raw reads and need to be filtered out. Among them is PCR duplication bias, that is, pairs of reads who originate from amplification of the same molecule in the biological sample and thus can be easily filtered out, as they align to identical positions on both ends [Kalhor et al., 2012], although this effect seems to be minimal [Lieberman-Aiden et al., 2009]. Another type of uninformative read pairs stems from circular ligations, which align closely to each other and can be filtered out based on their genomic distance, which would be smaller than the size of the largest restriction fragment. While the majority of cleavage sites corresponds to restriction sites, also unspecific cleavage occurs. Subsequent blunt-end fill up, biotin marking and ligation is unspecific as to whether cleavage occured at a restriction site or not. For this reason, unspecific ligation products contribute to noise in the HiC library. The contacts corresponding to these ligation events can be filtered out by calculating the sum of the distances from the ligation junction to the closest downstream restriction sites and demanding that it should be smaller than a

size selection parameter, because paired-end reads aligning too far away from a restriction site are most likely non-specific ligation events. Yaffe and Tanay showed using a size selection parameter of 500 bp that for the HindIII cutter, 22% of contacts in the Lieberman-Aiden et al. data are spurious ligations.

A further source of bias is the length of restriction fragments. Short and long restriction fragments will have different ligation efficiencies and it is clear that restriction fragments of different length might have different propensities for *trans* (inter-chromosomal) and *cis* (intra-chromosomal) ligations. The results by Yaffe and Tanay show that, indeed, *cis* ligations are enriched and *trans* ligations are depleted for rather short fragment lengths, although the effects are non-linear.

A third source of bias affects sequencing and PCR efficiency. DNA regions with a high GC content are harder to amplify by PCR and underrepresented in Illumina NGS readouts [Aird et al., 2011].

As HiC is a "all-vs-all" method, it requires alignment of the reads to the reference genome in order to determine to which loci reads correspond. For this reason, it is clear that the sequence uniqueness has an effect on detected interaction frequencies, as not uniquely mappable reads do not appear in the contact catalogue and regions with low mappability are thus underrepresented in a raw HiC matrix.

Several methods and software packages have been developed to correct for these biases and to normalize HiC data. In the original HiC publication [Lieberman-Aiden et al., 2009], the interaction frequency matrix is normalized by dividing a given entry by the expected number of reads for the corresponding pair, but more complete normalization procedures have been proposed.

Yaffe and Tanay [2011] use a probabilistic model to calculate prior probabilities for *cis* and *trans* contacts given the above reviewed biasing factors. A maximum likelihood estimate then gives fragment length and GC content correction factors for each combination of fragment ends.

In a method termed *ICE*, Imakaev et al. [2012] improve already on the alignment procedure by accumulating alignments over increasing truncation lengths, which results in significantly more successfully aligned reads compared to a

fixed truncation length. Further downstream in the pipeline, they assume and prove that the (source-unspecific) total bias in detecting a contact between two regions of a binned matrix factorizes in two seperate biases for each region and iteratively determine the MLE for all biases. The ICE-corrected contact maps then allow to directly proceed to an eigenvector analysis without the need for transformation in a correlation matrix as done in [Lieberman-Aiden et al., 2009].

A very recent software named *HiFive* [Sauria et al., 2015] first filters uninformative reads as described above, then iteratively filters out fragments with too low numbers of interaction partners and the corresponding interactions. As one of the few packages, HiFive calculates the signal in the interaction frequency matrix which is due to the distance dependence. For normalization, three different methods are implemented: one relying on matrix balancing, that is, turning a symmetric matrix into a doubly stochastic matrix with row and column sums equal to one, guaranteeing equal "visibility" of each bin; a probabilistic one based on modeling HiC data with a binomial distribution and a third one similar to the method used by Yaffe and Tanay [2011]. HiFive also emphasizes speed and usability.

Hi-Corrector [Li et al., 2015] implements a parallelized version of the ICE method with highly efficient memory usage. In *HiCNorm* [Hu et al., 2012], a Poisson regression model is used to normalize binned contact maps. This and the choice of a parametric over a non-parametric model leads to a fewer number of parameters and significantly reduced computation time compared to the approach of Yaffe and Tanay.

The development of high-throughput, genome-wide chromosome conformation capture resulted in a number of important discoveries, of which we only mention a few particularily striking ones. Lieberman-Aiden et al. [2009] analyzed the genome of human lymphoblastoid cells and found that intrachromosomal HiC heatmaps are partitioned into two types of compartments having a size of several Mbp in an alternating manner. Within each compartment, contacts are enriched, but contacts between the two different compartements are

depleted. This transfers to interchromosomal contacts: given any two chromosomes, labels A and B can be assigned to compartments on each chromosome so that *trans* contacts are enriched between compartments with equal labels and depleted between differently labeled compartments. This suggests a spatial partioning of the whole genome into two types of chromatin, termed A/B compartments. Analysis of genetic and epigenetic features associates one compartment with open, active euchromatin and the other one with closed, inactive heterochromatin. This landmark of large-scale genome organization is conserved across tissues.

By increasing sequencing depth, Dixon et al. [2012]; Nora et al. [2012] discovered topologically associating domains (TADs) in mouse embryonic stem (ES) cells, human stem cells and human IMR-90 cells. The locations of these $\approx$ Mbp-sized domains are not tissue-specific and occur in both compartments. A subset of detected TAD boundaries coincides with boundaries of other domain-like features of chromosomal organization, for example the already discussed A / B compartments and lamina-associated domains (LADs, Guelen et al. [2008]; Peric-Hupkes et al. [2010]). Furthermore, TADs seem to be evolutionarily mostly conserved and their boundary regions correlate with insulators, which block the the interaction between promoters and enhancers, and barrier insulators, which set heterochromatin boundaries. They might thus be related to transcription control.

A TAD can be disrupted by deletion or inversion of its boundaries [Dixon et al., 2012; Nora et al., 2012]. A very interesting, recent result is that disruption of a specific TAD with genes involved in limb development adjacent to one of its borders causes new enhancer-promoter interactions and misexpression and leads to certain limb development disorders in mice [Lupiáñez et al., 2015]. Moreover, 4C performed on fibroblasts of patients suffering from the same types of limb malformation showed the same chromatin reorganization and abnormal interactions.

But fine-scale structure of the human genome can be resolved even further. [Rao et al., 2014] produced contact maps in kb resolution using a modification of the orginal HiC protocol in which the ligation step is performed in intact

nuclei, which reduces random ligation events and allows to use a 4-cutter instead of the commonly used 6-cutters. They found a further level of domain organization, so-called *contact domains* with a size of $\approx 185\,\mathrm{kb}$, which again line the diagonals of the contact matrix. These contact domains segregate in at least six nuclear subcompartments, which all have distinct patterns of histone modifications. Furthermore, in maps with 5 kb resolution, it is possible to to distinguish between ordinary domains and loop domains with the latter being demarcated by off-diagonal peaks indicating enriched contact frequencies relative to their neighbourhood, while not all peaks demarcate contact domains. Contact domains are conserved across human cell lines and evolution (between human and mouse) and often so are peak loci. This hints to conservation of three-dimensional genome structure on a very fine scale across mammals and tissues. Many loops are associated with gene regulation and a major part of peak loci are bound by CTCF and two cohesin subunits. All three proteins bind to a specific CTCF-binding motif, which is found at both loci in contact and is convergently orientated. Furthermore, the analysis of the inactive X chromosome revealed homolog-specific features such as compartementalization of the paternal X chromosome in two super-domains and parent-of-origin specific loops.

4.3 Bayesian structure determination from Single Cell HiC data

We now proceed to the application of ISD on data from the aforementioned single cell HiC experiment [Nagano et al., 2013]. Nagano et al. applied the single cell HiC protocol to ten male mouse T helper cells. We infer structural ensembles and nuisance parameters for the X chromosome of the best-quality data set as determined by Nagano et al.. These contact data represent structural information about one single copy of a molecule and thus we do not have to worry about how to assign contacts to different copies of the molecule. A key property of single cell HiC data is that they are very sparse. In the already nor-

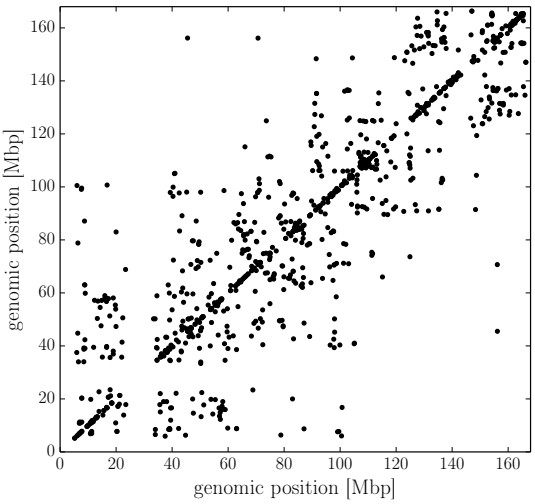

Figure 4.3: X chromosome single cell HiC contact data used in this work.

malized and filtered data set made available to us by Stevens [2013], there are in total 518 *cis* contacts for the X chromosome, which has a length of approximately 166 Mbp. It should be noted that the data deposited online (accessible at NCBI GEO database [Nagano et al., 2013], accession GSE48262) contain a few more contacts. The contact data we use are shown in Fig. 4.3. We note two regions in which no contacts are detected: up to a genomic position of 5 Mbp and between 48–66 Mbp. In the reference genome, these correspond to unknown sequences, to which obviously no reads can be aligned.

We describe every part of the application of the ISD structure determination framework to this system. First, we introduce the coarse-grained structural prior distribution and a simple model we represent a chromosome with. Then, we discuss three different likelihoods and the nuisance parameter prior distributions. We give details on how we draw samples from the resulting posterior distribution and finally discuss the results of our approach and compare it with the conformations calculated by Stevens using the methods described in Nagano et al. [2013].

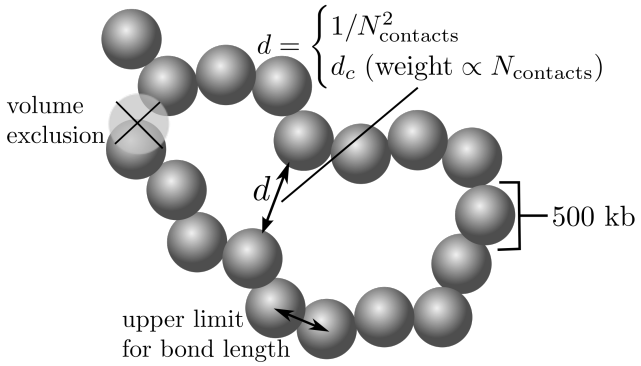

Figure 4.4: Coarse-grained polymer model employed as structural prior information. Bead overlaps are penalized quartically and backbone distance restraint violations quadratically.

4.3.1 Structural prior information: a beads-on-a-string model as a coarse-grained representation of the chromatin fiber

We represent a chromosome by a beads-on-a-string model made of N beads of equal diameter d_{VE}. Each bead corresponds to a bin of 500 kb in the binned experimental HiC matrix. Rosa and Everaers [2008] give a nuclear density of 12 Mbp/µm^3 for interphase chromosomes. We can thus calculate the diameter of one bead to $d_{\text{VE}} = 430\,\text{nm}$. Calculations were run in arbitrary modeling units and for the sake of clarity we come back to the biological meaningful distance unit only for analysis. The distance $d_{i,i+1}$ between two adjacent beads i, $i+1$ is restrained to an upper limit of d_0 with deviations penalized quadratically with a force constant k_{bb}. d_0 and d_{VE} are thus parameters determining the length scale in model coordinates. Overlaps between two beads (excluding sequential neibghbours) are penalized quartically with a (unit-less) force constant k_{ve}. We chose the force constants according to Nagano et al. [2013] and summarize them, along with likelihood parameters, in Table 4.1. When considering distance restraints derived from the contact data, we set $d_{\text{VE}} = d_0$ and regard d_0 as a prior hyperparameter, which needs to be estimated from

	lower-upper error model	log-normal error model	contact-based likelihood	Nagano et al. [2013]
basepairs / bead	500 kb	500 kb	500 kb	500 kb
backbone force constant k_{bb}	25.0	25.0	25.0	25.0
volume exclusion force constant k_{ve}	1.0	1.0	1.0	1.0
bead diameter	d_0 (inferred[1])	d_0 (inferred[1])	$d_{\mathrm{c}} = 430\,\mathrm{nm}$	const.[2]
backbone bead distance	d_0 (inferred[1])	d_0 (inferred[1])	d_0 (inferred[1])	const.[2]
target / contact distance	n_{ij}^{-2}	n_{ij}^{-2}	$1.5 d_{\mathrm{c}}$	n/a
distance restraint force constant k_{dr}	inferred	inferred	n / a	25.0

Table 4.1: Summary of the modeling parameters. [1]: for analysis rescaled to 430 nm. [2]: in accordance with Nagano et al., we rescale their models at 500 kb resolution to the average size of their models calculated at 50 kb resolution, which is 4.3 µm.

the data. Our conformational prior distribution is then given by

$$p_{\text{distance}}(\mathbf{x}|d_0) = \frac{1}{Z(d_0)} \exp\left\{-d_0^{-4} k_{\text{VE}} \sum_{i\neq j}^{N} \theta[d_0 - d_{ij}(\mathbf{x})][d_0 - d_{ij}(\mathbf{x})]^4\right\} \times \exp\left\{-\frac{k_{bb}}{2} d_0^{-2} \sum_{i=0}^{N-1} \theta[d_{i,i+1}(\mathbf{x}) - d_0][d_{i,i+1}(\mathbf{x}) - d_0]^2\right\} . \tag{4.1}$$

During ISD simulations from a posterior incorporating this prior, the bead diameter and thus the length of the molecule change constantly. For analysis, we recover physical units by dividing distances within a structure by its d_0 value.

When employing the contact-based likelihood, we keep the bead diameter fixed to a value d_{VE} and vary only the distance d_0 between adjacent beads in the polymer chain;

$$p_{\text{contact}}(\mathbf{x}|d_0) = \frac{1}{\tilde{Z}(d_0)} \exp\left\{-\frac{k_{\text{VE}}}{2d_{\text{VE}}^2} \sum_{i\neq j}^{N} \theta[d_{\text{VE}} - d_{ij}(\mathbf{x})][d_{\text{VE}} - d_{ij}(\mathbf{x})]^4\right\} \times \exp\left\{-\frac{k_{bb}}{2} d_0^{-2} \sum_{i=0}^{N-1} \theta\left[d_{i,i+1}(\mathbf{x}) - d_0\right]\left[d_{i,i+1}(\mathbf{x}) - d_0\right]^2\right\} . \tag{4.2}$$

The reason for this is the fact that estimating both the bead size and linear bead distance allows the contact restraints to be always fulfilled. This can be seen by regarding the bead diameter and the linear bead distance as parameters determining the size of a structure. For sufficiently small bead diameters and linear bead distances, beads will always be closer than the contact distance d_c.

Fig. 4.4 summarizes the coarse-grained representation of the chromatin polymer and the two different types of restraints used in this work.

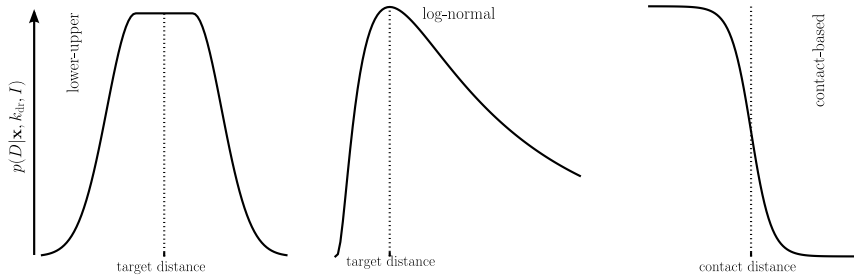

Figure 4.5: Likelihoods implemented for chromosome structure determination from HiC data.

4.3.2 Likelihoods: distance restraint- and contact-based data back-calculation

We implemented three different likelihoods to measure the compatibility of a structure with the data: a distance-restraint based likelihoods similar to the data energy used in [Nagano et al., 2013] in order to compare the results of the ISD approach to the structural ensemble obtained by Nagano et al. via pseudo-energy minimization, a variation of this likelihood with a log-normal error model, and a less ad-hoc, contact-based likelihood, in which models with inter-bead distances lower than a global contact distance are preferred, but weighted differently according to the contact count, similar in spirit to the contact-based data energy term used in [Trieu and Cheng, 2014]. In all cases we do not take into account sequential contacts, that is, contacts between neighbouring beads, as their distances are already restrained by the polymer model.

We split up the first two likelihoods in a forward and an error model as described in Sec. 1.1. As done by Nagano et al., we convert the total number of contacts n_{ij} between two bins i, j to a target distance $d_{ij} \propto {}^1\!/_{n_{ij}^2}$. The forward model then simply consists in back-calculating distances between restrained beads from the structure or, in mathematical terms, $f_{ij}(\mathbf{x}) = |\mathbf{x}_i - \mathbf{x}_j|$. We first employ a flat-bottom error model, which allows the back-calculated distances to deviate from the target distances in a given range depending on them and, if this range is exceeded, penalizes deviations quadratically with the same

force constant for all distances. While this is easy to understand as a scoring function, it is instructive to work out the interpretation of this combination of forward and error model as a likelihood for the experimental distances (calculated from contacts).

Focusing on a single distance between two beads i, j, the forward model back-calculates a theoretical, noise-free distance $\hat{d}_{ij}(\mathbf{x})$. In the error model we then assume that the experimental distance d_{ij} scatters in an interval $L_{ij} = [\hat{d}_{ij}(\mathbf{x}) - {}^{\Delta_{ij}}/_2, \hat{d}_{ij}(\mathbf{x}) + {}^{\Delta_{ij}}/_2] =: [L_{\min}, L_{\max}]$ of width $\Delta_{ij} = L_{\max} - L_{\min}$ centered on the theoretical distance and thus, in this interval, has uniform, maximum probability. As soon as the experimental distance exceeds the limits of L_{ij}, its probability is given by the left / right flank of a Gaussian distribution with variance $\sigma^2 =: {}^1/_{k_{\mathrm{dr}}}$ with its mean on the interval limits. The likelihood of the experimental distance d_{ij} then is

$$L(d_{ij}|\mathbf{x}, k_{\mathrm{dr}}) = \frac{1}{\sqrt{\frac{2\pi}{k_{\mathrm{dr}}}} + \Delta_{ij}} \exp\left\{-\frac{k_{\mathrm{dr}}}{2}\left[E_{\mathrm{lower}} + E_{\mathrm{upper}} + E_{\mathrm{interval}}\right]\right\}$$

with the single contributions given by

$$\begin{aligned} E_{\mathrm{lower}}(\mathbf{x}, d_{ij}) &= \theta\left[L_{\min} d_{ij} - \hat{d}_{ij}(\mathbf{x})\right]\left[\hat{d}_{ij}(\mathbf{x}) - L_{\min} d_{ij}\right]^2 , \\ E_{\mathrm{upper}}(\mathbf{x}, d_{ij}) &= \theta\left[\hat{d}_{ij}(\mathbf{x}) - L_{\max} d_{ij}\right]\left[\hat{d}_{ij}(\mathbf{x}) - L_{\max} d_{ij}\right]^2 , \\ E_{\mathrm{interval}}(\mathbf{x}, d_{ij}) &= 0 \ . \end{aligned}$$

To generalize to several distances, we assume equal force constants and independent measurements. We can then multiply all the single-distance likelihoods and arrive at

$$L(D|\mathbf{x}, k_{\mathrm{dr}}) = \frac{1}{Z(k_{\mathrm{dr}})} \exp\left[-\frac{k_{\mathrm{dr}}}{2}\chi^2(\mathbf{x})\right] \tag{4.3}$$

with

$$Z(k_{\mathrm{dr}}) = \prod_{(i,j)\in D}\left(\sqrt{\frac{2\pi}{k_{\mathrm{dr}}}} + \Delta\right) = \left(\sqrt{\frac{2\pi}{k_{\mathrm{dr}}}} + \Delta\right)^{n_{\mathrm{data}}}$$

and

$$\chi^2(\mathbf{x}) = \sum_{(i,j)\in D}\left[E_{\mathrm{upper}}(\mathbf{x}, d_{ij}) + E_{\mathrm{lower}}(\mathbf{x}, d_{ij})\right] \ .$$

In these equations, $(i,j) \in D$ denotes a pair of restrained beads i and j. The total number of distance restraints is n_{data}. We choose, in accordance with Nagano et al., $\Delta = {}^{2}\!/_{5}$. While the interval limits depend on the theoretical distance $\hat{d}_{ij}(\mathbf{x})$, the width Δ does not. This means that the normalization constant is independent of the structure x and thus does not need to be accounted for when simulating from the conditional posterior distribution for the structure.

The flat-bottom Gaussian error model for distances has the flaw of assigning finite probabilities to negative distances. This is unwanted and we thus implement a log-normal distribution as a suitable replacement. It is a probability distribution for a random variable whose logarithm is normally distributed and thus has strictly positive support. Distances larger than the target distance are less strongly penalized than shorter distances. The log-normal likelihood then takes the form

$$L(D|,k_{\text{dr}}) = \prod_{(i,j)\in D} \frac{1}{\sqrt{\frac{2\pi}{k_{\text{dr}}}} d_{ij}} \exp\left[-\frac{k_{\text{dr}}}{2} \log^2\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{d_{ij}}\right)\right] . \tag{4.4}$$

The log-normal distribution was proposed by Rieping et al. [2005b] to model errors in nuclear Overhauser effect (NOE) data in the context of protein structure determination and shown to lead to structures of higher quality.
In estimating the length scale of the polymer model relative to the target distances, we effectively rescale the experimental distances, as in physical units, the bead radius and thus polymer length is fixed. This scaling of experimental distances can also be done directly in the likelihood: we can take the opposite point of view and, keeping d_0 at a fixed value, regard the experimental crosslinking distances d_{ij} as unknowns which we can estimate from the data. In this case, the data would just be the information between which regions a crosslink has been measured.

In implementing these likelihoods we are mainly aiming for comparability with the structure calculation by Nagano et al. [2013], but nevertheless consider both of them as problematic for a number of reasons: first, the distance

between two beads decreases quadratically with the number of measured contacts, which is an unjustified assumption on the nature of the biological system. In polymer physics, different models exist, predicting different dependences of contact probabilities $P_c(d)$ on distances d. Lieberman-Aiden et al. [2009] use HiC to determine a fractal globule [Mirny, 2011] as the best fitting polymer model for interphase chromatin. For a fractale globule, simulations by Mirny showed $P_c(d) \propto d^{-1}$, which would fit Nagano et al.'s prescription for $n_{ij} = 1$, but does not neccessarily need to interpolate to a general $d_{ij} \propto n_{ij}^{-2}$ scaling. These results from polymer physics hold for homopolymers, in which each bead has identical properties. While one could model a 30 nm fiber *in silico* as such a free homopolymer, due to non-random, functional interactions, this approximation has to be questioned *in vivo*. Second, for more than one measured contact between two beads, the enforced distances are not compatible with the conformational prior distribution (Eq. 4.1), as beads with a target distance $< d_0$ will necessarily overlap.

As an alternative to distance-restraint based modeling, modeling strategies directly based on the contact information have been proposed in the literature [Kalhor et al., 2012; Trieu and Cheng, 2014]. Instead of assuming a dependence of the contact distance d_c on the number of contacts measured between two bins, we fix $d_c = 1.5d_0$. If two beads are closer than d_c, they are considered to be in contact. To take the variable number of measured crosslinking events between bins into account, we weigh each contact with the corresponding number of counts n_{ij} in the binned contact matrix. Because we employ a gradient-based method in our sampling approach (HMC, see Sec. 1.3.3), we need to transform the binary contact restraint into a smooth and differentiable function of the distance. To this end, we define a smoothing function $s(x, \alpha) = [1 + \exp(-\alpha x)]^{-1}$, which depends on parameter α determining how smeared out the contact is. We take α to be equal for all contacts. The likelihood then is

$$\hat{L}(D|x) = \prod_{i,j} s(d_c - |\mathbf{x}_i - \mathbf{x}_j|, \alpha)^{n_{ij}} \tag{4.5}$$

with n_{ij} the experimentally measured contact count between beads i and j. For one single contact, the likelihood can be regarded as a Bernoulli distribution. Let the random variable c_{ij} describe a single contact; $c_{ij} = 1$ if the contact is established and $c_{ij} = 0$ if it is not. The probability of obtaining $c_{ij} = 1$ is given by $p = s(d_c - |\mathbf{x}_i - \mathbf{x}_j|\,; \alpha)$, and the probability of obtaining $c_{ij} = 0$ by $1 - p$. Here, we are not interested in the latter, but in general one could use it to model anti-contact restraints. This is not sensible for single cell HiC data, as the absence of a contact in the contact matrix is most likely due to the inefficacy of the experimental procedure to read out higher number of contacts. It would be possible, though, to introduce anti-contact restraints from ensemble HiC data by regarding all interaction frequencies below a certain cut-off as noisy experimental zero-counts.

In theory, we could regard α as a nuisance parameter and estimate it along the structures, but ultimately we are interested in the limit $\alpha \to \infty$ to recover the forward model as a step function. For this reason we set α as a replica parameter similar to a temperature: for small α, the contact restraints are very soft and we are effectively simulating a freely moving polymer chain. For large α, violations of the contact restraints significantly decrease to the posterior probability. The only parameter in this likelihood then is the contact distance d_c. Estimating it would again be difficult, because a large d_c would always lead to perfect agreement with the experimental data. Note that we use a slightly modified prior distribution (Eq. 4.2) with fixed volume exclusion distance, because its estimation would lead to extremely compact structures trivially fulfilling all contact restraints.

4.3.3 Nuisance parameter prior distributions

A posterior distribution over both structures and nuisance paramaters also contains prior distributions for the nuisance parameters. As both the lower-upper and the log-normal error models are of the form given by Eq. 1.4, the force constant k_{dr} is a scaling parameter and we thus choose, according to our

discussion of prior distributions for scale parameters (Sec. 1.2),

$$p(k_{\mathrm{dr}}) \propto \frac{1}{k_{\mathrm{dr}}} \ .$$

For the distance scale d_0 we also choose a Jeffreys prior, that is,

$$p(d_0) \propto \frac{1}{d_0} \ .$$

4.3.4 Sampling from the Single Cell HiC posterior distributions

As discussed in Sec. 1.2, implementing the ISD approach is not trivial, because it is impossible to directly draw samples from the posterior distribution. We rely on Gibbs sampling (Sec. 1.3.2) to decompose sampling from the posterior distribution into consecutive sampling steps from conditional posterior distributions for the structure and the nuisance parameters. Furthermore, we embed Gibbs sampling in a Replica Exchange algorithm (Sec. 1.3.4). In the case of distance-based likelihoods, the schedule is designed such that with increasing replica index, the likelihood is more and more downweighted. For the contact-based likelihood, we change the smoothing parameter from large to very small values. In both cases, effectively, there is only a weak influence of the data in the "high-temperature" replicas and the prior distributions dominate. This prevents the sampling from getting trapped in high-probabilty regions, but requires more computational resources. In our simulations, 31 replicas sample posterior distributions with likelihood L^{λ} with decreasing λ such that sufficient exchange acceptance rates are sufficiently high to ensure good mixing. All sampling parameters are summarized in Table 4.2.

To draw representative samples from the posterior distribution for structures conditioned on the nuisance parameters is very challenging, for reasons discussed in Sec. 1.3.3 and we employ the HMC sampler discussed there. HMC usually takes three parameters: a mass matrix, the MD trajectory length and the timestep for the numerical integration. We set the mass matrix equal to the unity matrix. The timestep is automatically adapted in a preliminary run

	lower-upper error model	log-normal error model	contact-based likelihood
structure $\mathbf{x}$	HMC	HMC	HMC
k_{dr}	RWMH	Gamma distribution	n/a
d_0	RWMH	RWMH	RWMH
Replica schedule (31 replicas)	$\lambda \in [1.0, \ldots, 0.1]$; exponentially decreasing	$\lambda \in [1.0, \ldots, 0.1]$; exponentially decreasing	$\alpha \in [100, \ldots 0.01]$, exponentially decreasing

Table 4.2: Summary of samplers / sampling distributions for all model parameters and the Replica Exchange scheme. RWMH denotes a random walk Metropolis-Hastings scheme (Sec. 1.3) with a uniform proposal distribution and stepsizes adapted in a preliminary run to give acceptance rates of 50 %. HMC denotes Hamilonian Monte Carlo (Sec. 1.3.3) with a MD trajectory length of 100 steps and, as for RWMH, timesteps adapted in a preliminary run.

to keep a constant acceptance rate of 0.5 and we use the resulting values to set the timestep for the production run.

Drawing force constants from the conditional posterior distributions for the force constants is a much easier task. For the log-normal error model, the conditional distribution for the force constant k_{dr} is a Gamma distribution for which sampling routines are readily available in many programming languages, while for the lower-upper error model, we use a simple Metropolis-Hastings MCMC scheme (Sec. 1.3.1) to sample k_{dr}.

Just like for the lower-upper error model force constant, the sampling distributions for the distance scale d_0 is a non-standard distributions and we again resort to a random walk Metropolis-Hastings scheme. As the conditional posterior distributions for the nuisance parameters are one-dimensional, this is sufficient.

A little catch is that the normalization constants for the structural prior distribution depend on d_0/d_0'. Although they are unknown, because of $q(d\mathbf{x}|d) := p(d\mathbf{x}|d) \times Z(d) = q(\mathbf{x}|1)$, we know their dependence on d, that is, $Z(d) = d^{3N}Z(1)$. This argument holds for both $Z(d_0)$ and $\tilde{Z}(d_0')$ and thus allows us to use the Metropolis-Hastings algorithm, as the unknown $Z(1)$ cancels out in the acceptance criterion.

For all random walk Metropolis-Hastings samplers, we again adapt the stepsize in advance and use the fixed, optimized values for the production run.

4.3.5 Structural ensemble and nuisance parameters

While we performed calculations for the six single cell HiC datasets of best quality as indicated by Nagano et al. [2013], we illustrate our structure determination approach on the data set of best quality, which also contains the highest number of contacts.

Any ISD calculation results in not only one, but in an ensemble of structures and so does ours. Fig. 4.6 shows a superimposed subset of ensemble members for each likelihood. As expected, the sparsity of the data and the prior in-

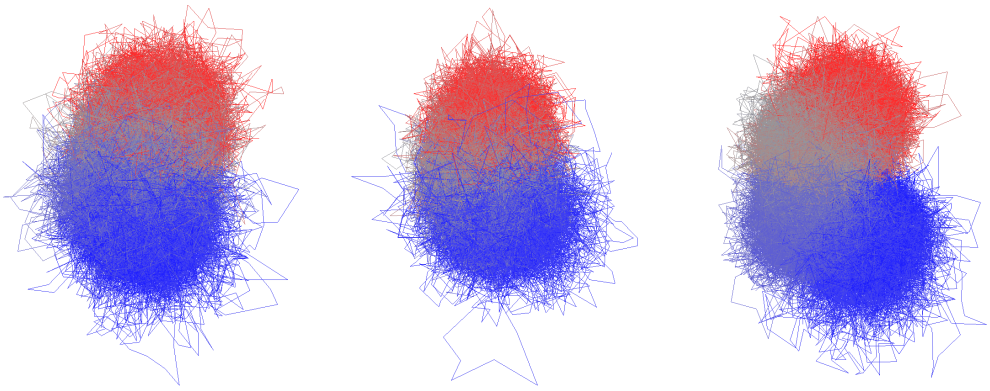

Figure 4.6: Structural ensembles calculated with ISD using the lower-upper (*left*) error model, the log-normal (*center*) error model and the contact-based likelihood (*right*). Shown is every 30th sample after discarding 3000 samples. For clarity, larger regions with no data (0 Mbp – 5 Mbp, 24 Mbp – 33 Mbp) are not shown. Color-coded is the genomic position from red over grey to blue.

formation are reflected in a wide spread of the structural ensembles, although it is interesting that the contact-based likelihood results in a better defined ensemble. Nevertheless, one basic feature is (more or less) visible in all three ensembles: the partition of the X chromosome in several super-domains, visible in Fig. 4.6 as blobs of similar color, which form crescent-shaped models of the chromosome. This partition is already evident in the data (Fig. 4.3) as large blocks on the diagonal. The radii of gyration for conformations sampled from the three different posterior distributions are $1175 \pm 79\,\mu m$ (lower-upper error model), $1084 \pm 64\,\mu m$ (log-normal error model) and $1408 \pm 63\,\mu m$ (contact-based likelihood). Although the radius of gyration is a conservative estimate of the spatial extension of a molecule, this shows that the inferred structures have sizes in the same order as the experimentally measured diameters of $\approx 3.7\,\mu m$ [Nagano et al., 2013].

Clustering of the ISD samples using self-organizing maps (SOMs, Bouvier et al. [2014]; Kohonen [1982]) shows that, for the log-normal and the contact-based likelihood, we have a continuum of structures without too distinct clusters.

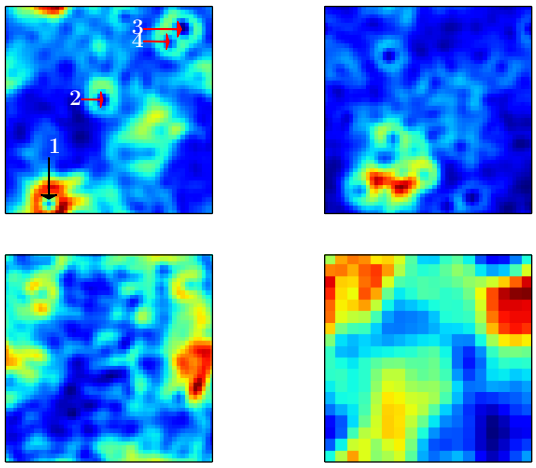

Figure 4.7: Self-organizing maps computed from a subset of the samples obtained during an ISD run. Shown is the U-matrix, whose entries are a measure for distance to neighbouring neurons. Each neuron is represented by one pixel in the matrix.
Top left: ISD; lower-upper error model. Four different clusters are annotated with numbers ranking the respective population. *Top right:* ISD; log-normal error model. *Bottom left:* ISD; contact-based likelihood. *Bottom right:* Structures obtained by Nagano et al.

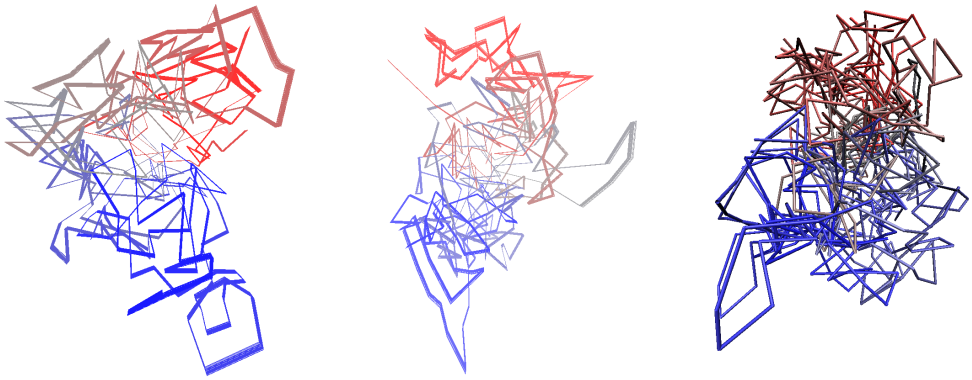

Figure 4.8: Superimposed structures of cluster 1 (*left*), 2 (*center*) and 3 (*right*) of the ISD ensemble calculated with the lower-upper error model.

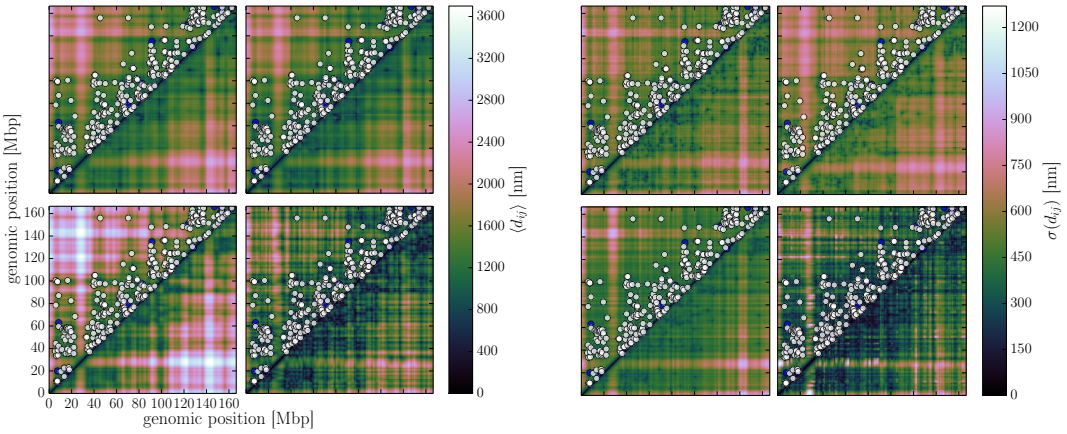

Figure 4.9: Pairwise distance matrices computed from a subset of the samples obtained during an ISD run. Left panel shows the average pairwise distance $\langle d_{ij} \rangle$, right panel the standard deviation. Circles represent entries in the Single Cell HiC contact matrix.
Top left: ISD; lower-upper error model. *Top right:* ISD; log-normal error model. *Bottom left:* ISD; contact-based likelihood. *Bottom right:* Structures obtained by Nagano et al.

The results for the lower-upper error model, on the other hand, are different: we identify several clusters of similar structures denoted, sorted by population, with numbers from 1 – 4 (Fig. 4.7). Cluster 3 contains several copies of a structure and its mirror images, which demonstrates that, as in NMR structure determination, also in our modeling approach mirror images can occur and have to be taken into account during analysis. The very well-aligning structures in cluster 1, on the other hand, resemble the conformations obtained by sampling from the posterior with a contact-based likelihood.

A qualitative validation of our structure determination approach can be performed by analyzing the ensemble-averaged pairwise distance matrix computed from the models obtained during sampling from the ISD posterior distribution. Fig. 4.9 (left) shows the average distances matrices of both structures resulting from both ISD and of structures obtained by Nagano et al. using an energy minimization-based modeling approach. We immediately observe that regions with a high number of contacts in the Single Cell HiC matrix correspond to

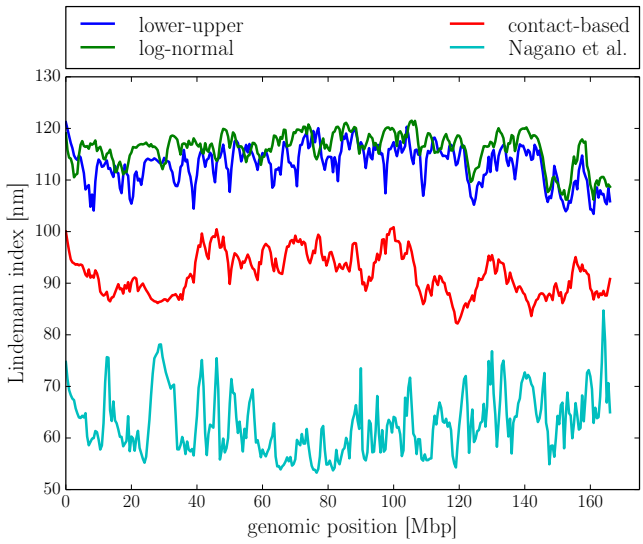

Figure 4.10: Local Lindemann indices of structures obtained by ISD and Nagano et al.

regions of low pairwise bead distances in the distance matrices. The ISD distance matrices and Nagano et al.'s distance matrices share this property, which makes us expect qualitatively similar structures.

Considering the matrices of pairwise distance standard deviations (Fig. 4.9; right), we confirm that both our ISD and the SA based modeling approach indeed restrain beadwise distances supported by corresponding data points in the contact matrix. Both the centromeric region (24 Mbp – 33 Mbp) without any data points and the first few Mbp in which no contacts were measured show significantly higher standard deviations in pairwise distances involving beads representing them. Interestingly, we observe that the pairwise distances are much more variable in the ISD ensembles as compared to the SA ensembles with the exception of the distances between the two regions not supported by data. This leads us to suspect that the minimization procedure employed by Nagano et al., if repeated several times, results in very similar configurations close to a pseudo-energy minimum, which appears to be located in a rather broad energy basin more exhaustively sampled by the MCMC methods we employ.

A more quantitative view on structural variation within a set of structures is

given by the (local) Lindemann index [Zhang et al., 2007]. While it actually is a measure of thermally driven disorder, but we divert it from its intended use to replace root mean square fluctuations, which in our case are not sensible because of the lack of a well-defined reference structure. The local Lindemann index is defined by

$$L_i = \frac{1}{N-1} \sum_{j \neq i} \frac{\sqrt{\langle |\mathbf{x}_i - \mathbf{x}_j|^2 \rangle - \langle |\mathbf{x}_i - \mathbf{x}_j| \rangle^2}}{\langle |\mathbf{x}_i - \mathbf{x}_j| \rangle} \ .$$

Fig. 4.10 shows L_i for all three sets of ISD simulations and the ensemble by Nagano et al.. We find that the ensemble resulting from a minimization procedure has a consistently lower local Lindemann index, which confirms that its structural variability is indeed lower than the ensembles resulting from ISD posterior sampling. Furthermore, sampling from the posterior distribution involving the contact-based likelihood results in lower Lindemann indices compared to the ensembles obtained by approximating the distance-restrained based posteriors. This makes sense, as one can easily imagine bead distances being trapped between the contact and the volume exclusion distance. Interestingly, the Lindemann index does not clearly distinguish between regions with little data and well-restrained regions.

We can verify that the ISD approach is self-consistent by back-calculating data from a structure and applying our method with these "fake" data as input. To this end, we take the MAP estimate of the structures and nuisance parameters from the contact-based ensemble and calculate the corresponding distance matrix. By comparing its entries with the contact distance d_{c}, we obtain a back-calculated, binary contact map. We then run an ISD simulation using the contact-based likelihood on this contact map and expect the resulting structural ensemble to reproduce, on average, the reference structure. Fig. 4.11 confirms this and a more quantitative measure is given by the Mantel test [Mantel, 1967], which gives a correlation of ≈ 0.81.

One of the main advantages of ISD is its capability to estimate nuisance parameters along with the structure. Fig. 4.12 shows the histograms of estimated nuisance parameters for all three likelihoods. The force constant k_{dr} takes sim-

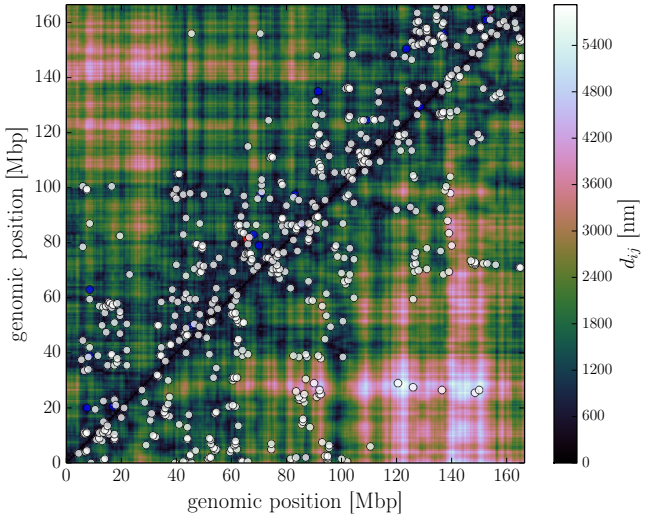

Figure 4.11: Distance matrix of the reference structure (*upper triangle*) and average distance matrix of structural ensemble using data back-calculated from it (*lower triangle*). Experimental single cell HiC / back-calculated contacts are shown as white dots in the lower / upper triangle.

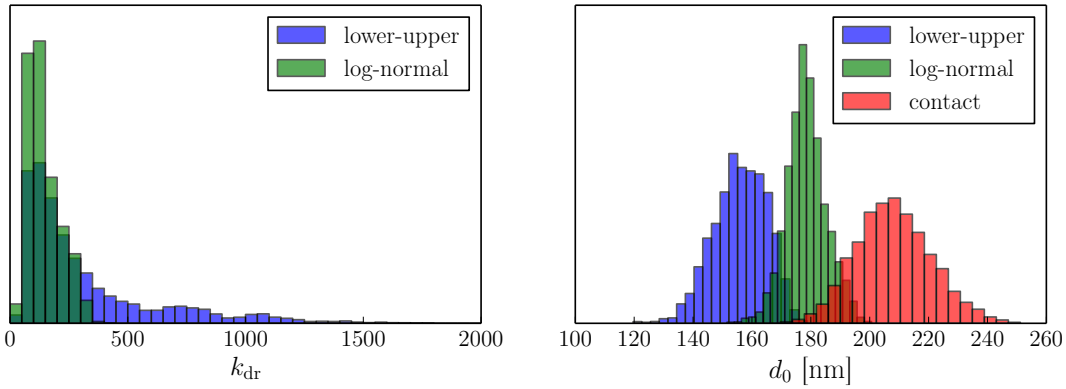

Figure 4.12: Histograms of sampled values for the nuisance parameters k_{dr} (left) and d_0 (middle) for simulation from the ISD posterior employing a lower-upper and a log-normal error model. *Right*: linear bead distances $\overline{d_0}$ obtained by sampling from the ISD posterior employing a contact-based likelihood.

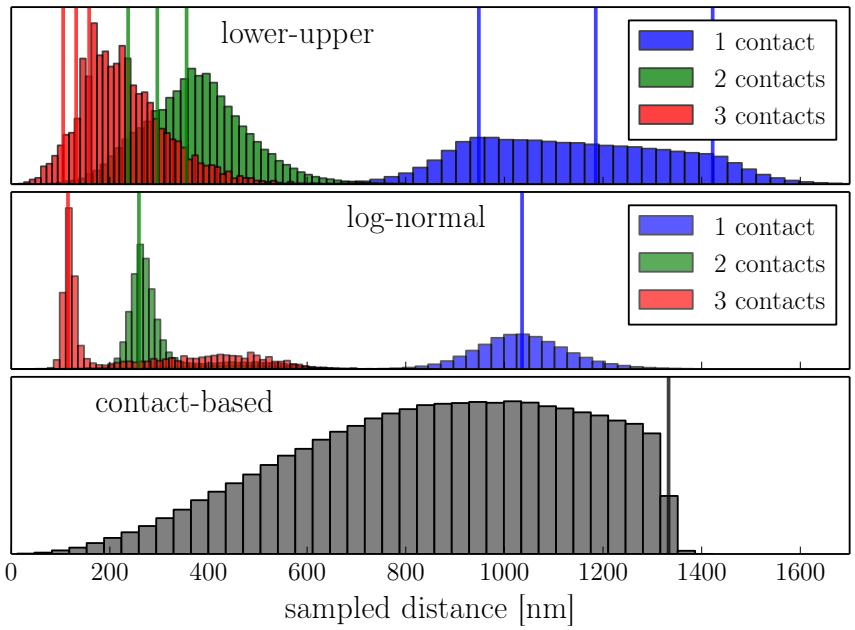

Figure 4.13: Histograms of restrained distances $d_{ij}^{(k)}(x)$ in structures obtained by a simulation from the ISD posterior distribution. Bead distances were restrained with an lower-upper error model to target distances $\hat{\propto} d_{ij}^{(k)} \in \{1/9, 1/4, 1\}$ (top, middle, bottom) with lower / upper bounds equal to $d_{ij}^{(k)} \pm 1/5 d_{ij}^{(k)}$. Red vertical lines indicate target distances, green vertical lines lower and upper bounds.

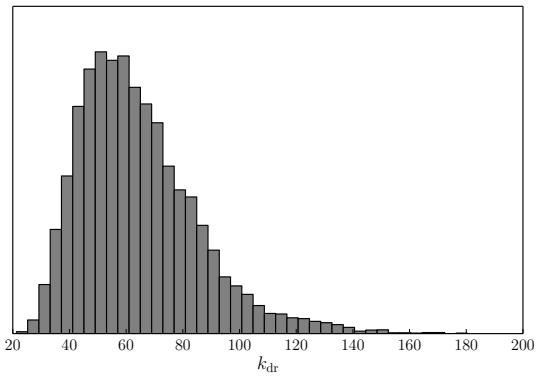

Figure 4.14: Force constants inferred from the cell 1 data set by ISD using the log-normal likelihood, but with the contact counts of two arbitrary entries multiplied by a factor of ten, rendering these data points barely compatible with the physical prior information. Force constants are considerably lower as compared to Fig. 4.12 and ISD thus downweights data inconsistent with the prior.

ilar values for both the log-normal and the lower-upper error model. The force constants are quite large (≈ 120), which indicates that most restraints are well satisfied, in agreement with the histogram of restrained distances (Fig. 4.13). This statement has to be seen in light of the fact that the great majority of bins in the experimental matrix contain only one contact and thus violations of distance restraints corresponding to two or three contacts are comparatively less frequent. Note furthermore that, for the ensemble obtained using the contact-based likelihood, we only show the histogram of distances for contacts with weight $n_{ij} = 1$. For clarity, the other two contact classes are not shown, but show a similar behaviour, that is, the contact restraints are basically always fulfilled.

For comparison with Nagano et al. we note that they used a data force constant equivalent to $k_{\mathrm{dr}} = 25$. Force constants estimated by ISD are thus an order of magnitude higher.

If the data were of bad quality, e.g., by containing contacts either inconsistent of with each other or with the prior information, ISD would mistrust the data and assign a lower force constant. This effect is demonstrated in Fig. 4.14,

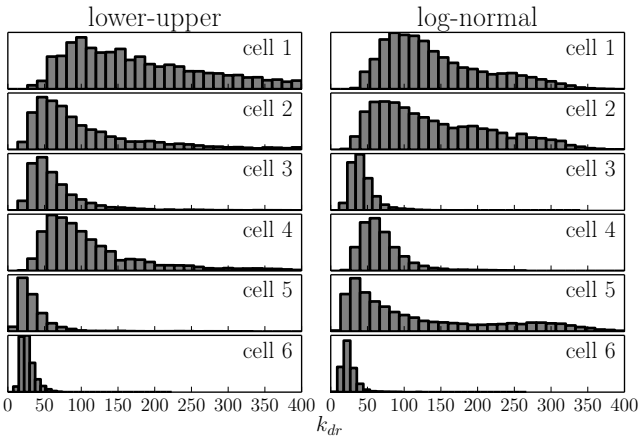

Figure 4.15: Histograms of sampled values for the nuisance parameter k_{dr} for a simulation from the ISD posterior.
Left: lower-upper error model, *right*: log-normal error model.

where we, for the case of a log-normal likelihood, chose two arbitrary contacts and multiplied their count by a factor of ten, thus leading to very small experimental distances, which are incompatible with the prior information of excluded volume.

As Nagano et al. sort the published data sets by quality (1: best, 10: worst), it is interesting to check whether force constants inferred from the different data sets reflect this change. Fig. 4.15 shows that this is indeed the case; we recognize a general trend to lower force constants for both the lower-upper and log-normal likelihood. ISD thus assigns a higher error to data of lower quality. The bead size and thus the polymer length in model coordinates, on the other hand, are different, but comparable in all three likelihoods. Employing the log-normal error model leads to larger beads than the lowerupper error model. This makes sense, because the log-normal error model allows for larger distances between restrained beads and thus, with respect to the volume exclusion, larger beads are still compatible with both the force field and the distance restraints. For the contact-based likelihood, Fig. 4.12 shows somewhat greater sequential bead distances. This comes as no surprise, because we fix the bead diameter and the contact distance to values which are larger than the estimated d_0 values and the target distances in the distance-restraint based likelihood.

4.3.6 Comparing likelihoods via Bayesian model comparison

We may now ask which of the three described posterior distributions describe the data best. Bayesian model comparison (see, e.g., Sivia and Skilling [2006]) can, in principle, give us the answer, as it allows to determine which of two models M_1, M_2 is favored by the data. In this context we mean by "model" our complete description of a modeling approach, encoded in the posterior distribution. We can formulate the posterior probability of model M_i by using Bayes' theorem;

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{P(D)} \ ;$$

and recognize $P(D|M_i)$ as the evidence already discussed in Sec. 1.2. If we do not prefer any model *a priori* by assigning both models the prior probability $P(M_i) = 1/2$, we see that the ration of the model posterior probabilities is given by

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)} =: K \ , \tag{4.6}$$

that is, the ratio of the evidences of both models. K is called the *Bayes factor* and if greater (lesser) than one tells us how much more strongly the data support M_1 (M_2). But calculating K is no easy task because the evidences are the normalization constants of the corresponding posterior distributions and as such usually are high-dimensional integrals over all model parameters. But with histogram reweighting (WHAM, Sec. 1.4) we have a powerful tool at hand to approximate evidences from MCMC samples from Replica Exchange simulations. As we are doing Replica Exchange simulations anyways in order to enhance sampling, WHAM gives us the evidences and thus the Bayes factor without expending a significant amount of additional computing time.

WHAM can only compare models describing identical data. This limits the application of WHAM to the comparison of the lower-upper and the log-normal likelihood, as both use the same data, namely, the distances calculated from contact counts. In the contact-based likelihood, on the other hand, we work

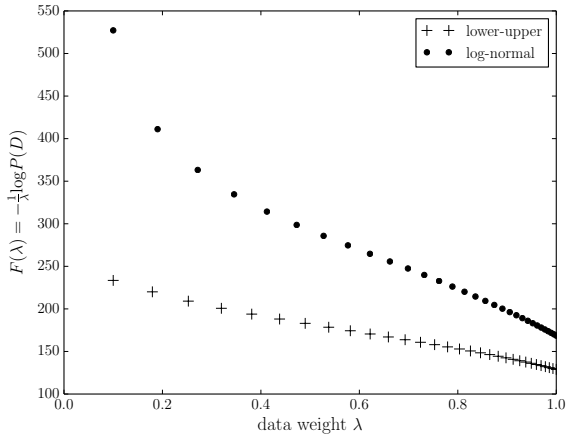

Figure 4.16: "Free energies" for all likelihood weights λ. The very right data points ($\lambda = 1$) correspond to the full posterior distributions of lower-upper and log-normal likelihood and show that, in the framework of Bayesian model comparison, the distance data strongly favor the lower-upper error model.

directly with the contact counts and thus, in the framework of Bayesian model comparison, cannot compare it to the distance-restraint based likelihoods. Applying WHAM to the samples calculated from all replicas, we find that the distance data strongly favor the lower-upper error model (Fig. 4.16).

4.3.7 Inferring the structure of a diploid chromosome

In NMR structure calculation, crosspeaks often cannot be assigned unambigously. Such a peak could stem from the interaction from a proton A with another proton B or, equally likely, from an interaction between A and a third proton C. If one then constructs a distance restraint by averaging arithmetically over the two possible target distances;

$$\overline{d} = \frac{1}{2}(d_{\mathrm{AB}} + d_{\mathrm{AC}}) \ ;$$

only a structure in which both distances $d_{\mathrm{AB}}, d_{\mathrm{AC}}$ fulfill the restraint will give a favorable likelihood contribution. If one distance is close to the target distance, but the other one is very large, the average distance $\overline{d}$ would erroneously

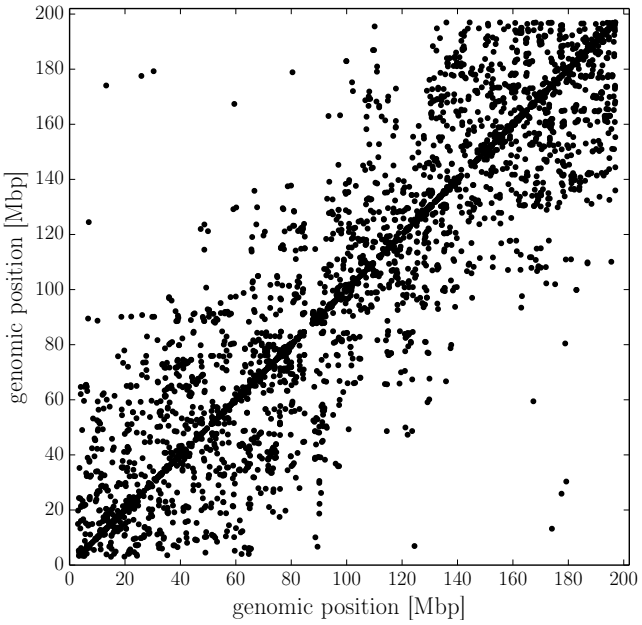

Figure 4.17: Two-copy mouse chromosome 1 contact map containing 1426 contacts.

violate the distance restraint. Nilges [1995] solved this problem by replacing the arithmetic average by an average in which large one distance does not contribute significantly. More specifically, they introduce an r^{-6}-average distance by

$$\overline{d} = \left(d_{\mathrm{AB}}^{-6} + d_{\mathrm{AC}}^{-6}\right)^{-\frac{1}{6}} \quad . \tag{4.7}$$

Although this is not done by Nilges, we divide this average distance by a factor of $2^{-1/6}$ to recover the correct distance for the case of $d_{\mathrm{AB}} = d_{\mathrm{AC}}$. By virtue of the strong decay of the -6 power, a large distance does not contribute and the average will be closer to the small distance. If, on the other hand, both distance are large, the r^{-6} average will also be large.

This average can be diverted from its intended use to infer the structure of a two-copy chromosome from single cell HiC data. In a contact map (Fig. 4.17) of a two-copy chromosome, say, chromosome 1 of the mouse genome, contacts may have been formed either in one of the copies or in both of them. Using the contact-based likelihood, we simulate two copies at the same time and ask for the contact restraints to be fulfilled in either one or both structures. This

amounts to demanding that the r^{-6}-averaged distance $\overline{\hat{d}}_{ij}$ between two beads i, j in the two structures be smaller than the contact distance d_c. In a sense, we now try to assign contacts in a logical *or*-operation instead of an exclusive *or*: the likelihood for a given contact will be close to one if the contact restraint is fulfilled in one or both structures. In the single-structure calculation we asked the two restrained beads to be close in exactly one structure, which can be incompatible with other restraints fulfilled in the second conformation.

We combine this two-copy likelihood with the structural prior given in Sec. 4.3.1 for each copy. Here, we assume that the two copies do not interact, but, at the time of crosslinking, occupied distinct territories [Babu et al., 2008; Khalil et al., 2007]. Under this assumption, the structural prior factorizes in separate contributions for each structure. In the following calculations, we kept the distance scale d_0 fixed.

Sticking with the first mouse chromosome, use now apply ISD using this posterior distribution on the single cell HiC contact map shown in Fig. 4.17, which we obtained from the NCBI GEO database (accession GSE48262). The likelihood based on the r^{-6} average is indeed able to assign contacts to either one of the structures or both (Fig. 4.18): while some contacts are shared between the two copys, a large part of the contacts is uniquely fulfilled either in the model for copy one or copy two. Fig. 4.18 also shows superimposed structures for each of the two copies. While structural variability is considerably high, two features can be discerned. Chromosome 1, too, seems to be organized in large super-domains. Furthermore, one copy shows a more compact structure.

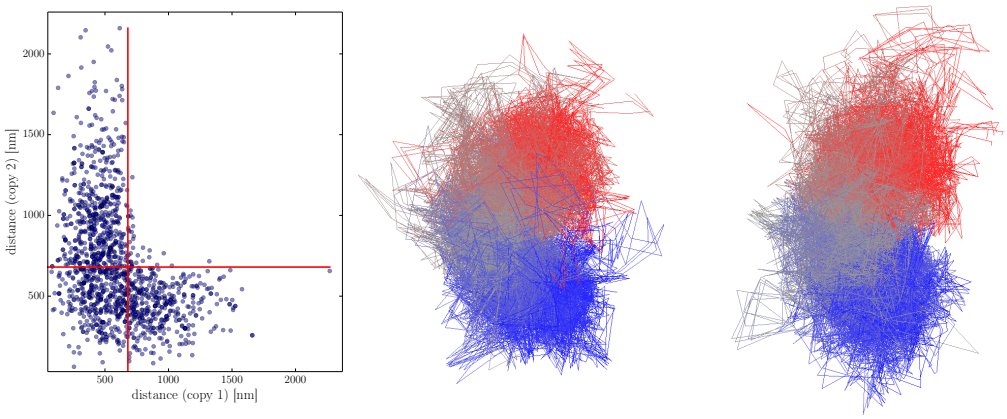

Figure 4.18: Modeling a two-copy chromosome using ambiguous distance restraints.

Left: restrained distances in copy one vs. the same distances in copy two. Red lines mark the contact distance. Some contacts are shared between the two copies (bottom left), while others are fulfilled in only one copy.

Middle / right: 50 superimposed samples of the model for copy one and two.

4.4 Modeling chromosomes from population HiC data

Up to this point, we focused on inference of structures from Single Cell HiC data. While interesting on their own because they allows to measure cell-to-cell variability and give insight into the structure of single molecules, they are very sparse and only allow for very coarse-grained modeling. Furthermore, to our knowledge, no other Single Cell HiC data sets exist and interest in Single Cell HiC experiments seems limited. Population HiC, on the other hand, owing to the wealth of fine-scale information it offers and most certainly the fact that it is less recent than its Single Cell HiC sibling, has become almost a routine tool to investigate nuclear architecture.

4.4.1 Existing approaches to structure determination from population contact data

Many methods have been proposed to translate a contact matrix obtained by a 5C or HiC experiment into three-dimensional structures. Several deterministic methods rely on minimization of a scoring function. Duan et al. [2010] convert interaction frequencies of a 4C variant to distances and use the minimization approach determine a consensus structure. In a similar fashion, work by Baù et al. [2011] on 5C data relates target distances between beads to inverse $\log_{10}$ Z-scores and uses the Integrated Modeling Platform (IMP, Russel et al. [2012]) to optimize a scoring function. A slightly different approach is taken by Varoquaux et al. [2014], who assume that the contact counts follow a Poisson distribution and relate the Poisson rate to the spatial distance between beads. They optimize a parameter of this relation along with the structures. Zhang et al. [2013] use semi-definite programming and also include the conversion factor for constructing distance restraints from interaction frequencies in the optimization procedure. Furthermore, they formulate a measure assessing whether the input distance matrix can be fulfilled by a single 3D structure. Distance geometry is used for genome structure reconstruction by Lesne et al. [2014] after calculating missing distances using a shortest-path algorithm on a graph whose nodes are the loci between ligation events where measured and whose edges are the inverse contact frequencies. Acknowledging the difficulties in converting interaction frequencies to distances, Trieu and Cheng [2014] propose a contact-based scoring function in which all measured contacts are assigned the same target distance, but weights according to the contact count. By minimizing this scoring function, they determine consensus structures for chromosomes of both healthy and cancerous human cells.

All these methods try to determine a single structure fitting both the data and, in most cases, some basic assumptions from the polymer physics of chromatin. But just as for Single Cell HiC structure reconstruction, these approaches fail to give a statistically well-defined measure of confidence in the optimization result. For this reason, several probabilistic methods have been proposed.

Rousseau et al. [2011] again convert contact frequencies to distances and use MCMC methods (among them Hamiltonian Monte Carlo discussed in Sec. 1.3.3) to sample from the posterior distribution for the unknown structure using flat priors. While they acknowledge the possibility that different interaction frequencies might have been measured with different noise levels, they do not sample the errors, but set them to interaction-frequency specific values. In Hu et al. [2013], an attempt is made to more realistically model the relationship between distance similar to [Varoquaux et al., 2014], but again sampling from the structural posterior distribution using MCMC methods. In a more general version of their method termed BACH-MIX, they assume a mixture of models and are thus able to assess how justified calculation of a consensus structure using their one-component method BACH is.

In one way or another, these methods focus on determining consensus structures from population HiC data, but it is highly unlikely that one physically realistic structure is representative of a population of million of molecules. To accomodate this, different population-based modeling approaches have been developed. Using IMP, Kalhor et al. [2012] model a human genome based on data from a modified HiC protocol with higher signal-to-noise ratio using a population of structures, in which contact restraints are fulfilled only in the fraction of structures corresponding to the interaction frequency. An ensemble of structures without violations is determined by minimization of an ensemble target function including these restraints and a basic polymer model. A Maximum Entropy approach is employed in [Zhang and Wolynes, 2015] to determine an optimized energy landscape for human chromosomes and, by simulation of Langevin dynamics, obtain structural ensembles. Furthermore, the topology of chromosomes is investigated with the result that their models are largely free of knots thanks to TADs which locally increase chain rigidity. Finally, Wang et al. [2015] construct chromosome ensembles by converting contact frequencies into distance restraints and, via expectation-maximization (see, e.g., Dempster et al. [1977]), obtain MAP estimates of both a representative structural ensemble and errors as well as conversion factors, which are treated as nuisance parameters. Their ensemble likelihood is a mixture of single-structure

likelihoods.

We now outline a generalization of the ISD principle to infer ensembles of structures from population HiC data. While being preliminary work, we think our idea has the potential to improve on the previous discussed population-based modeling approaches by taking advantage of exhaustive MCMC sampling of the conformational space and a unified treatment of structures and nuisance parameters.

4.4.2 Extension of ISD to model structural ensembles from population HiC data

In most applications, ISD was used to infer a consensus structure from various sources of averaged data. The heterogenity of the underlying ensemble of molecules was not explicitly captured. While this is a reasonable approximation for well-folded proteins, for systems exhibiting greater conformational flexibility, the ISD approach requires an extension. Olsson et al. [2013] propose a joint posterior distribution $p(\mathbf{x}, \mathbf{f}, \mathbf{e}|\mathbf{d})$ for the atomistic model of a structure $\mathbf{x}$, back-calculated data $\mathbf{f}$ and the average $\mathbf{e}$ of the simulated data conditioned on the noisy, averaged data $\mathbf{d}$. ISD is recovered from this framework if $\mathbf{f} = \mathbf{e}$ is assumed. In a previously discussed approach, Wang et al. [2015] calculate MAP estimates for structural ensembles from HiC data by using a likelihood based on a mixture of single-structure likelihoods.

We propose a different approach. If in ISD we infer an ensemble from data representing one molecule or a population of molecules with very similar conformations, then from data from a heterogenous population we can try to infer a population of ensembles. Each population will be a possible approximation of the real structural population and, as does each structure obtained in conventional ISD, will have its own probability weight. Nuisance parameters are treated in exactly the same manner as before, but we may introduce possible weights of population members as new nuisance parameters. This approach is limited to a fixed number of population members, which is not known *a priori*.

Assuming that molecules in the experimental population do not interact, the structural prior distribution of a candidate population factorizes in the multiplication of (identical) prior distributions for each population member. The forward model F, on the other hand, has to be designed specifically for the application in question. In general, it will back-calculate averaged data $\hat{D}$ from a population of structures $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^N)$. We can introduce weights w^k in the forward model, such that, in the case of an arithmetic average,

$$\hat{D} = F(\mathbf{X}; \mathbf{w}, \alpha) = \sum_{k=1}^{N} w^k f(\mathbf{x}^k; \alpha) \ , \tag{4.8}$$

where f denotes the single-structure forward model and $\mathbf{w}$ the vector of weights w^k. α are nuisance parameters on which the forward model might additionally depend. Introducing weights is useful for two reasons. First, we are unaware of the number of distinct structures required to reproduce the averaged experimental data and thus could expect that if a few ensemble members already can reproduce the structure correctly, the remaining ensemble members will be assigned a low weight. Also, we do not know the relative populations of conformations represented by structures in our discrete ensemble, but weights contain this information. In this forward model we could, for example, calculate average distances $\overline{d}_{ij} = F(\mathbf{X})$ between atoms i, j from distances in single structures by the single-structure forward model $d_{ij}^k = f\left(\mathbf{x}^k\right) = |\mathbf{x}_i^k - \mathbf{x}_j^k|$. As before, the forward model F is combined with an error model g to yield a likelihood for the population-averaged experimental data D:

$$L(D|\mathbf{X}, \sigma, \mathbf{w}, \alpha) = g\left[f(\mathbf{X}; \mathbf{w}, \alpha); \sigma\right] \tag{4.9}$$

If we neglect interaction between copies of the molecule, the structural prior distribution just factorizes into a product of prior distributions for the single ensemble members, that is,

$$p(\mathbf{X}) = \prod_{k=1}^{N} p(\mathbf{x}^k) \ . \tag{4.10}$$

Construction of the joint posterior distribution for the ensemble $\mathbf{X}$ and weights and other nuisance parameters then proceeds as described in Sec. 1.2.

Applying this idea to population-based HiC data requires the back-calculation of an interaction frequency matrix $\hat{F}_{ij}$ from an ensemble of structures. This is straightforward; we can just calculate binary contact matrices $\hat{F}_{ij}^k$ for each structure and sum them up. Again we have to replace the step-function θ, by which a contact is defined, by a smoothing function s in order to use HMC (Sec. 1.3.3) to sample from the posterior distribution. We thus have

$$\hat{F}_{ij} = \sum_{k=1}^{N} w^k \hat{F}_{ij}^k = \sum_{k=1}^{N} w^k s(d_c - |\mathbf{x}_i^k - \mathbf{x}_j^k|; \alpha) \ . \tag{4.11}$$

4.4.3 Technical aspects of sampling from the population HiC posterior distribution

In sampling from the single-cell HiC posteriors discussed in Sec. 4.3, low dimensionality and little data played to our advantage and sufficient sampling was comparatively easily achieved. Population HiC data, on the other hand, are more challenging. Dimensionality depends linearily on the number of ensemble members, but quadratically on the number of beads. The reason for the latter is the back-calculation of the single-structure contact matrices f_{ij}^k: as population HiC matrices, due to the large number of molecules in a sample, have basically no entries equal to zero, N^2 distances and (expensive) smoothing functions have to be evaluated for each likelihood gradient evaluation. Several measures can be taken to reduce the number of distance evaluations and thus decrease the computational burden. Quite obviously, we can decrease the number of beads in our models at the cost of lowering model resolution. Second, under the assumption that entries in the experimental interaction frequency matrix with a very low value represent noise, we can introduce a cut-off and thus effectively reduce the number of data points right in the beginning. A third possibility to reduce the number of distance calculations lies in the details of HMC: in calculation of the short MD trajectory, whose final state serves as a proposal, we do not neccessarily need to use the gradient of the negative log-posterior; bias introduced by a different gradient will be corrected in the

Metropolis criterion. This allows us to not use the full interaction frequency matrix, but only a subset of all contacts, thus reducing computation time. It is also very intuitive: most entries in the interaction frequency matrix are usually highly correlated due to the fact that sequential beads are restrained to be close to each other and thus do not neccessarily contain much information. By randomly sampling the data points which play in the gradient evaluation before calculating the short MD trajectory in HMC, we can make sure not to systematically neglect datapoints.

Due to the great dimensionality of the problem, the use of RE (Sec. 1.3.4) or another multicanonical algorithm is mandatory to explore the conformational space. Difficult sampling problems require optimized RE schedules, a problem we adressed in Sec. 3. This is especially the case if the system exhibits a phase transition at a certain value for the replica parameter(s). As the adaptive Replica Exchange method outlined before is not yet functional for general probability distributions, we heuristically determine a suitable schedule which ensures good acceptance rates.

A non-trivial nuisance parameter are the weights w^k. For a Gaussian error model, the conditional posterior distribution of w^k is a product of normal distributions, whose means depends on all other weights. The weights are thus highly correlated random variables and, depending on the number of structures chosen, will be hard to sample from using a simple random walk Metropolis-Hastings scheme. For this reason we use HMC to not only sample the structural ensemble, but also the weights. This is efficient, as expensive parts of the gradient do not depend on the weights and thus have to be computed only once per HMC sample.

4.4.4 Inferring chromosome ensembles from artificial data

We first apply the outlined method to artificial data obtained from single cell HiC structural ensembles. For both the cell 1 and cell 4 data sets, we calculate an ensemble of X chromosome models as described in Sec. 4.3. We then back-calculate binary contact matrices from all models and sum them element-wise

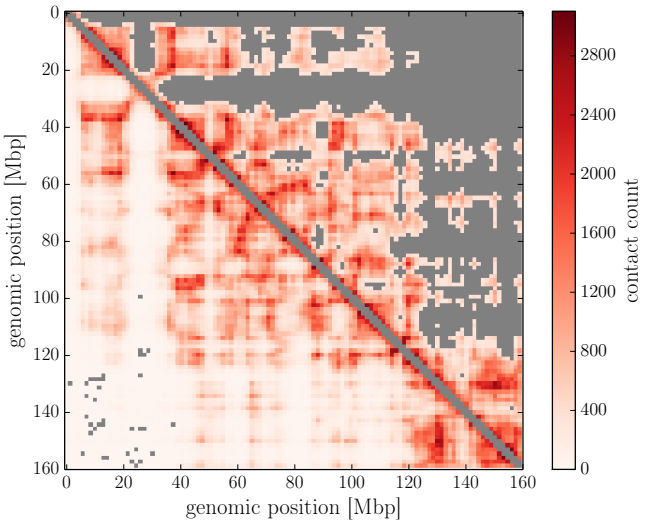

Figure 4.19: Back-calculated interaction frequency matrix used as fake data. Lower triangular matrix: full interaction frequency matrix. Upper triangular matrix: thinned interaction frequency matrix used in the calculations. 50 % of interactions with lowest counts were discarded and are shown in gray. Gray entries in the full data are bins with zero interactions and have not been taken into accout for the calcuations.

and again sum element-wise the two resulting matrices. This way we obtain a mock interaction frequency matrix: it contains contacts formed in models of the mouse X chromosome calculated from two different cells. The result is shown in Fig. 4.19. The rationale behind this is that we might be able to reproduce structures, which are likely in the single cell HiC ensembles, in the ensemble calculation.

As we only try to establish a proof-of-concept for our idea, we choose a low resolution of 111 beads corresponding to 1.5 Mbp each. Again refering to the nuclear density estimate of $^{12\,\mathrm{Mbp}}/_{\mu\mathrm{m}^3}$ given by Rosa and Everaers [2008], the diameter of one bead roughly corresponds to $d_0 = 620\,\mathrm{nm}$. At this resolution, our code runs sufficiently fast to test parameters and, most importantly, the RE schedule. To save computation time, as discussed before, we discard some of the low-frequency interactions. Only the 50 % of entries with the highest contact counts are taken into account. While this may sound rather harsh, we nevertheless retain 91 % of the total contact count; confirming that we indeed only disregard low-frequency interactions. Furthermore, in each gradient evaluation, we only take into accout a randomly chosen 10 % of the data. It is not exactly clear, whether this procedure in fact constitutes a valid MCMC sampling algorithm, as the proposal is not deterministic anymore. Empirically, we find that sampling converges faster and leads to the same results as choosing a subset of the data for each HMC move and keeping it constant during the MD trajectory.

Running ISD simulations with 159 replicas, we find that the sampled structural ensembles are able to reproduce the interaction frequency matrix; a Mantel test gives a Pearson's correlation of 98 %. But the structures also mostly comply with the structural prior distribution. Fig. 4.20 shows the reference interaction matrix and the back-calculated interaction matrix of the last of 45000 samples as well as histograms of $i, i+1$ and all other distances.

Visualization of the structural ensembles shows that during local sampling,

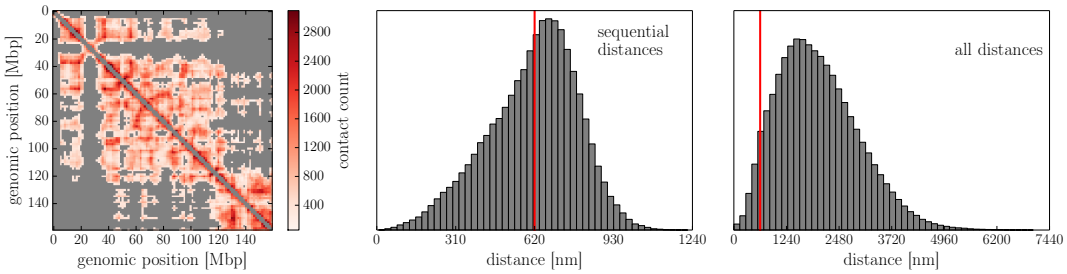

Figure 4.20: Qualitative validation of inferred structural ensembles with $M = 20$ members from back-calculated ensembles.

Left: experimental (lower triangular matrix) and back-calculated (upper triangular matrix) interaction frequency matrix. *Middle:* pairwise distances for chain neighbours and all other beads (*right*). These distances are restrained by the prior distribution; red lines show the bead diameter of 620 nm, which is the lower limit for distant interactions and the upper limit for nearest-neighbour distances.

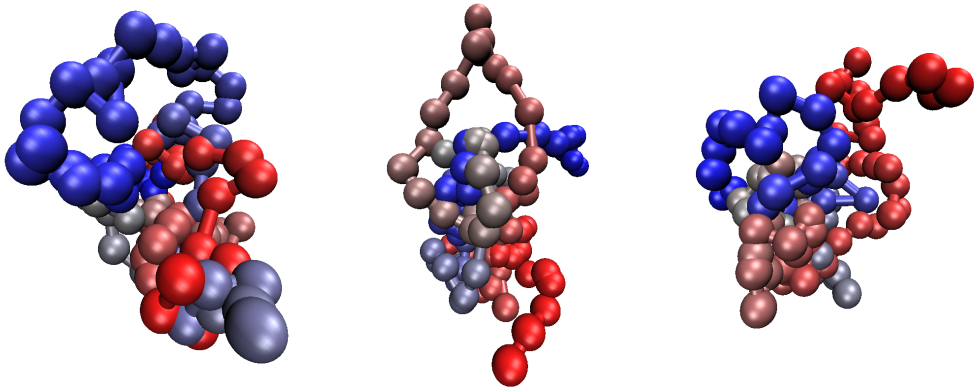

Figure 4.21: Three out of 20 distinct conformations in an inferred structural ensemble.

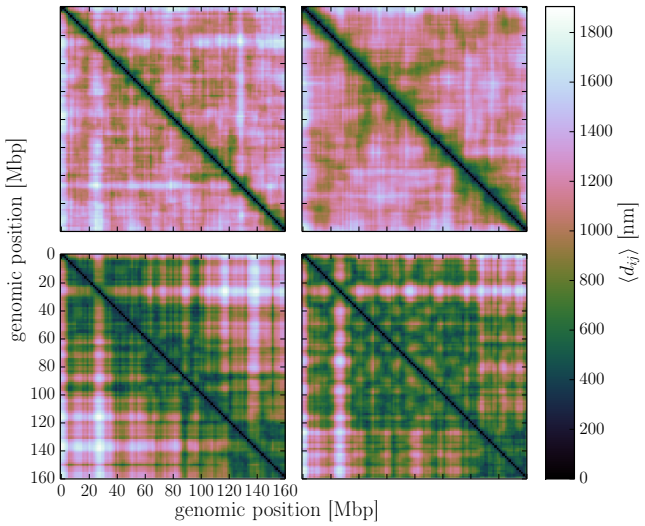

Figure 4.22: Average distance matrices of structures assigned to components of a two-component structural mixture model (*top*) and of reference ensembles (*bottom*).

each structure "slot" is occupied by a specific conformation, which changes only slightly. Fig. 4.21 shows three conformations of the same sample ensemble. These distinct slot occupancies only change through accepted Replica Exchange moves. We can now try to recover the original cell 1 and 4 X chromosome ensembles from our ensembles. To this end, we use a method which models structural ensembles as Gaussian mixtures [Hirsch and Habeck, 2008]. Its only parameter is the number of components, which we set to two. We find that the sub-populations for each component are very heterogenous. Average distance matrices (Fig. 4.22) show correlations calculated by Mantel tests [Mantel, 1967] between 40% and 65 %; in fact, the highest correlations are between the distance matrices of the reference ensembles and of the two clusters, respectively. Our method is thus not able to clearly recover two subpopulations corresponding to the reference ensembles, which is not too surprising given that structural variability in the latter is very high.

Considering the sampled weights for each structure, we find that weights span a broad range. Fig. 4.23 shows the histogram of all sampled weights and we notice that a significant part of them is, in comparison the rest, very small.

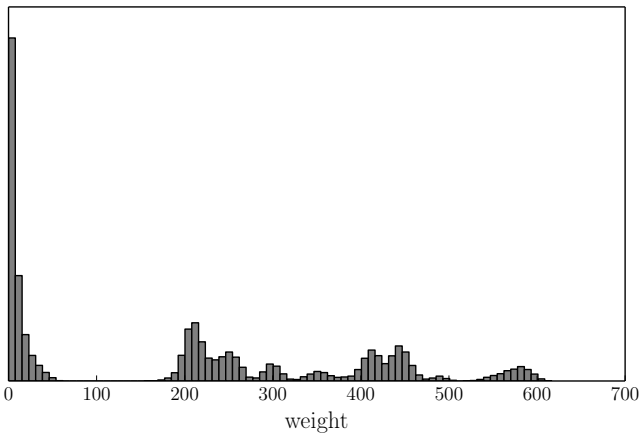

Figure 4.23: Histogram of sampled weights. Only a part of the population carries significant weight and thus contributes most to the likelihood.

43 % of all sampled weights are smaller than 50, while only 22 % take values greater than 400. Structures corresponding to the many small weights thus might only contribute to low-frequency interactions or are possibly not required at all.

5 Summary and outlook

5.1 Replica Exchange with Non-equilibrium Switches (RENS) tested on complex protein systems

In Sec. 2, we applied RENS to coarse-grained protein systems with varying number of degrees of freedom and to the problem of sampling from the ISD posterior of Ubiquitin. We were able to confirm that RENS is indeed able to increase acceptance rates for swaps between neighbouring replicas. This comes at the cost of significantly increased computation time due to the calculation of the non-equilibrium trajectories. For this reason, RENS is much less efficient than ordinary, instantaneous Replica Exchange.

But the field of non-equilibrium statistical mechanics is young and several avenues might lead to increased RENS efficiency. We tested three different ways to calculate thermostatted non-equilibrium trajectories dragging states from one ensemble into the other. These represent three classes of dynamics: Markov Chain Monte Carlo (HMCRENS), molecular dynamics with simulated friction and random collisions (LMDRENS) and molecular dynamics with only random collisions, which reset momenta completely (AMDRENS). All three classes rely on random processes. The Nosé-Hoover thermostat [Hoover, 1985; Martyna et al., 1992; Nosé, 1984], on the other hand, is completely deterministic. An expression for work performed during a Nosé-Hoover thermostatted trajectory is given in [Ballard, 2012]. These dynamics do not conserve phase space volume [Smit and Frenkel, 2002] and the work hence includes a Jaco-

bian, which in general will be difficult to evaluate analytically, but can be easily accumulated during discrete timesteps. Deterministic dynamics might avoid possible looping back of a trajectory due to MCMC move rejections or momentum draws and could thus be beneficial for RENS.

Another open question is the optimal dependence of the switching parameter $\lambda(t)$ on the simulation time. In this work we assumed a linear protocol, which most certainly is not an optimal choice in the sense that it minimizes the work expended to drive the system from state A to state B. The distance between two thermodynamic states in equilibrium can be measured by the thermodynamic length both for macroscopic [Ruppeiner, 1979; Salamon and Berry, 1983; Weinhold, 1975] and, as recently shown by Crooks, microscopic systems. Thermodynamic length is not a state function, but explicitly depends on the path taken through the space of thermodynamic states. It forms a Riemannian manifold, in which paths minimizing dissipation are geodesics for slow, but finite-time switches [Crooks, 2007; Nulton et al., 1985; Salamon and Berry, 1983]. The metric is given by the Fisher informations of the corresponding equilibrium distributions [Burbea and Rao, 1982]. If thus in a RENS non-equilibrium trajectory dissipation is minimized, a minimum amount of work is expended and we can expect an optimal acceptance rate.

Sivak and Crooks [2012] generalize the notion of thermodynamic length and derive it directly from linear response theory. They derive important properties of optimal paths. These trajectories of minimal work are, for example, independent of the non-equilibrium trajectory duration, to which the expended work is inversely proportional. Furthermore, excess work is accumulated at a constant rate. In Zulkowski et al. [2012], this approach was explored further by making use of results in Riemannian geometry to find, for the first time, a closed expression for the optimal path of a particular simple, stochastic system.

We can thus hope that, using these very recent results, approximative optimal switching protocols can be found for realistic systems, which may be able to unleash the full potential of RENS.

Finally, if finding optimal protocols using the thermodynamic length frame-

work should prove unsuccessful, RENS would most likely also profit from an optimized replica exchange schedule.

5.2 Outlook on an adaptive Replica Exchange scheme

The class of Replica Exchange (RE) methods, to which RENS belongs, requires setting a (temperature) schedule interpolating from the target temperature to a high temperature, at which a systems' probability distribution is more easy to sample from. This choice is not trivial. In Sec. 3, we suggested a new idea to solve this problem. Arguing that accurate estimates of normalization constants result from good sampling, we borrowed a result about statistically optimal interpolating distributions for thermodynamic integration [Gelman and Meng, 1998] and investigated whether these distributions improve sampling quality compared to a simple linear log-probability interpolation with a preliminary implementation.

To this end, we simulated a particle in a rough, one-dimensional energy landscape; which poses a problem for local sampling using a simple Metropolis-Hastings scheme. As a "high-temperature" distribution we chose a relatively flat normal distribution, from which sampling is easy. The statistically optimal interpolation schedule turned out to be qualitatively similar to the linear interpolation. We then performed RE simulations using both interpolation methods and noticed that acceptance rates are indeed comparable, but somewhat lower for the two lowest-temperature replicas; confirming that the schedules are similar, but not identical. Using the standard error as a measure of sampling quality, based on these RE simulations, we analyze both the accuracy of estimates of the mean extended ensemble log-probability and of the mean log-probability of the multimodal target distribution. Unfortunately, standard errors for both quantities were higher when using the statistically optimal interpolation schedule, indicating better log-probability estimates when using the naive linear log-probability interpolation.

A possible reasons for this result could be that the derivation of the statistically optimal schedule relies on independent samples drawn from all distributions, including the target distribution. This assumption is obviously violated already by construction, because local sampling by Random Walk Metropolis-Hastings is not able to cross the energy barriers and thus at least between successful replica swap attempts, sample are strongly correlated. Gelman and Meng already mention that the lower bound on the variance of the free energy difference (Eq. 3.4) in practice cannot be reached for exactly this reason.
Furthermore, our assumption that optimizing free energy estimates obtained by thermodynamic integration also means optimizing sampling quality might have to be questioned and investigated seperately. Then Gelman and Meng's optimality result might indeed only be useful for thermodynamic integration, which is at the heart of its derivation. In this case, we could divert our method from its intended use and regard it as an efficient, automatic scheme for free energy estimation.
The interpolating distributions presented here is not the only optimality result discussed by Gelman and Meng. They also give a set of Euler-Lagrange equations for the general case of several switching parameters λ if a fixed family $p_\lambda(x)$ of interpolating distributions is already given. This would translate to, e.g., choosing an exponential decay in the inverse temperature β in RE simulations of a Boltzmann ensemble and determining an optimal temperature spacing between adjacent replicas.

5.3 Bayesian structure determination from HiC data

We were able to demonstrate in Sec. 4 that the ISD framework is also applicable to chromatin structure determination from single cell HiC data. All advantages of ISD carry over to this new application: automatic estimation of nuisance parameters, most notably the error or force constant, and exhaustive sampling of the combined space of conformational degrees of freedom and

nuisance parameters by powerful MCMC methods allowed us to objectively determine an ensemble of conformations, which, under our chosen likelihoods and prior information, truly represents the information available.

This information is very sparse. There are very little data points compared to the size of the X chromosome; at our resolution we have only ≈ 1.6 contacts/bead restraining the structure. In comparison, in a typical NMR structure determination, data is much richer: at least 10, usually even more, restraints per residue [Kwan et al., 2011] determine, along with a molecular forcefield, the native conformation of a protein. This comparison is, of course, not fair, as NMR structure determination relies on rich data from an ensemble of molecules exhibiting similar folds.

But not only the data is sparse: at the extremely coarse resolution the sparse data forces us to choose, little can be said about the polymer physics of chromatin and the large beads may introduce artificial volume exclusion. One way to improve upon this is to choose a different representation of the discrete bins of a HiC map. Zylindrical or ellyptical elements could replace the spherical beads and so more accurately reproduce the actual proportions of the 30 nm fiber. This approach was taken in, e.g., [Wong et al., 2012]. The drawback is that it is harder to compute a soft volume exclusion for non-spherical monomers. This results in increased computation time.

There is also further information we can take into account: Nagano et al. [2013] performed FISH chromsome paint experiments to determine typical diameters of of X chromosome territories. This information can be easily included in an ISD calculation in the form of an additional likelihood. One can heuristically establish a relation between the radius of gyration and the size of a molecule and could then, by means of a Gaussian error model, ask a structure to have a gyration radius corresponding to the average X chromosome territory diameter. The corresponding error can be deduced from the experimental data, as FISH measurements have been performed on several cells. Furthermore, it might be worthwile to consider a forcefield with also attractive interactions, which would lead to the sampling of more compact structures. This might be especially useful for ISD calculations using a contact-based likelihood, as the

attractive force between two beads is very small when they are far away from each other.

We were also able to show that it is possible to infer structures of two-copy chromosomes from single cell HiC data by borrowing a method to deal with ambigous restraints from NMR structure determination. The problem of demixing a single cell HiC contact matrix is thus elegantly solved, at least for two non-interacting copies of a chromosome.

Demixing a population HiC contact matrix, on the other hand, is much more difficult, as data from millions of copies of the same molecule is contained in a single contact matrix. We thus proposed an extension of ISD to ensemble-based modeling from HiC data relying on sampling not single conformations, but small, hopefully representative ensembles from a posterior distribution. The result of this method is thus a "hyper-ensemble"; an ensemble consisting of ensembles, each of which has its own probability weights and nuisance parameters assigned. These comparatively small populations of structures were asked to reproduce the experimental interaction frequency matrix. The core of this approach is a likelihood, in which a mock interaction frequency matrix is back-calculated from some single structures by summing the respective quasi-binary contact matrices. We assigned each ensemble member a weight, which conveniently serves two purposes: first, it allows structures to influence the data back-calculation with different strengths and thus can tell us, which and how many structures in our test ensemble are essential to reproduce an interaction frequency matrix. Second, the weights accomodate for the fact that the size of the modeled ensembles is orders of magnitude smaller than the experimental one, thus effectively scaling the back-calculated matrix and so making it directly comparable to the experimental data.

We tested this approach on an interaction frequency matrix back-calculated from X chromosome structural ensembles obtained from single cell HiC data and found that already a few structures are sufficient to reproduce the input data. The model ensembles also fulfilled the prior information. These results make us optimistic that in real applications, too, our approach will produce structural ensembles which are compatible with both prior information and

data. Modeling of chromosomes from population HiC data will also give us a much broader choice of data sets and results from the literature to compare our results with. The fact that the computational cost for evaluating the likelihood increases quadratically with the number of beads would suggest it might be sensible to first validate this application of ISD on well-resolved population HiC data of only a specific region of a chromosome of biological interest. This would allow us to choose a modeling resolution adapted to the resolution of the data, while keeping computation time reasonable.

An obvious question is the optimal number of ensemble members: if we simulate too little structures per ensemble, we will not correctly reproduce the interaction frequency matrix. With a large number, on the other hand, reproducing the data will be easier, but computation time increases and we risk overfitting. We could use model comparison techniques such as the Akaike Information Criterion (AIC; Akaike [1998]) or the Bayesian Information Criterion (BIC; Schwarz [1978]), which include a penalty on the number of parameters used to fit a model.

We also proposed several ways to save computation time, which need to be investigated further. Instead of disregarding a fixed, arbitrary fraction of the data in the first place, operating under the assumption that very low contact counts are not as important to the structures as highly populated bins, one should find a more objective way to set this cut-off. Furthermore, as we showed that neglecting large parts of the data when calculating gradients in HMC improves performance, a systematic investigation should be carried out as to how many and which data points can be disregarded in this sampling step. This might allow for increased efficiency and thus for higher resolutions or the modeling of larger parts of the genome.

A further line of possible improvements of our method concerns the structural prior distribution: population HiC data are much richer than single cell HiC and, depending on the size of the genomic region modeled, we could incorporate more detailed prior information about, e.g., persistence lengths ($\approx 100\,\mathrm{nm}$). Similar as in the string-binder-switch model proposed in [Barbieri et al., 2012], we could introduce binding sites on our model polymer which can bind to other

regions of that type by means diffusing binind particles. These would mock the behaviour of, e.g., CTCF-binding factors or transcription factories and effectively introduce attractive interactions independent of the data. Depending on the resolution, these binding sites could be placed at positions in the beads-on-a-string model corresponding to their genomic position.

It may also be worth reconsidering our approach of just multiplying the single structural prior distributions: as soon as one ensemble member significantly violates the structural prior, it will downweight the whole ensemble, even if the other structures comply with the prior and the likelihood. We thus possibly are rejecting a lot of reasonable ensembles with one outlier structure in the MCMC moves, which might not only be a waste of computational resources, but also a sign of a too strict or even misspecified structural prior. After all, we are still coarse-graining and, e.g., the volume exclusion we introduce does not necessarily need to hold strictly.

Acknowledgements

This thesis would not exist if it was not for the help of countless people.
I would like to thank Burkhard Rost for the uncomplicated collaboration and for accepting me as an external PhD student in his lab.
During the last four and a half years, Michael Habeck was always available via Skype or in person to supervise my projects. Thank you for all your input and a lot of patience!
Michael Nilges kindly hosted me for three years in the Structural Bioinformatics lab at Institut Pasteur. I had a great time under the best working conditions imaginable and thank him for supervision and many valuable suggestions.
My PhD time would not have been such a good experience without my colleagues, both in Tübingen and in Paris. I am particularly indebted to Nathan Desdouits and Guillaume Bouvier for countless Python tips & tricks and to Isidro Cortés Ciriano, Silke Wieninger and Yannick Spill for many discussions on or off the topic of science. Tru Huynh was always willing to help and maintained the luxurious BIS computing infrastructure. Renée Communal and Maya Um were an indispensable help for dealing with the Pasteur administration. Merci à toutes et tous mes collègues francophones, qui m'ont appris une grande partie de mon français et qui m'ont aussi introduit aux mystères de la langue familière.
Pasteur and non-Pasteur friends from all over the world made my time in- and outside the labs in Paris and Tübingen very enjoyable. Special thanks to Eva Boritsch and Anncharlott Berglar for showing me what *real* biologists do all day.
By proof-reading almost all of it, Anna Howell prevented this document from becoming one giant garden path sentence and Anna Gueiderikh made sure I got the biology right.

Finally, I would like to thank my family and, most of all, my parents for their unconditional support - not only over the last few years.

Bibliography

Marc Adrian, Jacques Dubochet, Jean Lepault, and Alasdair W. McDowall. Cryo-electron microscopy of viruses. *Nature*, 308(5954):32–36, Mar 1984.

Daniel Aird, Michael Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2):R18, 2011.

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

Hans C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980.

Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and MichaelI. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50 (1-2):5–43, 2003.

M Madan Babu, Sarath Chandra Janga, Ines de Santiago, and Ana Pombo. Eukaryotic gene regulation in three dimensions and its impact on genome evolution. *Current Opinion in Genetics & Development*, 18(6):571 – 582, 2008.

Andrew J. Ballard. *Exploring Equilibrium Systems with Nonequilibrium Simulations*. PhD thesis, University of Maryland, College Park, 2012.

Andrew J. Ballard and Christopher Jarzynski. Replica exchange with nonequi-

librium switches. *Proceedings of the National Academy of Sciences*, 106(30): 12224–12229, 2009.

Andrew J. Ballard and Christopher Jarzynski. Replica exchange with nonequilibrium switches: Enhancing equilibrium sampling by increasing replica overlap. *The Journal of Chemical Physics*, 136(19):194101, 2012.

Mariano Barbieri, Mita Chotalia, James Fraser, Liron-Mark Lavitas, Josée Dostie, Ana Pombo, and Mario Nicodemi. Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences*, 109(40):16173–16178, 2012.

Christian Bartels and Martin Karplus. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *Journal of Computational Chemistry*, 18(12):1450–1462, 1997.

Alberto Bartesaghi, Alan Merk, Soojay Banerjee, Doreen Matthies, Xiongwu Wu, Jacqueline L. S. Milne, and Sriram Subramaniam. 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science*, 348(6239):1147–1151, 2015.

Davide Baù, Amartya Sanyal, Bryan R Lajoie, Emidio Capriotti, Meg Byron, Jeanne B Lawrence, Job Dekker, and Marc A Marti-Renom. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology*, 18(1):107–114, 2011.

JGJ Bauman, J Wiegant, P Borst, and P Van Duijn. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Experimental cell research*, 128 (2):485–490, 1980.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

Jay Boris. A vectorized "near neighbors" algorithm of order N using a monotonic logical grid. *Journal of Computational Physics*, 66(1):1 – 20, 1986.

Guillaume Bouvier, Nathan Desdouits, Mathias Ferber, Arnaud Blondel, and Michael Nilges. An automatic tool to analyze and cluster macromolecular conformations based on self-organizing maps. *Bioinformatics*, 2014.

Shelagh Boyle, Susan Gilchrist, Joanna M. Bridger, Nicola L. Mahy, Juliet A. Ellis, and Wendy A. Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human Molecular Genetics*, 10(3):211–219, 2001.

AT Brunger, GM Clore, AM Gronenborn, R Saffrich, and M Nilges. Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. *Science*, 261(5119):328–331, 1993.

Axel T. Brunger. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(6359):472–475, Jan 1992.

Axel T Brünger. Version 1.2 of the Crystallography and NMR system. *Nature protocols*, 2(11):2728–2733, 2007.

Axel T Brünger, Paul D Adams, G Marius Clore, Warren L DeLano, Piet Gros, Ralf W Grosse-Kunstleve, J-S Jiang, John Kuszewski, Michael Nilges, Navraj S Pannu, et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography*, 54(5):905–921, 1998.

Jacob Burbea and C.Radhakrishna Rao. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *Journal of Multivariate Analysis*, 12(4):575 – 596, 1982.

Giovanni Bussi and Michele Parrinello. Accurate sampling using Langevin dynamics. *Phys. Rev. E*, 75:056707, May 2007.

John Chodera. Private communication, 2012.

John D. Chodera, William C. Swope, Jed W. Pitera, Chaok Seok, and Ken A. Dill. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *Journal of Chemical Theory and Computation*, 3(1):26–41, 2007.

R. T. Cox. Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, 14(1):1–13, 1946.

Thomas Cremer and Marion Cremer. Chromosome Territories. *Cold Spring Harbor Perspectives in Biology*, 2(3), 2010.

Gavin E Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *Journal of Statistical Physics*, 90(5-6):1481–1487, 1998.

Gavin E Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3): 2721, 1999a.

Gavin E Crooks. *Excursions in Statistical Dynamics*. PhD thesis, University of California at Berkely, 1999b.

Gavin E Crooks. Path-ensemble averages in systems driven far from equilibrium. *Physical review E*, 61(3):2361, 2000.

Gavin E. Crooks. Measuring Thermodynamic Length. *Phys. Rev. Lett.*, 99: 100602, Sep 2007.

Elzo de Wit and Wouter de Laat. A decade of 3C technologies: insights into nuclear organization. *Genes & Development*, 26(1):11–24, 2012.

Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing Chromosome Conformation. *Science*, 295(5558):1306–1311, 2002.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B (Statistical Methodology)*, 39(1):1–38, 1977.

Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398): 376–380, May 2012.

Josée Dostie, Todd A. Richmond, Ramy A. Arnaout, Rebecca R. Selzer,

William L. Lee, Tracey A. Honan, Eric D. Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D. Green, and Job Dekker. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10): 1299–1309, 2006.

Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C Anthony Blau, and William S Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, 2010.

Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216 – 222, 1987.

Denis J. Evans and Debra J. Searles. Equilibrium microstates which generate second law violating steady states. *Phys. Rev. E*, 50:1645–1648, Aug 1994.

Alan M Ferrenberg and Robert H Swendsen. New Monte Carlo technique for studying phase transitions. *Physical review letters*, 61(23):2635, 1988.

Alan M Ferrenberg and Robert H Swendsen. Optimized Monte Carlo data analysis. *Physical Review Letters*, 63(12):1195, 1989.

Henrik Flyvbjerg and Henrik Gordon Petersen. Error estimates on averages of correlated data. *The Journal of Chemical Physics*, 91(1):461–466, 1989.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

Stuart Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721–741, Nov 1984.

Charles J Geyer. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proc. 23rd Symp. on the Interface*. Interface Foundation of North America, 1991.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Alan Grossfield and Daniel M. Zuckerman. Chapter 2: Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. In Ralph A. Wheeler, editor, *Annual Reports in Computational Chemistry*, volume 5 of *Annual Reports in Computational Chemistry*, pages 23 – 48. Elsevier, 2009.

Lars Guelen, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B. Faza, Wendy Talhout, Bert H. Eussen, Annelies de Klein, Lodewyk Wessels, Wouter de Laat, and Bas van Steensel. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, Jun 2008.

P. Güntert, C. Mumenthaler, and K. Wüthrich. Torsion angle dynamics for NMR structure calculation with the new program Dyana1. *Journal of Molecular Biology*, 273(1):283 – 298, 1997.

Peter Güntert. Automated NMR structure calculation with CYANA. In *Protein NMR Techniques*, pages 353–378. Springer, 2004.

Peter Güntert, Werner Braun, and Kurt Wüthrich. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *Journal of Molecular Biology*, 217(3):517 – 530, 1991.

M. Habeck. Ensemble annealing of complex physical systems. *ArXiv e-prints*, March 2015.

Michael Habeck. Evaluation of marginal likelihoods via the density of states. In *International Conference on Artificial Intelligence and Statistics*, pages 486–494, 2012a.

Michael Habeck. Bayesian Estimation of Free Energies From Equilibrium Simulations. *Phys. Rev. Lett.*, 109:100601, Sep 2012b.

Michael Habeck, Michael Nilges, and Wolfgang Rieping. Replica-Exchange Monte Carlo Scheme for Bayesian Data Analysis. *Phys. Rev. Lett.*, 94: 018105, Jan 2005.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric numerical integration illustrated by the Störmer-Verlet method. *Acta Numerica*, 12: 399–450, 5 2003.

David Hames and Nigel Hooper. *BIOS Instant Notes in Biochemistry*. Garland Science, 4 edition, 2011.

W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Torsten Herrmann, Peter Güntert, and Kurt Wüthrich. Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA. *Journal of Molecular Biology*, 319(1):209 – 227, 2002.

Michael Hirsch and Michael Habeck. Mixture models for protein structure ensembles. *Bioinformatics*, 24(19):2184–2192, 2008.

William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, Mar 1985.

Thomas A Hopf, Charlotta PI Schärfe, João PGLM Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre MJJ Bonvin, and Debora S Marks. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, 3:e03430, 2014.

Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S. Liu. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, Siddarth Selvaraj, Jennifer Fang, Bing Ren, and Jun S. Liu. Bayesian Inference of Spatial Organizations of Chromosomes. *PLoS Comput Biol*, 9(1):e1002893, 01 2013.

Jim R. Hughes, Nigel Roberts, Simon McGowan, Deborah Hay, Eleni Giannoulatou, Magnus Lynch, Marco De Gobbi, Stephen Taylor, Richard Gibbons, and Douglas R. Higgs. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet*, 46(2):205–212, Feb 2014.

Koji Hukushima and Koji Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

R. Hulspas and J.G.J. Bauman. The use of fluorescent in situ hybridization for the analysis of nuclear architecture by confocal microscopy. *Cell Biology International Reports*, 16(8):739 – 747, 1992.

Gerhard Hummer and Attila Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences*, 98(7):3658–3661, 2001.

Yukito Iba. Extended Ensemble Monte Carlo. *International Journal of Modern Physics C*, 12(05):623–656, 2001.

Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Meth*, 9(10):999–1003, Oct 2012.

Roland Jäger, Gabriele Migliorini, Marc Henrion, Radhika Kandaswamy, Helen E. Speedy, Andreas Heindl, Nicola Whiffin, Maria J. Carnicer, Laura Broome, Nicola Dryden, Takashi Nagano, Stefan Schoenfelder, Martin Enge, Yinyin Yuan, Jussi Taipale, Peter Fraser, Olivia Fletcher, and Richard S. Houlston. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun*, 6, Feb 2015.

Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690, 1997a.

Christopher Jarzynski. Equilibrium free-energy differences from nonequilib-

rium measurements: A master-equation approach. *Physical Review E*, 56 (5):5018, 1997b.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Herman Kahn and Ted Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Math Series*, 12:27–30, 1951.

Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotech*, 30(1):90–98, Jan 2012. Research.

Helmut G Katzgraber, Simon Trebst, David A Huse, and Matthias Troyer. Feedback-optimized parallel tempering Monte Carlo. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(03):P03018, 2006.

J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, 181(4610):662–666, Mar 1958.

A. Khalil, J.L. Grant, L.B. Caddle, E. Atzema, K.D. Mills, and A. Arneodo. Chromosome territories have a highly nonspherical morphology and nonrandom positioning. *Chromosome Research*, 15(7):899–916, 2007.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

David A Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *The Journal of chemical physics*, 117(15):6911–6914, 2002.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

Aminata Kone and David A. Kofke. Selection of temperature intervals for

parallel-tempering simulations. *The Journal of Chemical Physics*, 122(20): 206101, 2005.

Werner Kühlbrandt. Cryo-EM enters a new era. *eLife*, 3, 2014.

Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

Ann H Kwan, Mehdi Mobli, Paul R Gooley, Glenn F King, and Joel P Mackay. Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS journal*, 278(5):687–703, 2011.

S. Lan, V. Stathopoulos, B. Shahbaba, and M. Girolami. Lagrangian Dynamical Monte Carlo. *ArXiv e-prints*, November 2012.

Paul Langevin. Sur la théorie du mouvement Brownien. *C. R. Acad. Sci. (Paris)*, 146, 1908. [English translation: On the theory of Brownian Motion, Am. J. Phys. 65, 1079 (1997)].

Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3D genome reconstruction from chromosomal contacts. *Nat Meth*, 11(11):1141–1143, Nov 2014. Brief Communication.

Wenyuan Li, Ke Gong, Qingjiao Li, Frank Alber, and Xianghong Jasmine Zhou. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics*, 31(6):960–962, 2015.

Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive Map-

ping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293, 2009.

Jens P. Linge, Michael Habeck, Wolfgang Rieping, and Michael Nilges. ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics*, 19(2):315–316, 2003.

Blanca López-Méndez and Peter Güntert. Automated Protein Structure Determination from NMR Spectra. *Journal of the American Chemical Society*, 128(40):13112–13122, 2006. PMID: 17017791.

Darío G. Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M. Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel, and Stefan Mundlos. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, 161(5):1012 – 1025, 2015.

Edward Lyman and Daniel M Zuckerman. On the structural convergence of biomolecular simulations by determination of the effective sample size. *The Journal of Physical Chemistry B*, 111(44):12876–12882, 2007.

Wenxiu Ma, Ferhat Ay, Choli Lee, Gunhan Gulsoy, Xinxian Deng, Savannah Cook, Jennifer Hesson, Christopher Cavanaugh, Carol B. Ware, Anton Krumm, Jay Shendure, Carl Anthony Blau, Christine M. Disteche, William S. Noble, and Zhijun Duan. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat Meth*, 12 (1):71–78, Jan 2015.

Nathan Mantel. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, 27(2 Part 1):209–220, 1967.

Glenn J Martyna, Michael L Klein, and Mark Tuckerman. Nosé–Hoover chains:

the canonical ensemble via continuous dynamics. *The Journal of chemical physics*, 97(4):2635–2643, 1992.

William Mattson and Betsy M. Rice. Near-neighbor calculations using a modified cell-linked list method. *Computer Physics Communications*, 119 (2âĂŞ3):135 – 148, 1999.

Martin Mechelke and Michael Habeck. Calibration of Boltzmann distribution priors in Bayesian data analysis. *Phys. Rev. E*, 86:066705, Dec 2012.

Martin Mechelke and Michael Habeck. Bayesian Weighting of Statistical Potentials in NMR Structure Calculation. *PLoS ONE*, 9(6):e100197, 06 2014.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, June 1953.

Leonid A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, 19(1):37–51, 2011.

Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502 (7469):59–64, Oct 2013. Article.

Radford M. Neal. Annealed Importance Sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Michael Nilges. Calculation of Protein Structures with Ambiguous Distance Restraints. Automated Assignment of Ambiguous NOE Crosspeaks and Disulphide Connectivities. *Journal of Molecular Biology*, 245(5):645 – 660, 1995.

Michael Nilges, Aymeric Bernard, Benjamin Bardiaux, Thérèse Malliavin, Michael Habeck, and Wolfgang Rieping. Accurate {NMR} Structures Through Minimization of an Extended Hybrid Energy. *Structure*, 16(9): 1305 – 1312, 2008.

Jerome P. Nilmeier, Gavin E. Crooks, David D. L. Minh, and John D. Chodera. Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation. *Proceedings of the National Academy of Sciences*, 108(45): E1009–E1018, 2011.

Elphege P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Bluthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, May 2012.

Shūichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81(1):511–519, 1984.

J Nulton, P Salamon, Bjarne Andresen, and Qi Anmin. Quasistatic processes as step equilibrations. *The Journal of chemical physics*, 83(1):334–338, 1985.

Yosihiko Ogata. A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, 55(2):137–157, 1989.

Simon Olsson, Jes Frellsen, Wouter Boomsma, Kanti V. Mardia, and Thomas Hamelryck. Inference of Structure Ensembles of Flexible Biomolecules from Sparse, Averaged Data. *PLoS ONE*, 8(11):e79439, 11 2013.

Sheldon B. Opps and Jeremy Schofield. Extended state-space Monte Carlo methods. *Phys. Rev. E*, 63:056701, Apr 2001.

Daan Peric-Hupkes, Wouter Meuleman, Ludo Pagie, Sophia W.M. Bruggeman, Irina Solovei, Wim Brugman, Stefan Gräf, Paul Flicek, Ron M. Kerkhoven, Maarten van Lohuizen, Marcel Reinders, Lodewyk Wessels, and Bas van Steensel. Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. *Molecular Cell*, 38(4):603 – 613, 2010.

Max F Perutz, Michael G Rossmann, Ann F Cullis, Hilary Muirhead, and Georg Will. Structure of hæmoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. *Nature*, 185:416–422, 1960.

Ana Pombo and Niall Dillon. Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol*, 16(4):245–257, Apr 2015.

AJ Rader, C. Chennubhotla, L.W. Yang, I. Bahar, and Q. Cui. The Gaussian network model: Theory and applications. *Normal Mode Analysis–Theory and Applications to Biological and Chemical Systems*, pages 41–63, 2006.

Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665 – 1680, 2014.

Nitin Rathore, Manan Chopra, and Juan J. de Pablo. Optimal allocation of replicas in parallel tempering simulations. *The Journal of Chemical Physics*, 122(2):024111, 2005.

Wolfgang Rieping, Michael Habeck, and Michael Nilges. Inferential Structure Determination. *Science*, 309(5732):303–306, 2005a.

Wolfgang Rieping, Michael Habeck, and Michael Nilges. Modeling Errors in NOE Data with a Log-normal Distribution Improves the Quality of NMR Structures. *Journal of the American Chemical Society*, 127(46):16026–16027, 2005b. PMID: 16287280.

Wolfgang Rieping, Michael Habeck, Benjamin Bardiaux, Aymeric Bernard, Thérèse E. Malliavin, and Michael Nilges. ARIA2: Automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, 23(3):381–382, 2007.

Christian Robert and George Casella. A short history of Markov Chain Monte Carlo: subjective recollections from incomplete data. *Statistical Science*, pages 102–115, 2011.

Christian P. Robert and George Casella. Markov Chains. In *Monte Carlo Statistical Methods*, Springer Texts in Statistics, pages 205–265. Springer New York, 2004.

Angelo Rosa and Ralf Everaers. Structure and Dynamics of Interphase Chromosomes. *PLoS Comput Biol*, 4(8):e1000153, 08 2008.

Mathieu Rousseau, James Fraser, Maria Ferraiuolo, Josee Dostie, and Mathieu Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, 12(1):414, 2011.

George Ruppeiner. Thermodynamics: A Riemannian geometric model. *Phys. Rev. A*, 20:1608–1613, Oct 1979.

Daniel Russel, Keren Lasker, Ben Webb, Javier Velázquez-Muriel, Elina Tjioe, Dina Schneidman-Duhovny, Bret Peterson, and Andrej Sali. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLoS Biol*, 10(1):e1001244, 01 2012.

Peter Salamon and R. Stephen Berry. Thermodynamic Length and Dissipated Availability. *Phys. Rev. Lett.*, 51:1127–1130, Sep 1983.

Michael Sauria, Jennifer Phillips-Cremins, Victor Corces, and James Taylor. HiFive: a tool suite for easy and efficient HiC and 5C data analysis. *Genome Biology*, 16(1):237, 2015.

E. Schöll-Paschinger and C. Dellago. A proof of Jarzynski's nonequilibrium work theorem for dynamical systems that conserve the canonical distribution. *The Journal of Chemical Physics*, 125(5):054105, 2006.

A Schug, T Herges, and W Wenzel. All-atom folding of the three-helix HIV accessory protein with an adaptive parallel tempering method. *Proteins: Structure, Function, and Bioinformatics*, 57(4):792–798, 2004.

Gideon Schwarz. Estimating the Dimension of a Model. *Ann. Statist.*, 6(2): 461–464, 03 1978.

Yigong Shi. A Glimpse of Structural Biology through X-Ray Crystallography. *Cell*, 159(5):995–1014, 2015/11/09 2015.

Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, 129(12): 124105, 2008.

Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*, 38(11):1348–1354, Nov 2006.

David A. Sivak and Gavin E. Crooks. Thermodynamic Metrics and Optimal Paths. *Phys. Rev. Lett.*, 108:190602, May 2012.

David A. Sivak, John D. Chodera, and Gavin E. Crooks. Using Nonequilibrium Fluctuation Theorems to Understand and Correct Errors in Equilibrium and Nonequilibrium Simulations of Discrete Langevin Dynamics. *Phys. Rev. X*, 3:011007, Jan 2013.

Deviderjit Singh Sivia and John Skilling. *Data analysis : a Bayesian tutorial*. Oxford science publications. Oxford University Press, Oxford, New York, 2006.

Berend Smit and Daan Frenkel. *Understanding molecular simulation: from algorithms to applications*. Academic Press, 2002.

Jascha Sohl-Dickstein, Mayur Mudigonda, and Michael R DeWeese. Hamiltonian monte carlo without detailed balance. *arXiv preprint arXiv:1409.5191*, 2014.

Yannick G Spill, Guillaume Bouvier, and Michael Nilges. A convective replica-exchange method for sampling new energy basins. *Journal of computational chemistry*, 34(2):132–140, 2013.

S.M. Stack, D.B. Brown, and W.C. Dewey. Visualization of interphase chromosomes. *Journal of Cell Science*, 26(1):281–299, 1977.

Tim J. Stevens. Private communication, 2013.

Yuji Sugita, Akio Kitao, and Yuko Okamoto. Multidimensional replica-

exchange method for free-energy calculations. *The Journal of Chemical Physics*, 113(15):6042–6051, 2000.

Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.*, 57:2607–2609, Nov 1986.

Holger Then and Andreas Engel. Computing the optimal protocol for finite-time processes in stochastic thermodynamics. *Phys. Rev. E*, 77:041105, Apr 2008.

Bas Tolhuis, Robert-Jan Palstra, Erik Splinter, Frank Grosveld, and Wouter de Laat. Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. *Molecular Cell*, 10(6):1453 – 1465, 2002.

David Tong. Lectures on Kinetic theory, 2012.

Tuan Trieu and Jianlin Cheng. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Research*, 2014.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.

Mark E. Tuckerman, Bruce J Berne, and Glenn J Martyna. Reversible multiple time scale molecular dynamics. *The Journal of chemical physics*, 97(3):1990–2001, 1992.

P.N.T. Unwin and R. Henderson. Molecular structure determination by electron microscopy of unstained crystalline specimens. *Journal of Molecular Biology*, 94(3):425 – 440, 1975.

Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014.

V. Černý. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, 1985.

Loup Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, 159:98–103, Jul 1967.

A.E. Visser, F. Jaunin, S. Fakan, and J.A. Aten. High resolution analysis of interphase chromosome domains. *Journal of Cell Science*, 113(14):2585–2593, 2000.

John von Neumann. Various Techniques Used in Connection with Random Digits. *National Bureau of Standards Applied Math Series*, 12:36–38, 1951.

Siyu Wang, Jinbo Xu, and Jianyang Zeng. Inferential modeling of 3D chromatin structure. *Nucleic Acids Research*, 2015.

F Weinhold. Metric geometry of equilibrium thermodynamics. *Journal of Chemical Physics*, 63(6):2479–2483, 1975.

Hua Wong, Hervé Marie-Nelly, Sébastien Herbert, Pascal Carrivain, Hervé Blanc, Romain Koszul, Emmanuelle Fabre, and Christophe Zimmer. A Predictive Computational Model of the Dynamic 3D Interphase Yeast Nucleus. *Current Biology*, 22(20):1881 – 1890, 2012.

Kurt Wüthrich. The way to NMR structures of proteins. *Nat Struct Mol Biol*, 8(11):923–925, Nov 2001.

Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 43(11):1059–1065, Nov 2011.

Tomoji Yamada and Kyozi Kawasaki. Nonlinear effects in the shear viscosity of critical mixtures. *Progress of Theoretical Physics*, 38(5):1031–1051, 1967.

Bin Zhang and Peter G Wolynes. Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, 112(19):6062–6067, 2015.

Kaiwang Zhang, G Malcolm Stocks, and Jianxin Zhong. Melting and premelting of carbon nanotubes. *Nanotechnology*, 18(28):285703, 2007.

Xin Zhang, Divesh Bhatt, and Daniel M. Zuckerman. Automated Sampling Assessment for Molecular Simulations Using the Effective Sample Size. *Journal of Chemical Theory and Computation*, 6(10):3048–3057, 2010.

Zhi Zhuo Zhang, Guoliang Li, Kim-Chuan Toh, and Wing-Kin Sung. 3D chromosome modeling with semi-definite programming and Hi-C data. *Journal of computational biology*, 20(11):831–846, 2013.

Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjolinder, Anita Gondor, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh Sandhu, Umashankar Singh, Vinod Pant, Vijay Tiwari, Sreenivasulu Kurukuti, and Rolf Ohlsson. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, 38(11):1341–1347, Nov 2006.

Christian Zorn, Christoph Cremer, Thomas Cremer, and Jürgen Zimmer. Unscheduled DNA synthesis after partial UV irradiation of the cell nucleus. *Experimental Cell Research*, 124(1):111 – 119, 1979.

Patrick R Zulkowski, David A Sivak, Gavin E Crooks, and Michael R DeWeese. Geometry of thermodynamic control. *Physical Review E*, 86(4):041148, 2012.