J. Phys.: Condens. Matter **27** (2015) 064120 (12pp) doi:10.1088/0953-8984/27/6/064120

Efficient estimation of contact probabilities from inter-bead distance distributions in simulated polymer chains

Dario Meluzzi and Gaurav Arya

Department of NanoEngineering, University of California San Diego, 9500 Gilman La Jolla, CA 92093, USA

E-mail: garya@ucsd.edu

Received 17 June 2014, revised 31 July 2014
Accepted for publication 26 August 2014
Published 7 January 2015

Abstract

The estimation of contact probabilities (CP) from conformations of simulated bead-chain polymer models is a key step in methods that aim to elucidate the spatial organization of chromatin from analysis of experimentally determined contacts between different genomic loci. Although CPs can be estimated simply by counting contacts between beads in a sample of simulated chain conformations, reliable estimation of small CPs through this approach requires a large number of conformations, which can be computationally expensive to obtain. Here we describe an alternative computational method for estimating relatively small CPs without requiring large samples of chain conformations. In particular, we estimate the CPs from functional approximations to the cumulative distribution function (cdf) of the inter-bead distance for each pair of beads. These cdf approximations are obtained by fitting the extended generalized lambda distribution (EGLD) to inter-bead distances determined from a sample of chain conformations, which are in turn generated by Monte Carlo simulations. We find that CPs estimated from fitted EGLD cdfs are significantly more accurate than CPs estimated using contact counts from samples of limited size, and are more precise with all sample sizes, permitting as much as a tenfold reduction in conformation sample size for chains of 200 beads and samples smaller than 10^5 conformations. This method of CP estimation thus has potential to accelerate computational efforts to elucidate the spatial organization of chromatin.

Keywords: chromatin, interaction frequency, Monte Carlo simulation, EGLD, GLD

(Some figures may appear in colour only in the online journal)

1. Introduction

The tight confinement of chromatin within the cell nucleus and the presence of regulatory and structural DNA binding proteins within the same space naturally lead to the existence of contacts between genomically distant segments of the chromatin fiber [1]. Quantitative information about such contacts can be experimentally obtained from millions of intact cell nuclei by employing techniques based on chemical cross-linking of DNA [2]. For example, high-throughput experiments using the Hi-C technique or one of its variants produce large amounts of DNA sequencing data that can be analyzed to detect contacts across most loci of an entire genome [3–5]. The collection of such contacts yields contact probability (CP) maps that represent the frequency of interaction between different genomic loci, and therefore also contain information about the higher-order spatial organization of chromatin in the cells under study.

To recover such spatial organization from CP maps, various computational methods have been proposed [6]. Some of the most promising among these methods rely on the estimation of CPs from simulated conformations of a bead-chain polymer model representing chromosomes or the chromatin fiber. Comparing the estimated CPs to the corresponding experimental CPs then enables refining the spatial organization of chromatin inferred from the experimental CPs [5, 7, 8]. To improve the speed and scalability of these methods it is desirable that CPs be estimated efficiently from simulation data sets

0953-8984/15/064120+12$33.00

of limited size. One approach is to estimate CPs using contact counts from a sample of bead-chain conformations, so that the CP for a given pair of beads is given by the fraction of conformations where those beads are found to make contact [5, 7, 9]. However, to estimate small probabilities, this approach requires a large number of simulated conformations, which in turn often require substantial computational effort to generate.

Here we describe an alternative CP estimation method that does not require a large conformation sample to estimate relatively small CPs. Specifically, we estimate the CP for each pair of beads from a functional approximation to the cumulative probability distribution function (cdf) of the inter-bead distance. To obtain this approximation we fit the extended generalized lambda distribution (EGLD) [10, 11] to inter-bead distances determined from a sample of simulated bead-chain conformations. The EGLD provides a great variety of distribution shapes by using four adjustable parameters, which can be determined from sample data using the well known method of moments. We found that, for chains of up 200 beads, estimating CPs from fitted EGLD cdfs yields significantly more accurate values than estimating CPs from contact counts if the sample size M is less than about 10^5 conformations and the CPs being estimated are less than about $100/M$. Thus, the proposed method of estimating CPs from simulated conformations of a bead-chain polymer model should be effective in accelerating other computational methods that use such CPs to deduce the spatial organization of the genome from experimental data.

2. Methods

2.1. Contact probabilities from inter-bead distance distributions

Our model system, representing the chromatin fiber, consists of a single linear chain of N beads in some thermodynamic ensemble, and our aim is to estimate the probability of contact between any two beads in the chain. A thermodynamic ensemble corresponds to a population of chain conformations consistent with a given set of macroscopic constraints on the system, such as number of chains, volume, and temperature. In each chain conformation from such a population, any two beads i and j, $1 < j-i < N$, may or may not be making contact. A contact is defined by the condition that the spatial distance $d_{i,j} = |\mathbf{r}_j - \mathbf{r}_i|$ between the beads is smaller than a predefined contact distance d_c. This condition defines the subset of conformations where beads i and j make contact. If we know both the size of such subset and the size of the population, then we can compute the contact probability (CP) $p_{i,j}$ for beads i and j as the size of the subset divided by the size of the population. In practice, both the population and the subset of interest may be too large or complicated to determine their sizes and perform the division. Thus, $p_{i,j}$ can often only be approximated from a representative sample of the population using an appropriate estimation method. Note that in the present study the term sample denotes not a single chain conformation, but a representative collection of conformations extracted from the population.

An obvious way to estimate the contact probability $p_{i,j}$ for a given bead-chain is to obtain a sample of M chain conformations from the population, for example by periodically observing the chain during a sufficiently long molecular dynamics, Brownian dynamics, or Monte Carlo simulation in the thermodynamic ensemble of interest. Then, the conformations in the sample are examined to determine the number of conformations where $d_{i,j} \leqslant d_c$. Dividing this number by the size of the sample M yields an estimate of the contact probability [5, 7],

$$\tilde{p}_{i,j} = \frac{1}{M} \sum_{k=1}^{M} \Theta\left(d_c - d_{i,j}^k\right), \tag{1}$$

where $d_{i,j}^k$ is the distance between beads i and j in the kth conformation, and $\Theta(x)$ is the Heaviside step function, which equals 1 when $x > 0$ and 0 otherwise. A similar definition was used to compute looping probabilities in a polymer chain representing the chromatin fiber [9]. We refer to these contact probabilities as being estimated from contact counts. As the sample size increases, the CP estimate approaches the true CP,

$$\lim_{M\to\infty} \tilde{p}_{i,j} = p_{i,j}. \tag{2}$$

However, in applications where the sample size is limited, the estimation of small CPs with this method may not be reliable or even possible. In fact, equation (1) cannot be used to estimate CPs smaller than $1/M$ from a sample of M conformations.

Another way to estimate $p_{i,j}$ is suggested by defining the true CP not through equation (2), but through the cumulative distribution function (cdf) of inter-bead distances, i.e.

$$p_{i,j} = F_{i,j}(d_c) = \int_0^{d_c} f_{i,j}(x)\,\mathrm{d}x, \tag{3}$$

where $f_{i,j}(x) = P(x < d_{i,j} \leqslant x + \mathrm{d}x)$ is the probability density function of the distance $d_{i,j}$ between beads i and j, and $F_{i,j}(x) = P(d_{i,j} \leqslant x)$ is the corresponding cdf. An analogous formulation in terms of radial distribution functions was used to compute the average number of pairwise contacts per monomer across cluster formations in a linear multiblock copolymer chain under poor solvent conditions [12, 13].

However, for realistic polymer models, an analytical expression of $F_{i,j}(x)$ is generally not available or practical to compute [14, 15], especially when the chain is subjected to arbitrary additional restraints [7]. In this case, if an approximation $\hat{f}_{i,j}(x)$ for $f_{i,j}(x)$, or $\hat{F}_{i,j}(x)$ for $F_{i,j}(x)$, can be obtained from the available sample of conformations, then a suitable estimate of the CP for beads i and j may be obtained from

$$\hat{p}_{i,j} = \hat{F}_{i,j}(d_c) = \int_0^{d_c} \hat{f}_{i,j}(x)\,\mathrm{d}x. \tag{4}$$

We therefore propose an alternative method for estimating inter-bead CPs from samples of bead-chain conformations (figure 1). This method consists of fitting an appropriate functional form for $\hat{F}_{i,j}(x)$ to sampled distance data and then obtaining $\hat{p}_{i,j}$ from equation (4).

2.2. Extended generalized lambda distribution

To obtain $\hat{F}_{i,j}(x)$, we use either the generalized lambda distribution (GLD) [16] or the generalized beta distribution

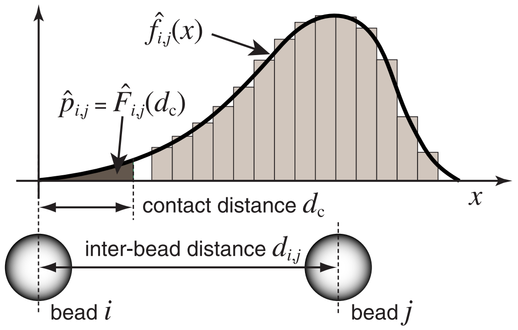
$\hat{f}_{i,j}(x)$
$\hat{p}_{i,j} = \hat{F}_{i,j}(d_c)$
x
contact distance d_c
inter-bead distance $d_{i,j}$
bead i
bead j

Figure 1. Using a fitted cumulative probability distribution function (cdf) to estimate contact probabilities (CPs) from simulated conformations of a bead-chain polymer model. The spatial distance between beads i and j is observed in a sample of M conformations, and an appropriate functional form $\hat{F}_{i,j}(x)$ for the cdf is fitted to the inter-bead distance observations. The CP for beads i and j can then be estimated as $\hat{p}_{i,j} = \hat{F}_{i,j}(d_c)$, where d_c is the contact distance. The black curve represents an approximation $\hat{f}_{i,j}(x)$ of the actual probability density function.

(GBD) [10] depending on the shape of the sampled inter-bead distance distribution. Together, the GLD and GBD are referred to as the extended generalized lambda distribution (EGLD) [10, 11].

The GLD is defined by its quantile function, which is the inverse of the cdf and is also known as percentile function,

$$Q^{(\mathrm{GLD})}(x) = \lambda_1 + \frac{x^{\lambda_3} - (1-x)^{\lambda_4}}{\lambda_2}, \tag{5}$$

where λ_1, λ_2, λ_3, and λ_4 are adjustable parameters. The parameters λ_1 and λ_2 control the location and scale of the distribution, respectively, while λ_3 and λ_4 control its shape, and are thus referred to as shape parameters.

The GBD is defined by its probability density function (pdf), and also contains four adjustable parameters,

$$f^{(\mathrm{GBD})}(x) = \begin{cases} \dfrac{(x-\beta_1)^{\beta_3}(\beta_1+\beta_2-x)^{\beta_4}}{\beta(\beta_3+1,\beta_4+1)\beta_2^{(\beta_3+\beta_4+1)}}, & \text{for } \beta_1 \leqslant x \leqslant \beta_1+\beta_2, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $\beta(a, b)$ is the beta function,

$$\beta(a,b) = \int_0^1 x^{(a-1)}(1-x)^{(b-1)}\,\mathrm{d}x. \tag{7}$$

Here again the parameters β_1 and β_2 control the location and scale of the distribution, respectively, while β_3 and β_4 are the shape parameters.

The four adjustable parameters available in both the GLD and the GBD allow a wide variety of distributions to be represented analytically and, therefore, concisely by each family of distributions. Also, using the EGLD allows a greater range of distribution shapes to be fitted than using only the GLD or only the GBD. Neither the GLD nor the GBD, however, offer an explicit expression for the cdf, which must therefore be evaluated with numerical methods. But before the cdf $\hat{F}_{i,j}(x)$ can be evaluated, its four parameters must be estimated from the available data.

2.3. Method of moments

To estimate the parameters of $\hat{F}_{i,j}(x)$ from a given sample of M bead-chain conformations, we fit the EGLD to the distribution of inter-bead distances $d_{i,j}$ using the method of moments [11]. The first four moments of a random variable X are known as the mean, variance, skewness, and kurtosis of X and are defined as

$$\alpha_1 = E(X), \tag{8}$$

$$\alpha_2 = \sigma^2 = E[(X-\alpha_1)^2], \tag{9}$$

$$\alpha_3 = \frac{E[(X-\alpha_1)^3]}{\sigma^3}, \tag{10}$$

$$\alpha_4 = \frac{E[(X-\alpha_1)^4]}{\sigma^4}, \tag{11}$$

where $E(\cdot)$ is the expectation operator. The corresponding sample moments are given by

$$\hat{\alpha}_1 = \hat{m}_1, \quad \hat{\alpha}_2 = \hat{c}_2, \quad \hat{\alpha}_3 = \frac{\hat{c}_3}{\hat{c}_2^{3/2}}, \quad \hat{\alpha}_4 = \frac{\hat{c}_4}{\hat{c}_2^2}, \tag{12}$$

where the sample moments about the mean [17]

$$\hat{c}_k = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{m}_1)^k = \sum_{j=0}^{k}\binom{k}{j}(-1)^j \hat{m}_{k-j}\hat{m}_1^j, \tag{13}$$

can be computed from the sample non-central moments $\hat{m}_k$, which in turn are computed from a sample of M observations $x_1, x_2, \ldots, x_M$ of the random variable X,

$$\hat{m}_k = \frac{1}{M}\sum_{i=1}^{M} x_i^k. \tag{14}$$

To fit the EGLD using the method of moments with data from a sample of M bead-chain conformations, we first collect the first four sample non-central moments $\hat{m}_k$ of the random variable $X \equiv d_{i,j}$ from the given conformations. Then we compute the sample moments $\hat{c}_k$ and solve the system of four non-linear equations obtained by equating each moment to the corresponding sample moment,

$$\alpha_k = \hat{\alpha}_k, \quad \text{for } k = 1, 2, 3, 4, \tag{15}$$

where the left-hand side of each equation is a function of the four GLD or GBD parameters. Actually, the equations for α_3 and α_4 involve only the shape parameters, i.e. λ_3 and λ_4 or β_3 and β_4, and can therefore be solved as a system of two equations,

$$\begin{cases} \alpha_3 = \hat{\alpha}_3 \\ \alpha_4 = \hat{\alpha}_4. \end{cases} \tag{16}$$

This system has no solutions for the GLD parameters when $1.8(\hat{\alpha}_3^2 + 1) < \hat{\alpha}_4$ [11]. In this case, the GBD can be used instead of the GLD to fit the data. The equations for α_1 and α_2 do involve all four parameters, but can easily be solved for the location and scale parameters, i.e. λ_1 and λ_2 or β_1 and β_2, once the two shape parameters are known. However, solving the system (16) for the shape parameters is a non-trivial task, especially for the GLD. This task must be carried out numerically, as explained in [11].

To determine the GLD parameters from given data, several other fitting methods, besides the method of moments, have been proposed, including the use of percentiles [18], the use of L-moments [19], the starship method [20], discretized methods [21, 22] and maximum likelihood estimation [23]. Usable implementations of these methods are available in the package GLDEX [24] of the R system for statistical computing [25]. However, we do not use these alternative fitting methods in the present study because they require that all observations of inter-bead distance $d_{i,j}$ be fed at once to the fitting procedure. To meet this requirement it would be necessary to record all M bead-chain conformations for each sample obtained by simulation, and such recording in turn would require large amounts of data storage and processing time, when N and M are large. Instead, by using the method of moments we only need to record the four non-central moments (14) of $d_{i,j}$ for each bead pair (i, j), and such moments can be computed incrementally as each conformation is generated by simulation, without having to record all conformations.

2.4. Monte Carlo simulations of bead-chains

To generate samples containing uncorrelated bead chain conformations suitable for estimating CPs using contact counts, with equation (1), or using fitted EGLD cdfs, with equation (4), we performed configurational bias Monte Carlo (MC) simulations [26] of a bead-chain polymer model at constant temperature, as described in [27, 28]. Specifically, we simulated a single chain of N beads connected by rigid bonds, and we investigated chain lengths of N = 25, 50, 100, and 200 beads. The potential energy of each chain,

$$U = \sum_{i=1}^{N-2} U_{\text{bend}}(i) + \sum_{1\leqslant i<j\leqslant N} U_{\text{excl}}(i, j) \tag{17}$$

included contributions for chain stiffness and excluded volume, namely

$$U_{\text{bend}}(i) = \frac{1}{2} k_\theta {\theta_i}^2 \tag{18}$$

and

$$U_{\text{excl}}(i, j) = \begin{cases} 4\varepsilon\left[\left(\dfrac{\sigma}{d_{i,j}}\right)^{12} - \left(\dfrac{\sigma}{d_{i,j}}\right)^{6} + \dfrac{1}{4}\right], & d_{i,j} \leqslant 2^{1/2}\sigma \\ 0, \quad \text{otherwise,} \end{cases} \tag{19}$$

where k_θ is the bending constant, θ_i is the angle between the two bonds connecting beads i, $i + 1$, and $i + 2$, ε is the Lennard–Jones energy parameter, and σ is the bond length. To approximate the physical properties of the 30 nm chromatin fiber, with one bead corresponding to roughly 3 kbp of DNA [29], and with contacts mediated by proteins of roughly 15 nm diameter, the parameters of this model were chosen to be $d_c = 1.5\sigma$, $k_\theta = 4$, $\sigma = 1$, and $\varepsilon = k_B T = 1$, in reduced units [7].

To ensure that the conformations in each sample were uncorrelated, we extracted the conformations periodically from each MC simulation with a sampling period of n_s MC steps, so that the kth conformation in a sample was generated at step kn_s of the MC simulation. To determine an appropriate sampling period n_s for each chain length N, we performed $N_s = 10$ sets of independent MC simulations, using a different value of n_s for each set and generating $M = 10^6$ conformations from each simulation. Using contact counts from each sample, we estimated $\tilde{p}_{i,j}$ from equation (1) for each bead pair (i, j). Thus each set of simulations yielded a sample of N_s independent observations for each $\tilde{p}_{i,j}$. We then compared the sample variance $\tilde{s}^2$ of these observations to the variance ${\sigma_B}^2 = p(1-p)/M$ of the average of M independent Bernoulli random variables with success probability $p = p_{i,j}$. Because $p_{i,j}$ is not known, we used $p \approx \overline{\tilde{p}_{i,j}}$, where the over-line denotes the sample average of $\tilde{p}_{i,j}$ over the N_s observations. Finally, we chose the smallest value of n_s such that $\langle \tilde{s}/\sigma_B \rangle \approx 1$, where the angle brackets denote averaging over all possible bead pairs (i, j), with $1 < j - i < N$ and $\tilde{p}_{i,j} > 10/M$.

2.5. Errors in estimated CPs

2.5.1. Reference CPs. To assess the error performance of the two CP estimation methods considered in this study, one using contact counts and the other fitted EGLD cdfs, we obtained close approximations to the unknown true CPs. To compute these approximations, which we refer to as reference CPs and denote with $p^*_{i,j}$, we estimated CPs from contact counts collected over $N_s = 10$ independent conformation samples, each consisting of a large number $M = 10^7$ of uncorrelated conformations, and we averaged those CPs over the N_s samples, i.e.

$$p^*_{i,j} = \frac{1}{N_s}\sum_{k=1}^{N_s} \tilde{p}^k_{i,j}, \tag{20}$$

where $\tilde{p}^k_{i,j}$ is the CP for beads i and j estimated using equation (1) from the kth conformation sample. Thus, effectively, each reference CP was estimated using contact counts from 10^8 uncorrelated conformations. To confirm the low variability of each $p^*_{i,j}$, and therefore its suitability for use as reference CP, we calculated the standard deviation of the CPs $\tilde{p}^k_{i,j}$ estimated using contact counts from the N_s independent samples of $M = 10^7$ uncorrelated conformations.

2.5.2. Root mean squared fractional deviation. To obtain a measure of average systematic error, or bias, in the CP estimates $\hat{p}_{i,j}$, from fitted EGLD cdfs, relative to the corresponding reference CP $p^*_{i,j}$, we computed the average root mean squared fractional deviation (RMSFD) of $\hat{p}_{i,j}$ using

$$\text{RMSFD} = \sqrt{\frac{1}{N_p}\sum_{i,j}\left(\frac{\hat{p}_{i,j}}{p^*_{i,j}} - 1\right)^2}, \tag{21}$$

where $\hat{p}_{i,j}$ was estimated from a sample of M uncorrelated conformations, N_p is the number of bead pairs (i, j), such that $1 < j - i < N$ and $p^*_{i,j} > 0$, and the summation under the square root is over all such pairs. The RMSFD for the estimates $\tilde{p}_{i,j}$ from contact counts was obtained using the same formula after replacing $\hat{p}_{i,j}$ with $\tilde{p}_{i,j}$. To assess the variability of the RMSFD across conformation samples, we calculated the mean and standard deviation of RMSFD values obtained from $N_s = 10$ independent conformation samples.

2.5.3. Mean ratio of standard deviations. To obtain an average measure of random error in the CP estimates $\hat{p}_{i,j}$ from fitted EGLD cdfs relative to the reference CPs $p^*_{i,j}$, we calculated the mean ratio of standard deviations (MRSD) for $\hat{p}_{i,j}$ using

$$\text{MRSD} = \frac{1}{N_\text{p}} \sum_{i,j} \frac{\hat{s}_{i,j}}{\sigma_\text{B}}, \tag{22}$$

where $\hat{s}_{i,j}$ is the sample standard deviation of $\hat{p}_{i,j}$ calculated using estimates from $N_\text{s} = 10$ independent samples of M conformations, and $\sigma_\text{B}{}^2 = p(1-p)/M$ is the variance of the average of M independent Bernoulli random variables with success probability p equal to the reference CP $p^*_{i,j}$. The ratio $\hat{s}_{i,j}/\sigma_\text{B}$ was averaged over the N_p bead pairs (i, j) such that $1 < j - i < N$, $p^*_{i,j} > 0$, and $\hat{p}_{i,j} > 0$. To assess the variability of the ratio $\hat{s}_{i,j}/\sigma_\text{B}$ across bead pairs, we calculated the standard deviation of that ratio over the same N_p bead pairs used to compute the MRSD. The MRSD of CP estimates $\tilde{p}_{i,j}$ from contact counts was similarly calculated from the above expression by using the sample standard deviation $\tilde{s}_{i,j}$ of $\tilde{p}_{i,j}$ in place of $\hat{s}_{i,j}$, and redefining N_p in terms of $\tilde{p}_{i,j}$.

3. Results

3.1. Monte Carlo simulations yield uncorrelated bead-chain conformations

In the present work we addressed the problem of efficiently estimating contact probabilities (CPs) for pairs of beads in a simulated bead-chain. To this end, we compared two computational methods for estimating CPs, one using contact counts, through equation (1), the other using fitted EGLD cdfs, through equation (4). Both methods take as input a sample of M uncorrelated bead-chain conformations. To generate several such samples, we periodically extracted the conformations from configurational bias Monte Carlo (MC) simulations [26] of a bead-chain polymer model (figure 2).

Because conformations from successive steps of a Markov-chain MC simulation are generally correlated [26], it is necessary to use a sampling period $n_\text{s} > 1$. To determine n_s, we calculated the variance $\tilde{s}^2$ of $\tilde{p}_{i,j}$ over several sets of $N_\text{s} = 10$ independent MC simulations, using a different value of n_s for each set. If $\tilde{p}_{i,j}$ was estimated from independent conformations, then $\tilde{s}^2$ should match the variance $\sigma_\text{B}{}^2 = p(1-p)/M$ of the average of M independent Bernoulli random variables with success probability $p \approx p_{i,j}$, Indeed, we found the average of $\tilde{s}/\sigma_\text{B}$ over bead pairs to approach 1 as n_s increases (figure 3), indicating that, with a sufficiently large sampling period, our MC simulations could produce samples of uncorrelated bead-chain conformations.

In particular, sampling periods of $n_\text{s} = 3$ and $n_\text{s} = 10$ appeared adequate for chains of 100 and 200 beads, respectively. We therefore used sampling periods of 3, 4, 5, and 10 to obtain uncorrelated conformations from all our subsequent MC simulations of chains containing $N = 25$, 50, 100, and 200 beads, respectively.

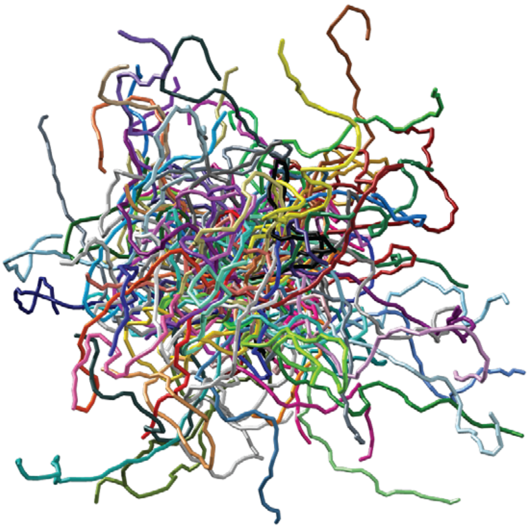

Figure 2. Representative collection of 100 conformations obtained from configurational bias MC simulations of a chain containing $N = 50$ beads. This collection was extracted from a much larger sample of $M = 10^7$ conformations, which were all grown starting at the origin and were used to compute the reference CPs $p^*_{i,j}$ for the simulated chain using equation (20). The image in this figure was generated using the program UCSF Chimera [30].

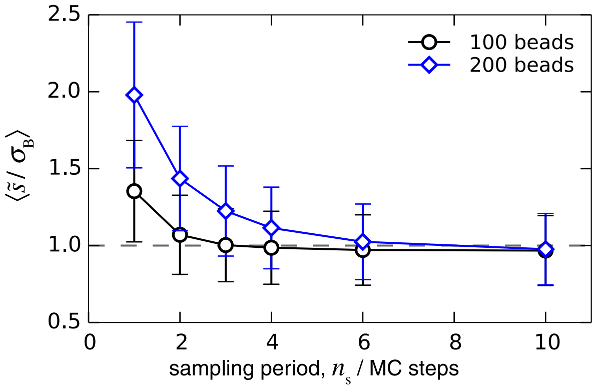

Figure 3. Ratio of sample standard deviation of CP estimates $\tilde{p}_{i,j}$, obtained using contact counts from $M = 10^6$ conformations, to sample standard deviation σ_B of the average of M independent Bernoulli random variables with probability of success equal to the sample mean of the estimated CPs, for chains containing 100 beads (circles and black line) and 200 beads (diamonds and blue line). Sample mean and standard deviation of CPs were calculated from 10 independent MC simulations. The ratio was averaged over all bead pairs (i, j), with $j - i > 1$ and $\tilde{p}_{i,j} > 10/M$, and plotted against the conformation sampling period n_s. Error bars are standard deviations of the ratio over the bead pairs considered. All plots in figures 3–9 were generated using the Python extension modules NumPy and matplotlib [31, 32].

3.2. Log-squared distance moments yield better CPs than distance moments

We next investigated whether CPs can be reliably estimated from EGLD cdfs fitted to inter-bead distance distributions. To obtain estimates $\hat{p}_{i,j}$ of the CP for each pair of beads i and j using a fitted EGLD cdf, we computed the first four non-central sample moments of the inter-bead distance $d_{i,j}$ between beads i and j from a sample of $M = 10^6$ uncorrelated conformations

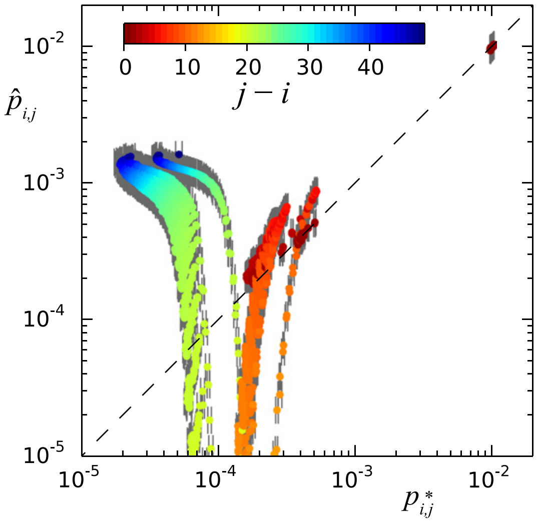

Figure 4. Plot of CPs $\hat{p}_{i,j}$ estimated for a chain of $N = 50$ beads from EGLD cdfs fitted using sample moments of inter-bead distances $d_{i,j}$ versus corresponding reference CPs $p^*_{i,j}$ calculated using contact counts from a large sample of 10^8 uncorrelated conformations. The vertical position of each point is the CP estimated for a particular pair of beads (i, j), $1 < j - i < N$, from a single sample of $M = 10^6$ uncorrelated conformations. The horizontal position of each point is $p^*_{i,j}$. Horizontal error bars are standard deviations over $N_s = 10$ estimates $\tilde{p}_{i,j}$, each calculated from a subsample obtained by splitting the sample of 10^8 conformations into N_s subsamples of equal size. Colors vary over blue, cyan, green, yellow, orange, and red to indicate decreasing separation $j - i$ of the contacting beads along the chain.

generated by MC simulation. We then determined the EGLD parameters using the method of moments [11] and calculated $\hat{p}_{i,j} = \hat{F}_{i,j}(d_c)$, where $\hat{F}_{i,j}(x)$ is the cdf of the fitted EGLD, $d_c = 1.5\sigma$ is the contact distance, and σ is the bond length.

To assess the accuracy of the estimates $\hat{p}_{i,j}$ obtained from EGLD fits, we compared those estimates to corresponding reference CPs $p^*_{i,j}$ that were in turn calculated using contact counts from a large sample of $M = 10^8$ uncorrelated bead-chain conformations. Unfortunately, plotting the estimated CPs against the corresponding reference CPs revealed a rather poor agreement between the two sets of CPs (figure 4). Whereas CPs around 10^{-2}, corresponding to pairs of beads (i, j) with separation $j - i = 2$, were estimated quite accurately, the CPs for most of the other bead pairs were significantly larger or smaller than the corresponding reference CPs (figure 4).

To investigate the cause of this poor agreement between estimated and reference CPs, we collected histograms of inter-bead distances $d_{i,j}$ from the same conformation samples used to compute the sample moments of $d_{i,j}$. We then plotted the fitted EGLD cdf over the corresponding cumulative histogram observed for selected pairs of beads (figures 5(*a*)–(*f*)). For many bead pairs, we found that the fitted EGLD cdf crosses the contact distance d_c either above (figures 5(*b*)–(*d*)) or below (figures 5(*e*) and (*f*)) the trend implied by the cumulative histogram. These results suggest that the poor correlation between reference CPs and corresponding CPs estimated from EGLD cdfs is likely due to a poor fit between the EGLD cdfs and the actual cumulative distribution function at short inter-bead distances.

To improve the fit of the EGLD cdf at short inter-bead distances, we stretched the distribution of such distances by collecting non-central sample moments of $\log(d_{i,j}/\sigma)^2$, rather than $d_{i,j}$. We thus determined EGLD parameters from sample moments of $\log(d_{i,j}/\sigma)^2$ and computed CPs for all pairs of beads by evaluating the corresponding fitted EGLD cdf at $\log(d_c/\sigma)^2$. This simple modification to the fitting procedure resulted in a much better visual agreement between the fitted EGLD cdfs and the corresponding cumulative histograms of inter-bead distance in the region around the log-squared contact distance $\log(d_c/\sigma)^2$ (figures 5(*g*)–(*l*)). Consistently, using sample moments of $\log(d_{i,j}/\sigma)^2$ instead of sample moments of $d_{i,j}$ to determine the EGLD parameters also resulted in a dramatically improved correlation between the reference CPs and the corresponding CPs estimated from fitted EGLD cdfs (figure 6). Although still not perfect, the greater agreement between the two sets of CPs motivated us to investigate the error performance of CPs estimated from EGLD relative to CPs estimated from contact counts.

3.3. EGLD fits incur smaller estimation errors at intermediate sample sizes

To determine how the errors in the CP estimates $\hat{p}_{i,j}$ obtained from fitted EGLD cdfs compare to errors in the CP estimates $\tilde{p}_{i,j}$ obtained from contact counts, we computed both $\hat{p}_{i,j}$ and $\tilde{p}_{i,j}$ from conformation samples of increasing size M. In particular, we performed MC simulations of chains containing $N = 25$, 50, 100, and 200 beads, and from these simulations we obtained samples containing $M = 10^3$, 10^4, 10^5, and 10^6 uncorrelated conformations. We then compared the estimates $\hat{p}_{i,j}$ and $\tilde{p}_{i,j}$ for each sample size M to the corresponding reference CPs $p^*_{i,j}$, which were computed using contact counts from a large sample of 10^8 conformations.

Plotting $\tilde{p}_{i,j}$ against $p^*_{i,j}$ for all bead pairs with $1 < j-i < N$ in a chain containing 200 beads confirmed the absence of CP estimates $\tilde{p}_{i,j}$ less than $1/M$ and revealed the presence of evident quantization errors for CP values in the range from $1/M$ to $10/M$ (figures 7(*a*)–(*d*)). The same quantization of $\tilde{p}_{i,j}$ was also observed for chains of $N = 25$, 50, 100 beads (data not shown). In contrast, the estimates $\hat{p}_{i,j}$ obtained from fitted EGLD cdfs were all non-zero and were not affected by such quantization errors, even with the relatively small sample size of $M = 10^3$ (figures 7(*e*)–(*h*)). Thus, estimating CPs from fitted EGLD cdfs yields viable estimates even for CPs in the range from $0.1/M$ to $1/M$, where the use of contact counts produces CP estimates $\tilde{p}_{i,j}$ that are either too large or zero.

The plots also indicate that the CPs estimated from EGLD fits generally deviate less from the reference CPs than do the CPs estimated from contact counts, for all tested sample sizes M. However, the progression of plots with increasing sample size M shows that while the CP estimates from contact counts eventually converge to the reference CPs as M increases, the CP estimates from EGLD fits do not appear to converge, indicating the presence of small systematic errors in the latter estimates. These errors are not surprising because the EGLD is used here as an approximation to, rather

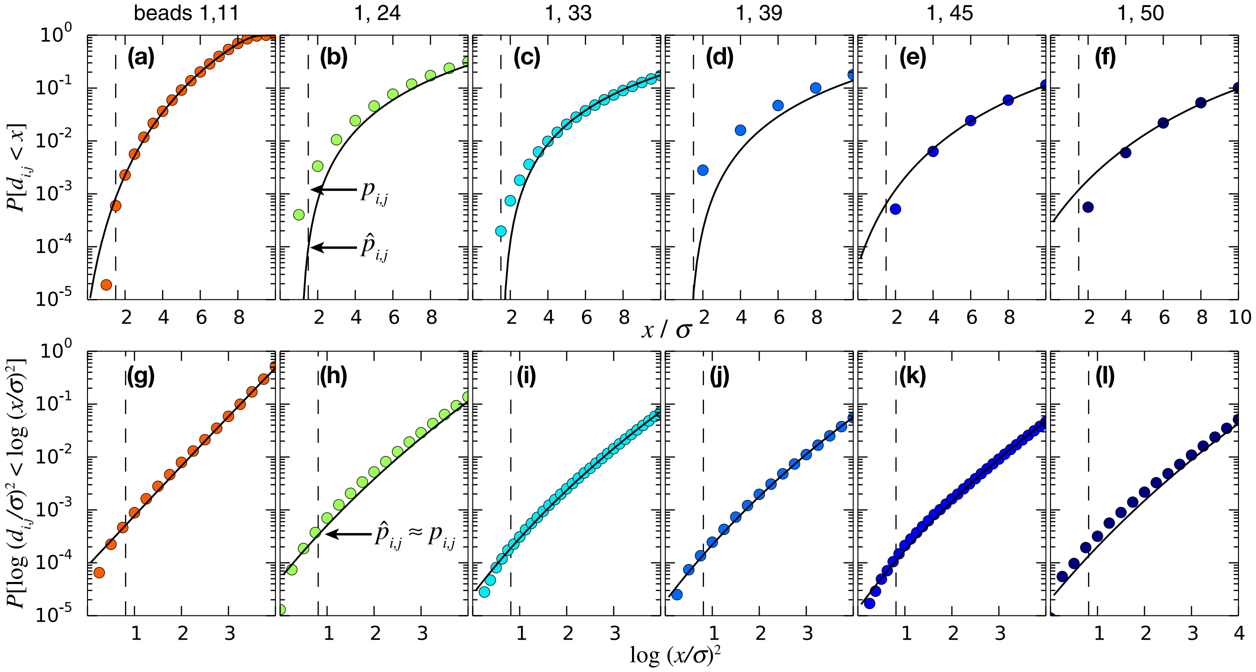

Figure 5. Fitted EGLD cdfs (black lines) overlaid onto the corresponding cumulative histograms (circles) of (*a*)–(*f*) inter-bead distance $d_{i,j}$ and (*g*)–(*l*) log-squared distance $\log(d_{i,j}/\sigma)^2$ for representative bead pairs (i, j) in a chain of 50 beads. Histograms in (*a*)–(*f*) and (*g*)–(*l*) were collected from the same conformation samples used to obtain the estimates $\hat{p}_{i,j}$ shown in figures 4 and 6, respectively. The vertical dashed lines indicate (*a*)–(*f*) the contact distance $d_c = 1.5\sigma$, or (*g*)–(*l*) the log-squared contact distance $\log(d_c/\sigma)^2 \approx 0.81093$, where σ is the bond length. The height of the intersection between the vertical dashed line and the fitted EGLD cdf is equal to the estimate $\hat{p}_{i,j}$. In (*b*) this estimate is at least a factor of 10 smaller than the true CP $p_{i,j}$, whereas in (*h*) the estimate closely approximates the true CP.

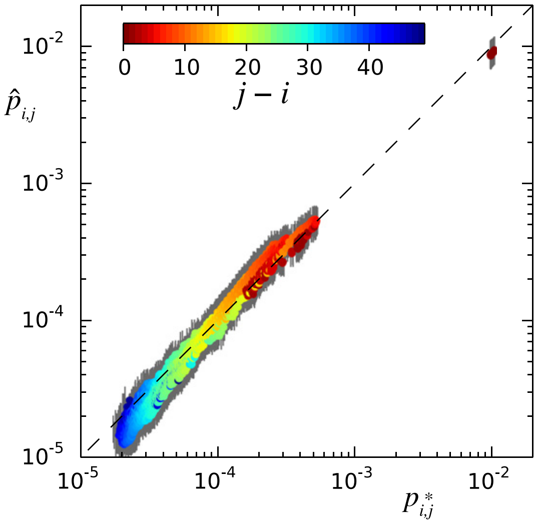

Figure 6. Plot of CP estimates $\hat{p}_{i,j}$ obtained for a chain of 50 beads from EGLD cdfs fitted using sample moments of log-squared distances $\log(d_{i,j}/\sigma)^2$ versus corresponding reference CPs $p^*_{i,j}$. For more details, see figure 4.

than an exact formulation for, the actual distribution of log-squared distances. Nevertheless, figure 7 shows that, with conformation samples of intermediate sizes, say from $M = 10^3$ to $M = 10^5$, fitting EGLD cdfs yields, on average, smaller CP errors than using contact counts.

3.3.1. Systematic errors in the CP estimates. To investigate quantitatively the systematic errors in the CP estimates $\hat{p}_{i,j}$ and $\tilde{p}_{i,j}$, we computed the root mean squared fractional deviation (RMSFD, equation (21)) for both sets of CP estimates. We used the RMSFD because it provides an average measure of fractional rather than absolute errors in the estimated CPs, thus providing equal sensitivity to errors in both large and small CPs.

Plotting the RMSFD of $\tilde{p}_{i,j}$ against sample size M for each bead-chain length confirmed that CPs estimated from contact counts are not biased, because their RMSFD approaches 0 as M increases (circles and black lines in figure 8). Therefore, the estimates $\tilde{p}_{i,j}$ contain only random errors that on average are proportional to $M^{-1/2}$ (figure 8(*d*)). In contrast, the RMSFD trends for $\hat{p}_{i,j}$ appear to approach finite values at sample sizes $M > 10^5$ (diamonds and blue lines in figure 8). These limiting values reflect an average fractional bias in $\hat{p}_{i,j}$. The bias increases with chain length because longer chains have a greater number of bead pairs with small CPs and small CPs have greater fractional bias than large CPs. However, when M is less than a threshold that varies from 10^5 for a chain of 25 beads to 10^6 for a chain of 200 beads, the RMSFD of $\hat{p}_{i,j}$ is consistently lower than the RMSFD of $\tilde{p}_{i,j}$, indicating that the estimates $\hat{p}_{i,j}$ are, on average, more accurate than the estimates $\tilde{p}_{i,j}$ when CPs are estimated from configuration samples of limited size M.

3.3.2. Random errors in the estimated CPs. To quantify the random errors in the estimates $\hat{p}_{i,j}$ and $\tilde{p}_{i,j}$, we computed the mean ratio of standard deviations (MRSD, equation (22)) for increasing sample sizes M. The MRSD compares the sample variance of the CP estimates to the theoretical variance of CPs estimated using contact counts from a sample of M

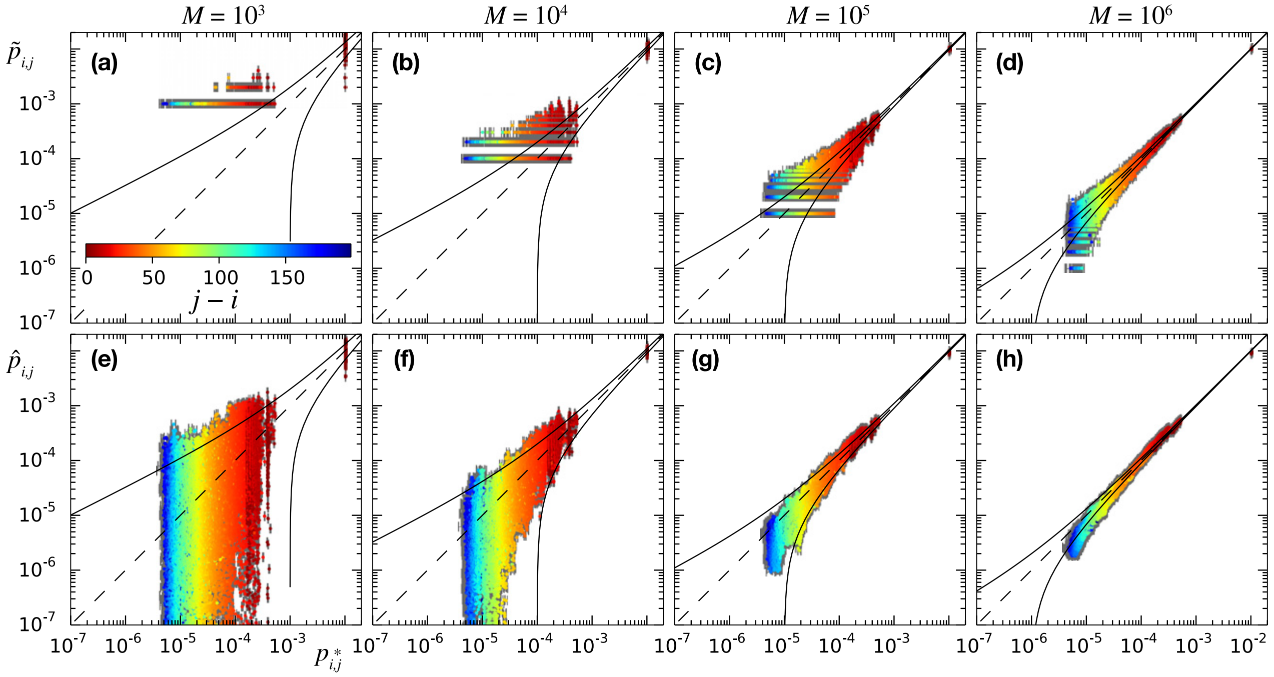

Figure 7. Plots of (*a*)–(*d*) CP estimates $\tilde{p}_{i,j}$ from contact counts and (*e*)–(*h*) CP estimates $\hat{p}_{i,j}$ from EGLD fits against corresponding reference CPs $p^*_{i,j}$, for a chain of $N = 200$ beads and for conformation samples of varying size M. The interpretation of each point is the same as in figures 4 and 6. Each dashed diagonal line corresponds to $y = p^*_{i,j}$, where $y \equiv \tilde{p}_{i,j}$ or $\hat{p}_{i,j}$, and the curves above and below each diagonal line correspond to $y = p^*_{i,j} + \sigma_B$ and $y = p^*_{i,j} - \sigma_B$, respectively, where σ_B is defined in the text and in figure 3. Colors are used as in figure 4.

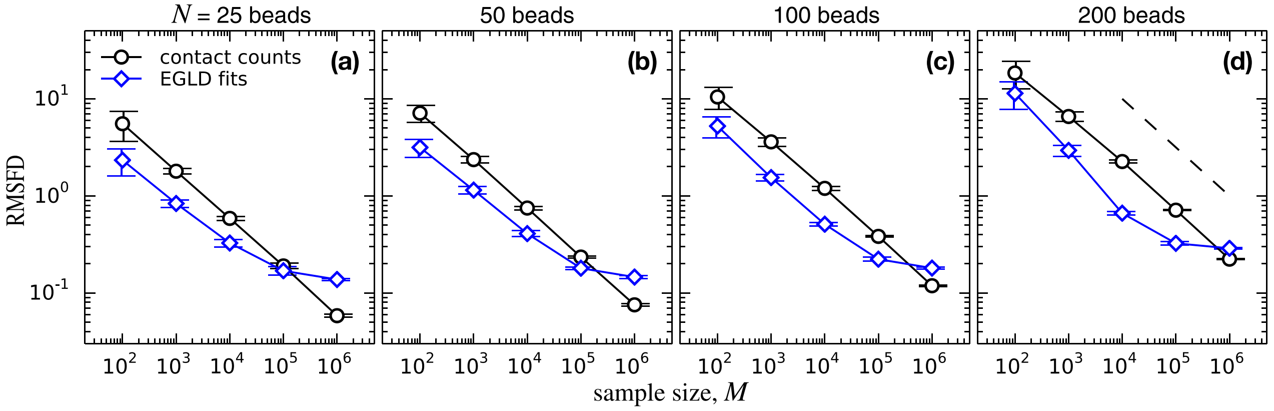

Figure 8. Comparison of average systematic errors in CPs estimated from fitted EGLD cdfs (diamonds and blue lines) to corresponding errors in CPs estimated from contact counts (circles and black lines), as the size M of the conformation samples used to obtain the estimates varies. Each point is the average of 10 values of root mean squared fractional deviation (RMSFD) for CPs estimated from 10 independent conformation samples. Error bars are standard deviations of those 10 RMSFD values. The dashed line in (*d*) indicates the power law $y \sim M^{-1/2}$.

uncorrelated conformations. The latter variance is the variance ${\sigma_B}^2$ of the average of M independent Bernoulli random variables with success probability equal to the reference CP. Similarly to the RMSFD, the MRSD gives equal importance to random errors over all magnitudes of CP estimates.

Plotting the MRSD of $\tilde{p}_{i,j}$ against M for different chain lengths N confirmed that the sample variance of $\tilde{p}_{i,j}$ approaches the theoretical variance of such estimates when M is sufficiently large (circles and black lines in figure 9). The larger average variance of $\tilde{p}_{i,j}$ seen at smaller sample sizes is an artifact due to the tendency of contact counts to yield zero CP estimates when M is small, thus decreasing the number N_p of pairs used to calculate the MRSD. For all tested chain lengths, the MRSD of $\hat{p}_{i,j}$ also appears to approach a limit as M increases (diamonds and blue lines in figure 9). However, at each sample size M, the MRSD of $\hat{p}_{i,j}$ is lower than the MRSD of $\tilde{p}_{i,j}$ by a factor that increases with chain length and is approximately 3 for a chain of 200 beads. These results indicate that the values of $\hat{p}_{i,j}$ are on average more precise than the corresponding values of $\tilde{p}_{i,j}$ estimated from the same sample of chain conformations. In other words, the CP estimates obtained from EGLD fits are less sensitive to variation across data sets than the estimates obtained from contact counts.

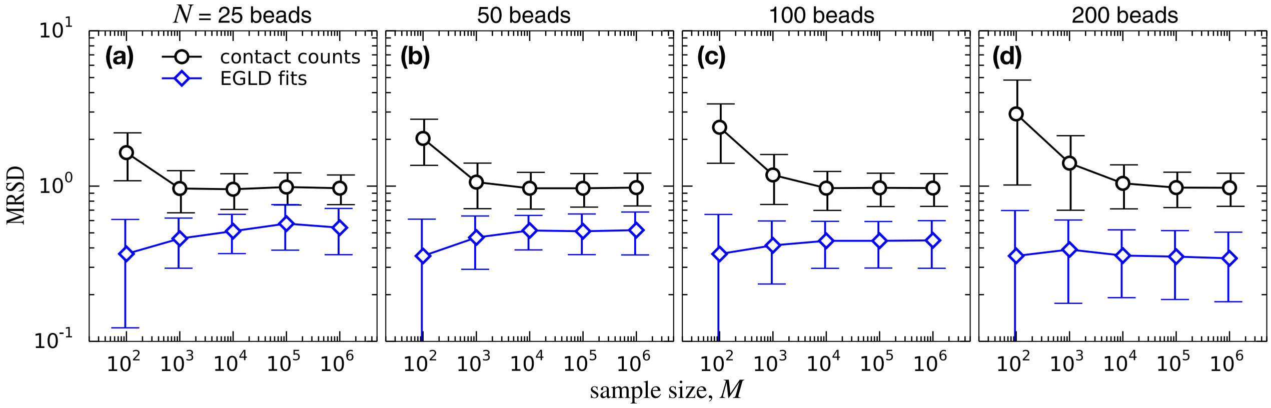

Figure 9. Comparison of average random errors in CPs estimated from EGLD cdfs (diamonds and blue lines) to corresponding errors in CPs estimated from contact counts (circles and black lines), as the size M of the conformation samples used to obtain the estimates varies. Each point is the average over bead pairs of the mean ratio of standard deviations (MRSD), which is defined by equation (22) and effectively compares the sample variance of the CP estimates to the theoretical variance of CP estimates obtained using contact counts from a sample of M uncorrelated conformations. Error bars are standard deviations of the MRSD over all bead pairs considered.

The lower variance of $\hat{p}_{i,j}$ compared to that of $\tilde{p}_{i,j}$ can be explained by noting that the estimation of CPs through fitted EGLD cdfs requires collecting four sample moments from a given conformation sample for each bead pair, whereas the estimation of CPs from contact counts requires collecting one contact count per bead pair. Therefore, CPs estimated from EGLD fits are derived from at least four times as much information as CPs estimated from contact counts. In summary, the plots in figures 8 and 9 provide quantitative evidence that both the accuracy and precision of CPs estimated from contact counts degrade more quickly than those of CPs estimated from fitted EGLD cdfs as the size M of the conformation sample decreases, confirming that the use of EGLD cdfs to estimate CPs is a better choice in applications where confirmation samples have limited size M.

3.3.3. Fractions of bead pairs with sufficiently accurate CPs. The RMSFD and MRSD provide convenient summary measures of average errors in CP estimates relative to reference CPs. It is also of interest to assess the error performance of CP estimates in terms of number of bead pairs with sufficiently accurate CP estimates. The definition of sufficient accuracy necessarily depends on the particular application at hand. For our purpose of comparing the CP estimates $\hat{p}_{i,j}$ and $\tilde{p}_{i,j}$ obtained from EGLD fits and contact counts, respectively, we deem each such estimate to be sufficiently accurate if it does not deviate from the corresponding reference CP $p^*_{i,j}$ by more than the standard deviation $\sigma_{\rm B} = \sqrt{p(1-p)/M}$ of the average of M independent Bernoulli random variables with success probability $p = p^*_{i,j}$.

Table 1 reports the fractions of bead pairs with sufficiently accurate CP estimates obtained from fitted EGLD cdfs and from contact counts, for different chain lengths N, different conformation sample sizes M, and different intervals of CP values. Our tallies indicate that, with sufficiently large conformation samples, i.e. $M = 10^6$, contact counts yield greater numbers of bead pairs with accurate CP estimates than do fitted EGLD cdfs. This outcome is due to the presence of a small bias in the estimates $\hat{p}_{i,j}$, as already seen in figure 7(h). As M increases, $\sigma_{\rm B}$ decreases in proportion to $M^{-1/2}$ and eventually becomes smaller than the bias in $\hat{p}_{i,j}$, at which point $\hat{p}_{i,j}$ is no longer considered to be sufficiently accurate according to our criterion.

Conversely, at intermediate sample sizes of up to 10^5 conformations, much larger numbers of bead pairs with accurate CP can be obtained from EGLD fits than from contact counts. A notable exception occurs for the most frequently interacting bead pairs, those with CPs in the interval $(10^{-3}, 10^{-1}]$, for a chain of 25 beads, or $(10^{-2}, 10^{-1}]$ for chains of 50 or more beads. When using large conformation samples, i.e. $M \geqslant 10^4$ for a chain of 25 beads, and $M \geqslant 10^5$ for a chain of 50 or more beads, fewer of these frequently interacting bead pairs are assigned sufficiently accurate CP estimates through EGLD fits than through contact counts. This discrepancy is again caused by $\sigma_{\rm B}$ becoming smaller than the bias in $\hat{p}_{i,j}$. These results indicate that using fitted EGLD cdfs is preferable to using contact counts when estimating CPs from samples of $M = 10^3$–10^5 conformations and when the CP magnitude is less than roughly $100/M$.

4. Discussion

We have described and tested a computational method for efficiently estimating contact probabilities (CPs) from samples of simulated bead-chain conformations. Our method relies on fitting the extended generalized lambda distribution (EGLD) to inter-bead distance distributions using the method of moments [11]. We have compared the average systematic and random errors in the CPs estimated with this method to corresponding errors in CPs estimated from contact counts. We found that CPs estimated from fitted EGLD cdfs are preferable to CPs estimated from contact counts if the conformation samples used to obtain the estimates are limited in size and if the specific application can tolerate some bias in the larger CP estimates.

One such application is the CP estimation described in [7] as part of an iterative procedure that optimizes certain additional restraints on a bead-chain polymer model of the 30 nm chromatin fiber. In this case, the speed of CP estimation

Table 1. Fractions of bead pairs (i, j), $1 < j - i < N$, with sufficiently accurate CP estimates obtained from contact counts and from fitted EGLD cdfs.[a]

			Conformation sample size M							
			10^3		10^4		10^5		10^6	
N[b]	p^* interval[c]	N_p[d]	counts[e]	fit[f]	counts	fit	counts	fit	counts	fit
25	$(0, 10^{-3}]$	253	5.2 (1.1)	**95.8** (1.1)	70.2 (1.4)	**92.8** (3.0)	67.4 (3.5)	**73.0** (5.2)	**69.1** (2.6)	36.3 (2.1)
	$(10^{-3}, 10^{-2}]$	15	68.7 (9.0)	**78.0** (10.8)	**73.3** (12.3)	68.0 (10.2)	**73.3** (10.7)	7.3 (7.6)	**62.0** (14.6)	
	$(10^{-2}, 10^{-1}]$	8	71.2 (18.6)	**76.2** (10.4)	**72.5** (15.6)	58.8 (14.8)	**75.0** (16.8)	3.8 (5.7)	**68.8** (14.0)	
50	$(0, 10^{-4}]$	112		**95.0** (4.5)	35.6 (4.6)	**97.7** (2.6)	69.2 (4.0)	**94.4** (4.3)	**66.6** (6.5)	47.9 (7.7)
	$(10^{-4}, 10^{-3}]$	1016	1.9 (0.4)	**94.1** (1.4)	67.1 (1.6)	**92.0** (2.2)	68.9 (1.3)	**75.6** (1.9)	**68.2** (1.5)	35.2 (2.4)
	$(10^{-3}, 10^{-2}]$	40	68.3 (6.0)	**78.5** (4.8)	67.2 (9.6)	**71.8** (4.6)	**69.0** (5.8)	4.8 (2.4)	**71.2** (8.5)	
	$(10^{-2}, 10^{-1}]$	8	72.5 (17.5)	**82.5** (11.5)	**77.5** (13.5)	72.5 (13.5)	**71.2** (19.4)	6.2 (6.2)	**77.5** (16.6)	
100	$(0, 10^{-4}]$	2338		**95.6** (1.3)	20.4 (1.0)	**97.9** (0.7)	67.9 (1.0)	**97.3** (0.9)	**69.0** (0.6)	46.5 (2.7)
	$(10^{-4}, 10^{-3}]$	2415	0.9 (0.2)	**94.6** (1.0)	68.0 (1.1)	**92.1** (1.4)	68.5 (0.4)	**71.9** (1.7)	**68.6** (0.9)	29.9 (1.5)
	$(10^{-3}, 10^{-2}]$	90	67.2 (4.7)	**76.7** (3.4)	66.3 (4.4)	**68.6** (4.2)	**67.8** (4.2)	4.1 (1.9)	**67.9** (3.2)	
	$(10^{-2}, 10^{-1}]$	8	76.2 (19.7)	**80.0** (10.0)	65.0 (14.6)	**70.0** (13.9)	**58.8** (12.6)	2.5 (5.0)	**70.0** (15.0)	
200	$(0, 10^{-5}]$	3491		**94.5** (2.5)		**99.1** (0.5)	32.8 (0.9)	**100.0** (0.1)	**66.0** (0.9)	15.3 (2.2)
	$(10^{-5}, 10^{-4}]$	10 885		**95.0** (0.7)	10.2 (0.5)	**98.0** (0.5)	63.5 (0.3)	**98.5** (0.3)	**65.9** (0.5)	47.6 (2.5)
	$(10^{-4}, 10^{-3}]$	5127	0.4 (0.1)	**93.6** (1.0)	65.4 (0.7)	**90.4** (1.1)	65.8 (0.5)	**69.4** (1.7)	**66.2** (0.8)	26.3 (1.1)
	$(10^{-3}, 10^{-2}]$	1.0	63.3 (3.3)	**74.0** (2.7)	**66.6** (2.4)	66.6 (2.4)	**65.0** (2.6)	5.1 (1.3)	**65.9** (2.4)	
	$(10^{-2}, 10^{-1}]$	8	70.0 (17.0)	**76.2** (21.3)	58.8 (22.4)	**70.0** (12.7)	**70.0** (17.9)	5.0 (6.1)	**62.5** (15.8)	

[a] Fractions are reported as percentages. Values in parentheses are standard deviations of fractions computed from 10 independent conformation samples. Omitted values are equal to zero.

[b] Number of beads in the chain.

[c] Each row in the table reports fractions for bead-pairs whose reference CPs $p^*_{i,j}$ are within the specified interval.

[d] Number of bead pairs (i, j), with $1 < j - i < N$, having reference CPs within the interval indicated in the previous column. Note that bead pairs (i, j) with $j - i = 1$ are of no interest, because $d_{i,j} < d_c$ for these beads, and are therefore not counted.

[e] Fraction of bead pairs with sufficiently accurate CP estimates from contact counts.

[f] Fraction of bead pairs with sufficiently accurate CP estimates from fitted EGLD cdfs. The larger of the two fractions reported for contact counts and fitted EGLD cdfs is highlighted in bold to facilitate comparison.

is more important than CP accuracy, because the iterative procedure employed to adjust the additional restraints on the polymer model requires tens of iterations to achieve good convergence, but is also resilient to errors in the estimated CPs. Our results indicate that, using fitted EGLD cdfs instead of contact counts, such procedure could achieve up to a tenfold reduction in the number of conformations needed to estimated CPs with a given average error quantified by the RMSFD (figure 8).

We found, however, that as CP magnitude increases, as sample size M increases, or as chain length N decreases, estimates obtained from contact counts become sufficiently accurate for more bead pairs than do estimates obtained from fitted EGLD cdfs (table 1). Thus, applications requiring sufficiently accurate estimates also for relatively large CPs may benefit from a hybrid approach that estimates large CPs from contact counts and small CPs from fitted EGLD cdfs. Such a hybrid approach might represent an interesting direction for future work.

Although the functional forms of the EGLD offer great flexibility in representing a variety of probability distributions, efficiently fitting such functional forms with the method of moments is complicated by the complexity of the resulting mathematical expressions and by the lack of global convergence in the space of shape parameters. Following the suggestions in [11], we addressed this challenge by implementing a table look-up of initial solutions that are known a priori to yield correct solutions for the shape parameters when using the Newton–Raphson method.

On the other hand, several alternatives to the method of moments for fitting the GLD, which is one component of the EGLD, have been proposed [20, 22, 23, 33–36] and ready-to-use implementations of these methods are available [24, 37, 38]. However, such methods require entire data sets, rather than just the moments, and appear to be more computationally demanding than an approach based on table look-up of a precomputed initial solution followed by a single root-finding iteration. Moreover, the latter approach is equally suitable to determine the shape parameters of both the GLD and the GBD. The details of our EGLD fitting procedure and its comparison to other methods will be provided in a separate publication.

The present work did not include any evaluation of the 'quality of fit' for the EGLD in the context of inter-bead distance distributions. Although such evaluation represents an interesting topic for future studies, it was not essential in addressing the main concern of the present study, which is the efficient estimation of reasonably accurate and precise CPs from simulated bead-chain conformations. In fact, we found that this requirement can be met by ensuring that a good match between the fitted EGLD cdf and the actual cdf is achieved for inter-bead distances on the order of the contact distance d_c. Therefore, maximizing a goodness-of-fit measure based on a standard test, such as a chi-square test or the Kolmogorov–Smirnov test, would not necessarily guarantee the optimal estimation of CPs from the fitted EGLD cdf. Instead, our direct comparison of estimated CPs to reference CPs provides a better criterion for assessing the ability of the fitted EGLD cdfs to yield reliable CP estimates.

Future studies could investigate alternative methods for fitting the EGLD and assess their ability to produce CPs with lower systematic and random errors, while minimizing computation time and sample size. Another interesting direction for future work could be to fit other families of probability distributions, either generic [39] or theoretical [14, 15], and to determine whether the CP estimates obtained from such distributions are more reliable and easier to compute than those obtained from fitted EGLD cdfs. Although the method we presented is aimed at accelerating efforts to elucidate the spatial organization of chromatin, the ability to estimate CPs efficiently from simulations of polymer chains may be also beneficial to research in topics as diverse as globule formation in multiblock copolymers [40] or reaction kinetics in macromolecules with reactive groups [41].

Acknowledgments

This work was partially supported by the American Cancer Society [Instructional Research Grant 70-002], and by the National Science Foundation [Research Grant 1200850].

References

[1] Harmston N and Lenhard B 2013 Chromatin and epigenetic features of long-range gene regulation *Nucl. Acids Res.* **41** 7185–99

[2] De Wit E and De Laat W 2012 A decade of 3C technologies: insights into nuclear organization *Genes Dev.* **26** 11–24

[3] Belton J-M, McCord R P, Gibcus J H, Naumova N, Zhan Y and Dekker J 2012 Hi–C: a comprehensive technique to capture the conformation of genomes *Methods* **58** 268–76

[4] Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A and Cavalli G 2012 3D folding and functional organization principles of the drosophila genome *Cell* **148** 458–72

[5] Kalhor R, Tjong H, Jayathilaka N, Alber F and Chen L 2012 Genome architectures revealed by tethered chromosome conformation capture and population-based modeling *Nat. Biotechnol.* **30** 90–8

[6] Dekker J, Marti-Renom M A and Mirny L A 2013 Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data *Nat. Rev. Genet.* **14** 390–403

[7] Meluzzi D and Arya G 2013 Recovering ensembles of chromatin conformations from contact probabilities *Nucl. Acids Res.* **41** 63–75

[8] Giorgetti L, Galupa R, Nora E P, Piolot T, Lam F, Dekker J, Tiana G and Heard E 2014 Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription *Cell* **157** 950–63

[9] Rosa A, Becker N B and Everaers R 2010 Looping probabilities in model interphase chromosomes *Biophys. J.* **98** 2410–9

[10] Karian Z A, Dudewicz E J and Mcdonald P 1996 The extended generalized lambda distribution system for fitting distributions to data: history, completion of theory, tables, applications, the 'final word' on moment fits *Commun. Stat.—Simul. Comput.* **25** 611–42

[11] Karian Z A and Dudewicz E J 2000 *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods* (New York: Chapman and Hall)

[12] Fytas N G and Theodorakis P E 2011 Analysis of the static properties of cluster formations in symmetric linear multiblock copolymers *J. Phys. Condens. Matter* **23** 235106

[13] Theodorakis P E and Fytas N G 2012 A study for the static properties of symmetric linear multiblock copolymers under poor solvent conditions *J. Chem. Phys.* **136** 094902

[14] Oono Y and Ohta T 1981 The distribution function for internal distances in a self-avoiding polymer chain *Phys. Lett.* A **85** 480–2

[15] Wittkop M, Kreitmeier S and Goritz D 1996 The distribution function of internal distances of a single polymer chain with excluded volume in two and three dimensions: a Monte Carlo study *J. Chem. Phys.* **104** 351–8

[16] Ramberg J S and Schmeiser B W 1974 An approximate method for generating asymmetric random variables *Commun. ACM* **17** 78–82

[17] Zwillinger D and Kokoska S 1999 *CRC Standard Probability and Statistics Tables and Formulae* (Boca Raton, FL: CRC Press)

[18] Karian Z A and Dudewicz E J 1999 Fitting the generalized lambda distribution to data: a method based on percentiles *Commun. Stat.—Simul. Comput.* **28** 793–819

[19] Asquith W H 2007 L-moments and TL-moments of the generalized lambda distribution *Comput. Stat. Data Anal.* **51** 4484–96

[20] King R A R and MacGillivray H L 1999 A starship estimation method for the generalized lambda distributions *Aust. N. Z. J. Stat.* **41** 353–74

[21] Su S 2005 A discretized approach to flexibly fit generalized lambda distributions to data *J. Mod. Appl. Stat. Methods* **4** 408–24

[22] Su S 2007 Fitting single and mixture of generalized lambda distributions to data via discretized and maximum likelihood methods: GLDEX in R *J. Stat. Softw.* **21** 1–17

[23] Su S 2007 Numerical maximum log likelihood estimation for generalized lambda distributions *Comput. Stat. Data Anal.* **51** 3983–98

[24] Su S 2010 Fitting GLD to data using gldex 1.0.4 in R *Handbook of Fitting Statistical Distributions with R* ed Z A Karian and E J Dudewicz (London: Chapman and Hall) chapter 15, pp 585–608

[25] R Core Team 2013 *R: A Language, Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing)

[26] Frenkel D and Smit B 2002 *Understanding Molecular Simulation* 2nd edn (San Diego, CA: Academic)

[27] Arya G 2009 Energetic and entropic forces governing the attraction between polyelectrolyte-grafted colloids *J. Phys. Chem.* B **113** 15760–70

[28] Arya G 2010 Chain stiffness and attachment-dependent attraction between polyelectrolyte-grafted colloids *J. Phys. Chem.* B **114** 15886–96

[29] Rosa A and Everaers R 2008 Structure and dynamics of interphase chromosomes *PLoS Comput. Biol.* **4** e1000153

[30] Pettersen E F, Goddard T D, Huang C C, Couch G S, Greenblatt D M, Meng E C and Ferrin T E 2004 UCSF chimera—a visualization system for exploratory research and analysis *J. Comput. Chem.* **25** 1605–12

[31] Oliphant T E 2007 Python for scientific computing *Comput. Sci. Eng.* **9** 10–20

[32] Hunter J D 2007 Matplotlib: a 2D graphics environment *Comput. Sci. Eng.* **9** 90–5

[33] Karvanen J and Nuutinen A 2008 Characterizing the generalized lambda distribution by L-moments *Comput. Stat. Data Anal.* **52** 1971–83

[34] Karian Z A and Dudewicz E J 2003 Comparison of GLD fitting methods: superiority of percentile fits to moments in L2 norm *J. Iran. Stat. Soc.* **2** 171–87

[35] Fournier B, Rupin N, Bigerelle M, Najjar D, Iost A and Wilcox R 2007 Estimating the parameters of a generalized lambda distribution *Comput. Stat. Data Anal.* **51** 2813–35

[36] Lakhany A and Mausser H 2000 Estimating the parameters of the generalized lambda distribution *Algo Res. Q.* **3** 47–58

[37] King R 2014 GLD: estimation and use of the generalised (tukey) lambda distribution. R package version 2.2.1

[38] Upadhyay R and Ezekoye O 2012 libMoM: a library for stochastic simulations in engineering using statistical moments *Eng. Comput.* **28** 83–94

[39] Karian Z A and Dudewicz E J (ed) 2011 *Handbook of fitting statistical distributions with R* (New York: Chapman and Hall)

[40] Rissanou A N, Tzeli D S, Anastasiadis S H and Bitsanis I A 2014 Collapse transitions in thermosensitive multi-block copolymers: a Monte Carlo study *J. Chem. Phys.* **140** 204904

[41] Friedman B and O'Shaughnessy B 1994 Scaling and universality in polymer reaction kinetics *Int. J. Mod. Phys.* B **08** 2555–91