Published online 11 November 2012 *Nucleic Acids Research, 2013, Vol. 41, No. 1* **63–75**
doi:10.1093/nar/gks1029

Recovering ensembles of chromatin conformations from contact probabilities

Dario Meluzzi[1,2] and Gaurav Arya[1,*]

[1]Department of NanoEngineering and [2]Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Received July 26, 2012; Revised October 2, 2012; Accepted October 4, 2012

ABSTRACT

The 3D higher order organization of chromatin within the nucleus of eukaryotic cells has so far remained elusive. A wealth of relevant information, however, is increasingly becoming available from chromosome conformation capture (3C) and related experimental techniques, which measure the probabilities of contact between large numbers of genomic sites in fixed cells. Such contact probabilities (CPs) can in principle be used to deduce the 3D spatial organization of chromatin. Here, we propose a computational method to recover an ensemble of chromatin conformations consistent with a set of given CPs. Compared with existing alternatives, this method does not require conversion of CPs to mean spatial distances. Instead, we estimate CPs by simulating a physically realistic, bead-chain polymer model of the 30-nm chromatin fiber. We then use an approach from adaptive filter theory to iteratively adjust the parameters of this polymer model until the estimated CPs match the given CPs. We have validated this method against reference data sets obtained from simulations of test systems with up to 45 beads and 4 loops. With additional testing against experiments and with further algorithmic refinements, our approach could become a valuable tool for researchers examining the higher order organization of chromatin.

INTRODUCTION

Eukaryotic cells need to accommodate their long genomic DNA within a relatively small nucleus. This remarkable feat is accomplished through several levels of 3D spatial organization (1). The first level consists of wrapping the DNA duplex around octamers of histone proteins to form nucleosomes. The resulting string of nucleosomes is then folded into a thicker fiber known as chromatin. Subsequent levels of folding ultimately lead to the territorial arrangement of chromosomes within the nucleus. These additional levels of folding, referred to as higher order organization of chromatin, are not only essential for efficient DNA packaging but are also believed to play a role in several other biological processes. For example, the formation of chromatin loops facilitates interactions between distant portions of DNA and these interactions are essential for regulating transcription and recombination (2–5). Also, the transcriptional activity of genes tends to be inversely correlated with the spatial density of chromatin fibers (6–9). Furthermore, growing evidence suggests that spatially proximal regions of the genome are more likely to be functionally correlated, leading to the concepts of ‘factories’, ‘globules’ and ‘territories’ (10–14). Unfortunately, owing to the limitations of current experimental methods in visualizing chromatin *in vivo*, the 3D higher order organization of chromatin is not well understood. In particular, state-of-the-art microscopy approaches, such as fluorescence *in situ* hybridization (FISH) (15) and super-resolution fluorescence microscopy (16), do not simultaneously provide the spatial resolutions and the measurement throughput necessary to discern and locate individual chromatin fibers within the nucleus.

During the past decade, however, increasingly higher resolution and throughput have been achieved by a number of sophisticated experimental techniques—including 4C (17,18), 5C (19), GCC (20) and Hi-C (21)—that are based on the original method of chromosome conformation capture (3C) (22). These techniques do not directly capture the 3D spatial organization of chromatin. Instead, they measure the frequency of interactions between different fragments of genomic DNA in fixed cells (23). To detect such interactions, spatially proximal segments of DNA are covalently cross-linked by treating millions of intact nuclei with chemical agents, such as formaldehyde. The DNA is then cleaved into small fragments by digestion with appropriate restriction enzymes. Next, the resulting pairs of cross-linked fragments are

*To whom correspondence should be addressed. Tel: +1 858 822 5542; Fax: +1 858 534 9553; Email: garya@ucsd.edu enzymatically ligated and the cross-links are chemically removed. Finally, the ligation products are amplified by polymerase chain reaction and sequenced by high-throughput methods. Analysis of the sequences allows one to identify the pairs of fragments that were originally cross-linked. Counting the number of times that each pair was identified from the sequences yields a 2D map of contact probabilities (CPs) for the examined pairs of fragments.

Although CP maps provide abundant information to help researchers infer the higher order organization of chromatin through theoretical and computational models (24,25), such a task is rather challenging. To tackle this problem, several approaches have already been proposed. Dekker *et al.* (22) presented the first such approach to deduce a coarse 3D structure for the 320-kb chromosome III in NKY2997 cells. To obtain the structure, 78 CPs were measured by 3C and converted to spatial distances through a theoretical expression for worm-like chains (26). The resulting distances were presumably used to solve a molecular distance geometry problem. Later, Fraser *et al.* (27) assumed the inverse proportionality relation $d \propto 1/p$ to calculate spatial distances d from hundreds of CPs p, obtained by 5C experiments on the *HoxA* gene cluster in THP-1 leukemia cells. The resulting distances d were then used as targets to optimize a piecewise linear curve representing the gene cluster under study. The same relation $d \propto 1/p$ was used by Duan *et al.* (28) to infer the 3D structure of the budding yeast genome from over 65 000 CPs obtained by 4C. In addition, they modeled chromatin as a chain of beads, each representing 10 kb of DNA, and defined various constraints to enforce known geometric and topological features of yeast chromatin. Nonlinear constrained optimization methods were then used to find an optimal structure. Another full genome, that of fission yeast, was studied by Tanizawa *et al.* (29) using a Hi-C variant with enrichment of ligation products. To determine the 3D structure of this genome, the authors used a bead-chain model and a method similar to that of (28). This time, however, spatial distances were calculated from CPs through a calibration curve obtained by fitting a double exponential decay function to distance measurements obtained by FISH.

A bead-chain model of chromatin was also employed by Baù *et al.* (12), who used 5C to analyze the 500-kb ENm008 domain (α-globin gene) on human chromosome 16 in K562 cells and in GM12878 cells. In this case, though, each bead represented a DNA restriction fragment, with bead radius proportional to fragment length. The beads interacted through harmonic restraints with strengths and equilibrium distances derived from experimental CPs. A combination of optimization and clustering algorithms was then used to determine a conformation ensemble and corresponding centroid structure for the ENm008 domain in each cell type. Again seeking conformation ensembles, Rousseau *et al.* (30) used a probabilistic approach to analyze 5C data on the 142-kb *HoxA* cluster in THP-1 and HB-1119 cell lines, and Hi-C data from (21) on the 88.4-Mbp long arm of human chromosome 14. In particular, they applied a Markov chain Monte Carlo sampling method to generate ensembles of structures consistent with a posterior distribution of spatial distances between restriction fragment midpoints, where the distances were again obtained from experimental CPs by assuming an inverse power law relation. Another effort to obtain chromatin conformation ensembles, but without using a distance-CP relationship, was recently reported by Gehlen *et al.* (31). To generate such ensembles for the entire *S. cerevisiae* genome, the authors performed multiple molecular dynamics simulations of a bead-chain polymer model and included within each simulation a randomly selected subset of intra- and inter-chromosomal interactions experimentally determined through GCC.

Although the above computational approaches are remarkable in their ability to handle large numbers of interacting fragments, almost all of them rely on converting the measured CPs to spatial distances between interacting fragments. Such conversion is achieved by assuming a functional relation that describes the behavior of free linear chains. For example, polymer theory predicts that $p \propto d^{-3}$ for ideal random walk chains (32), whereas more elaborate relations have been derived for worm-like chains (33). These relations, however, may not be valid for polymers subjected to looping and other external constraints. Also, several of the above approaches ignore the mechanical properties of the chromatin fiber or determine only a single average structure from a given set of CPs, which are in fact the result of cross-linking events over an ensemble of chromatin conformations sampled from millions of cells. Finally, none of the above studies validate their proposed computational methods against known chromatin conformation ensembles.

Here, we describe and validate a computational approach to obtain ensembles of chromatin conformation consistent with a given set of reference CPs. This approach does not require assuming a functional relation between spatial distances and CPs. Instead, we estimate new CPs by simulating a coarse-grained polymer model that approximates the physical behavior of a 30-nm chromatin fiber. We then iteratively adjust the parameters of this polymer model until a good match is achieved between the CPs estimated from the simulations and those in the given reference set. The result is an ‘optimal’ ensemble of conformations that is most consistent with the given reference CPs. Our initial validation of this approach against several simulated test systems produced good agreement of average spatial properties between reference and recovered conformation ensembles.

MATERIALS AND METHODS

Our goal is to generate an ensemble of conformations consistent with a given set of probabilities of contact between different segments of a chromatin fiber. To achieve this goal, we propose a computational method that consists of three main components (Figure 1): (i) a coarse-grained polymer model approximating the physical properties of chromatin; (ii) a procedure to generate an ensemble of conformations for the polymer model and

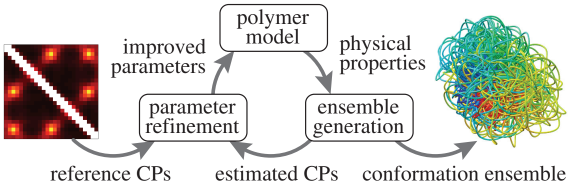

Figure 1. Main components of the proposed computational approach to recover a conformation ensemble from a given set of reference CPs.

(iii) a procedure to refine the parameters of the polymer model in such a way that the generated conformation ensemble is consistent with the given set of CPs.

Coarse-grained polymer model of chromatin

We assume that chromatin exists as a fiber with an average diameter of 30 nm, and that the conformation of this fiber is determined primarily by its stretching resistance, bending stiffness and excluded volume. To approximate these physical properties, we use a bead-chain model, with each bead representing a chromatin segment of 3–6 kb (34). A similar model was proposed by Rosa *et al.* (35) and was recently used to simulate the entire genome of budding yeast (36). Following Baù *et al.* (12), we also assume that the chromatin fiber is subjected to unknown external constraints, e.g. due to looping interactions or confinement, and that the average effects of these constraints can be approximated by additional harmonic restraints connecting particular beads in the chain (Figure 2a).

Thus, the potential energy U of the bead chain can be expressed as the sum of four terms,

$$U = U_{\mathrm{bond}} + U_{\mathrm{bend}} + U_{\mathrm{excl}} + U_{\mathrm{rest}}. \tag{1}$$

The first term U_{bond} accounts for the chain's resistance to stretching and results from connecting adjacent beads with harmonic springs,

$$U_{\mathrm{bond}} = \sum_{i=1}^{N-1} \frac{1}{2} k_{\mathrm{s}} \left(d_{i,i+1} - d_0\right)^2, \tag{2}$$

where $d_{i,j} = \left|\mathbf{r}_j - \mathbf{r}_i\right|$ is the distance between beads i and j, N is the number of beads in the chain, k_{s} is the spring constant, $\mathbf{r}_i$ is the position vector for bead i, $d_0 = \sigma$ is the equilibrium bond length and $\sigma = 30\,\mathrm{nm}$ is the unit of length used in our simulations (Table 1).

The second potential energy term U_{bend} accounts for the chain's resistance to bending and results from subjecting each triplet of adjacent beads to a harmonic bending potential (37),

$$U_{\mathrm{bend}} = \sum_{i=1}^{N-2} \frac{1}{2} k_\theta \theta_i^2, \tag{3}$$

where θ_i is the angle between the displacement vectors $\mathbf{r}_{i+1} - \mathbf{r}_i$ and $\mathbf{r}_{i+2} - \mathbf{r}_{i+1}$, $k_\theta = k_{\mathrm{B}} T L_{\mathrm{p}} / \sigma$ is an angular 'spring constant', L_{p} is the persistence length of the chain (38), k_{B} is the Boltzmann constant and T is the absolute temperature.

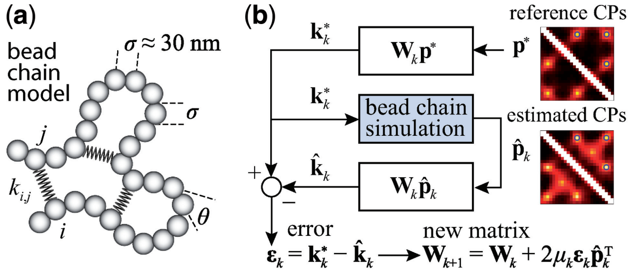
(a)
bead chain model
$\sigma \approx 30$ nm
σ
j
$k_{i,j}$
i
θ
(b)
reference CPs
$\mathbf{k}_k^*$
$\mathbf{W}_k\mathbf{p}^*$
$\mathbf{p}^*$
$\mathbf{k}_k^*$
bead chain simulation
estimated CPs
$\hat{\mathbf{p}}_k$
$\hat{\mathbf{k}}_k$
$\mathbf{W}_k\hat{\mathbf{p}}_k$
+
−
error
new matrix
$\boldsymbol{\varepsilon}_k = \mathbf{k}_k^* - \hat{\mathbf{k}}_k \longrightarrow \mathbf{W}_{k+1} = \mathbf{W}_k + 2\mu_k \boldsymbol{\varepsilon}_k \hat{\mathbf{p}}_k^{\mathrm{T}}$

Figure 2. Schematic representations of (**a**) restrained bead-chain polymer model used for BD simulations of a 30-nm chromatin fiber subjected to looping constraints and (**b**) application of the LMS algorithm to the optimization of the parameters in the general linear model (Equation 9) used to predict restraint spring constants from reference CPs.

The third potential energy term, U_{excl}, accounts for the excluded volume, or effective thickness, of the chain and is treated using the repulsive part of the Lennard–Jones potential,

$$U_{\mathrm{excl}} = \sum_{2 \le i+1 < j \le N} 4\epsilon\Theta\left(2^{1/6} - \frac{d_{i,j}}{\sigma}\right) \times \left[\left(\frac{\sigma}{d_{i,j}}\right)^{12} - \left(\frac{\sigma}{d_{i,j}}\right)^{6} + \frac{1}{4}\right], \tag{4}$$

where $\epsilon = k_{\mathrm{B}} T$ is the unit of energy in our simulations, $\sigma = 30\,\mathrm{nm}$ is the effective thickness of the fiber (Table 1) and $\Theta(x)$ is the Heaviside step function, which equals 1 when $x > 0$ and 0 otherwise.

The last potential energy term, U_{rest}, accounts for the presence of external forces or constraints that affect the shape of the chromatin fiber as well as the probability of contact between different segments of the fiber. In particular, we assume that the average effects of these constraints can be reproduced reasonably well by including a sufficient number of harmonic restraints that connect a subset of the beads in the chain,

$$U_{\mathrm{rest}} = \sum_{(i,j) \in R} \frac{1}{2} k_{i,j} \left[d_{i,j} - d_{i,j}^0\right]^2, \tag{5}$$

where R is the set of pairs of beads connected by harmonic restraints and $k_{i,j}$ ($d_{i,j}^0$) is the spring constant (equilibrium distance) for the restraint connecting beads i and j. The actual members (i, j) of the set R and the corresponding values of $k_{i,j}$ and $d_{i,j}^0$ are adjustable parameters in this model of restrained chromatin.

Generation of conformation ensembles

To obtain optimal values for these adjustable parameters, we compare a reference set of probabilities of contact between the beads in the chain with a set of corresponding probabilities estimated from an ensemble comprising a large number of bead-chain conformations.

Simulations of bead chain

To obtain such an ensemble, we start by minimizing the potential energy of an initial conformation. To this end, we use the Polak–Ribiere modification of the conjugate gradient algorithm (39). Next, we equilibrate the

Table 1. Parameter values used to simulate the restrained bead-chain polymer model of chromatin and to provide a physically realistic approximation of the mechanical properties of chromatin, as currently known from experiments

Parameter	Symbol	Reduced units	SI units
Thermal energy[a]	$k_B T$	1.0	4.1×10^{-21} J
Bead mass[b]	m	1.0	7.0×10^{-21} kg
Lennard–Jones size parameter	σ	1.0	30 nm
Lennard–Jones energy parameter	ϵ	$1.0\, k_B T$	4.1×10^{-21} J
Bead separation	d_0	$1.0\,\sigma$	30 nm
Contact distance[c]	d_c	$1.5\,\sigma$	45 nm
Bond spring constant[d]	k_s	$500\, k_B T/\sigma^2$	2.3×10^{-3} J m^{-2}
Persistence length[e]	L_p	$4.0\,\sigma$	120 nm
Bending energy constant	k_θ	$4.0\, k_B T/\text{rad}^2$	1.7×10^{-20} J rad^{-2}
Time step/damping constant[f]	$\Delta t/\gamma$	$3.3 \times 10^{-4}\, \sigma^2 m/k_B T$	5.1×10^{-19} s^2

[a] Energy per bead per degree of freedom at $T = 300$ K.
[b] Representative value based on the experimental measurement of 23.3 MDa for a 15.5-kb fragment of 30-nm chromatin upstream of the chicken β-globin locus (42).
[c] Following Rosa *et al.* (35), equivalent to assuming that contacts between chromatin fibers are mediated by proteins of 15-nm diameter.
[d] From experiments, the stretching modulus is $d_0 k_s \approx$ 5–150 pN (43), hence k_s ranges from 1.7×10^{-4} to 2.5×10^{-2} Jm^{-2}.
[e] From experiments, $L_p \approx 30-200$ nm (43).
[f] To maximize conformation sampling efficiency, we used the largest value of $\Delta t/\gamma$ found to maintain stability of the BD simulations. A lower bound for Δt can be estimated by considering a chromatin sphere of radius r = 15 nm and using $\gamma = 6\pi\eta r/m$ with the viscosity of water η = 890 µPa s at 25°C and 1 bar (44). Then, $\Delta t \approx$ 18 ns.

energy-minimized chain by performing 10^6 steps of Brownian dynamics (BD) simulation. We then perform an additional BD simulation during which we collect one conformation every 100 integration steps. The set of conformations collected from a single simulation trajectory constitute a conformation ensemble.

To perform the BD simulations, we apply a second-order algorithm (40,41), which we simplify to neglect the effects of hydrodynamic interactions. Specifically, for each bead i, we calculate a tentative new position at time $t+\Delta t$ using the position $\mathbf{r}_i(t)$ of the bead and the force $\mathbf{f}_i(t)$ on the bead at time t,

$$\widetilde{\mathbf{r}}_i(t+\Delta t) = \mathbf{r}_i(t) + \frac{\Delta t}{\gamma m}\mathbf{f}_i(t) + \sqrt{\frac{2k_B T \Delta t}{\gamma m}}\mathbf{N}(t), \qquad \textbf{(6)}$$

where $\gamma = k_B T/D_s m$ is the damping constant, D_s is the self-diffusion coefficient, Δt is the integration time step, m is the mass of each bead and $\mathbf{N}(t)$ is a random displacement vector whose components are normally distributed with mean 0 and variance 1. Next, we use the tentative bead positions to calculate a tentative new force $\widetilde{\mathbf{f}}_i(t+\Delta t)$ for each bead i. Then, for each bead i, we calculate a more accurate position using the tentative position and tentative force at time $t+\Delta t$ and the force at time t,

$$\mathbf{r}_i(t+\Delta t) = \widetilde{\mathbf{r}}_i(t+\Delta t) + \frac{\Delta t}{2\gamma m}[-\mathbf{f}_i(t) + \widetilde{\mathbf{f}}_i(t+\Delta t)]. \qquad \textbf{(7)}$$

Finally, these latter bead positions are used to calculate more accurate forces $\mathbf{f}_i(t+\Delta t)$ at time $t+\Delta t$.

Estimation of CPs

To estimate the bead CPs from an ensemble of bead chain conformations, we analyze each member of the ensemble and check for the occurrence of contacts within all possible pairs of beads in the chain. Following Rosa *et al.* (35), a contact between two beads is defined to occur whenever the distance between the beads is less than a predefined 'contact' distance, d_c (Table 1). Hence, we estimate the probability of contact $p_{i,j}$ between beads i and j by calculating the proportion of conformations in which a contact occurs between those beads,

$$\hat{p}_{i,j} = \frac{1}{N_c}\sum_{l=1}^{N_c} \Theta(d_c - d^l_{i,j}), \qquad \textbf{(8)}$$

where N_c is the total number of conformations in the ensemble and $d^l_{i,j}$ is the distance between beads i and j in conformation l of the ensemble.

Refinement of model parameters

In this work, we assume that a set of reference CPs, denoted $p^*_{i,j}$, is available for $1 \le i < j \le N$, and the problem is to find an ensemble of bead-chain conformations consistent with those CPs. To this end, we need to optimize the adjustable parameters of the bead-chain model so that simulating such a model yields a conformation ensemble whose estimated CPs $\hat{p}_{i,j}$ match as closely as possible the corresponding reference CPs $p^*_{i,j}$ for $1 \le i < j \le N$.

The adjustable parameters to be optimized are the pairs of indexes $(i,j) \in R$ and the values of $k_{i,j}$ and $d^0_{i,j}$ for each (i, j). To begin to tackle this complex problem, we choose to reduce the number of adjustable parameters by fixing the members (i, j) of the set R at the start of the optimization procedure and by using zero as the equilibrium distance for the harmonic restraints, i.e. $d^0_{i,j} = 0, \forall\ (i,j) \in R$. Although $d^0_{i,j} = 0$, excluded volume interactions (Equation 4) prevent beads from overlapping.

Placement of harmonic restraints

To determine the pairs $(i,j) \in R$ of beads that must be connected by harmonic restraints, we analyze the given set of reference CPs $p^*_{i,j}$, for $1 \le i < j \le N$, by using a peak detection algorithm. Specifically, we construct a smooth surface $z = g(x,y)$ such that $g(i,j) \approx p^*_{i,j}$ for

$1 \leq i < j \leq N$. Each peak on this CP 'surface' corresponds to a pair of beads that interact more frequently than their neighbors. Thus, we find the pair of integers (i, j) closest to the location (x_p,y_p) of each peak in the CP surface, and we add (i, j) to the set R of pairs of beads connected by harmonic restraints. To find the location (x_p,y_p) of each peak, we slice the surface at every point (i,j), for $1 \leq i < j \leq N$, using the four vertical planes $x = i$, $y = j$, $x+y = i+j$ and $x - y = i - j$. Next, we find the local maxima of the curve generated by each slice. If the curves on all four slices have a local maximum close to $(x_p,y_p) \approx (i,j)$, then we deem (x_p,y_p) to be the location of a peak on the CP surface.

Optimization of restraint spring constants

The remaining group of adjustable parameters are the spring constants $k_{i,j}$ that determine the strength of the harmonic restraints on bead pairs $(i,j) \in R$. To predict these spring constants from the known reference CPs, we use the general linear model

$$\mathbf{k}^* = \mathbf{W}\mathbf{p}^*. \tag{9}$$

Here, **k*** is a vector containing n predicted spring constants $k_{i,j}$ of the harmonic restraints, where n is the number of bead pairs in R; **W** is an $n \times (n+1)$ matrix of model parameters and **p*** is a vector containing $n+1$ elements, where the first n elements are the reference CPs $p^*_{i,j}$ for the pairs of beads connected by the n harmonic restraints, and the last element is a non-zero constant c that allows **W** to map the background CPs of an unrestrained chain to zero spring constants. As c is multiplied by appropriate weights, its value can be arbitrary. To minimize roundoff errors, however, we use $c = [\sum_{(i,j)\in R} p^*_{i,j}]/n$.

Now the problem of finding optimal values for the spring constants $k_{i,j}$ becomes a problem of determining the optimal elements of the matrix **W**. This is not a trivial problem, because each spring constant $k_{i,j}$ may in general affect not only the CP for the pair (i, j) of beads connected by that spring but also the CPs for other pairs of beads in the chain, including those connected by other restraints. Also, because Equation 9 is an approximation, the optimal **W** will not necessarily yield valid spring constants $k_{i,j}$ when **p*** changes. Thus, in general, an optimal **W** must be determined for each given **p***.

One could argue that predicting the $k_{i,j}$s through Equation 9 is an unnecessary complication, because optimal $k_{i,j}$s could be found more simply by using a standard optimization algorithm that adjusts the $k_{i,j}$s to minimize the sum of squared differences $\sum_{(i,j)\in R}(\hat{p}_{i,j} - p^*_{i,j})^2$. We did not pursue such a blind approach, however, because we suspected that it would be less efficient than alternative methods that take advantage of additional information about the underlying physical system. Such information, in the proposed approach, is the hypothesis that there exists an 'inverse' system that converts the CPs to spring constants according to the general linear model of Equation 9.

To find optimal elements for **W** in Equation 9, we apply the least mean squares (LMS) algorithm developed by Widrow and colleagues (45–47) (see the Appendix). This simple yet powerful algorithm has been extensively used in the field of adaptive signal processing to optimize a digital filter structure known as adaptive linear combiner (ALC). An ALC performs a dot product between a time-varying weight vector $\mathbf{w}_k$ and a time-varying input vector $\mathbf{x}_k$, thus obtaining a scalar output $y_k = \mathbf{w}_k{}^{\mathrm{T}}\mathbf{x}_k$, which is required to approximate a given desired signal d_k at each discrete time step k. To meet this requirement, the LMS algorithm uses a steepest descent scheme that iteratively adjusts the elements of the weight vector $\mathbf{w}_k$ at each time step k using

$$\mathbf{w}_{k+1} = \mathbf{w}_k + 2\mu\varepsilon_k\mathbf{x}_k, \tag{10}$$

where $\varepsilon_k = d_k - y_k$ is the error at time step k and μ is a gain factor that affects the speed of convergence and the stability of the algorithm.

To apply the LMS algorithm toward the optimization of the parameter matrix **W** in Equation 9, we allow this matrix, the CPs and the predicted spring constants to vary with iteration index k, i.e. $\mathbf{k}_k = \mathbf{W}_k\mathbf{p}_k$. We then treat the elements of $\mathbf{k}_k$ and the rows of $\mathbf{W}_k$ as the outputs and transposed weight vectors, respectively, of n ALCs,

$$\mathbf{k}_k = \left[y_{1,k}\ y_{2,k} \cdots {}_{n,k}\right]^{\mathrm{T}}, \tag{11}$$

$$\mathbf{W}_k = \left[\mathbf{w}_{1,k}\ \mathbf{w}_{2,k}\ \ldots\ \mathbf{w}_{n,k}\right]^{\mathrm{T}}. \tag{12}$$

To complete this application of the LMS algorithm, we must provide appropriate inputs to the ALCs and obtain appropriate errors, which are necessary to adjust the weight vectors. To obtain an input vector $\mathbf{x}_k$ for all ALCs, we first predict a set of restraint spring constants using the parameter matrix available at iteration k and the constant vector of reference CPs (first block in Figure 2b), i.e. $\mathbf{k}^*_k = \mathbf{W}_k\mathbf{p}^*$. Next, we use this set of spring constants to generate, through BD simulations, an ensemble of bead-chain conformations, and we use this ensemble to estimate the CPs for the restrained bead pairs $(i,j) \in R$ (second block in Figure 2b). The resulting vector $\hat{\mathbf{p}}_k$ of estimated CPs is now used as input for all n ALCs, which produce a corresponding output vector $\hat{\mathbf{k}}_k = \mathbf{W}_k\hat{\mathbf{p}}_k$ at iteration k (third block in Figure 2b). If the weights of the ALCs were optimal, then the ALC outputs in $\hat{\mathbf{k}}_k$ at iteration k would be very close to the spring constants in $\mathbf{k}^*_k$, which were used to generate the ensemble of conformations from which $\hat{\mathbf{p}}_k$ was estimated. Therefore, the ALC errors are the elements of the vector $\varepsilon_k = \mathbf{k}^*_k - \hat{\mathbf{k}}_k$, which we can finally use to compute a better estimate of the parameter matrix for the next iteration,

$$W_{k+1} = \mathbf{W}_k + 2\mu_k\varepsilon_k\hat{\mathbf{p}}_k^{\mathrm{T}}. \tag{13}$$

To ensure stability of the LMS algorithm, we need a gain factor $\mu < 1/\mathrm{tr}[\mathbf{R}]$, where tr[**R**] is the trace of the input correlation matrix (47). Thus, to calculate a safe value for μ, we let $\mu = 0.1/\mathrm{tr}[\mathbf{R}]$ and we use the approximation

$$\mathrm{tr}[\mathbf{R}] \approx \frac{1}{2}[\hat{\mathbf{p}}_k^{\mathrm{T}}\hat{\mathbf{p}}_k + \mathbf{p}^{*\mathrm{T}}\mathbf{p}^*], \tag{14}$$

where we account for both current and reference CPs in order to decrease μ when $\hat{\mathbf{p}}_k$ is large and to bound μ from above when $\hat{\mathbf{p}}_k$ is small. To determine the first vector of predicted restraint spring constants $\mathbf{k}_1^*$, we assume a linear relationship between each $k_{i,j}$ and the corresponding reference CP $p^*_{i,j}$. Specifically, we set $k_{i,j} = a_0 + a_1 p^*_{i,j}$ for $(i,j) \in R$, where $a_0 = 2 - 2p^*_{\min}/(p^*_{\max} - p^*_{\min})$, $a_1 = 2/(p^*_{\max} - p^*_{\min})$ and $p^*_{\min}(p^*_{\max})$ is the minimum (maximum) value of the reference CPs $p^*_{i,j}$ for $(i,j) \in R$. This choice yields initial spring constants ranging from 20 to 40% of the maximum value of 10 that we allow $k_{i,j}$ to take. Then, to begin the restraint optimization procedure with the first vector of estimated CPs $\hat{\mathbf{p}}_1$, we set the diagonal elements of $\mathbf{W}_1$ equal to 1 and all other elements equal to 0.

Selection of optimal ensemble

After a sufficient number of iterations, the restraint optimization procedure described above should yield a set of predicted spring constants $k_{i,j}$ that produce a good match between the CPs estimated for the 'restrained' bead pairs and the corresponding reference CPs, i.e. $\hat{p}_{i,j} \approx p^*_{i,j}$ for $(i,j) \in R$. Our goal, however, is to generate an optimal ensemble of bead-chain conformations such that the CPs estimated for 'all' bead pairs, not just the restrained ones, closely match the corresponding reference CPs, i.e. we want $\hat{p}_{i,j} \approx p^*_{i,j}$ for $1 \le i < j \le N$. To quantify the goodness of match between estimated and reference CPs, we calculate the root mean-squared deviation (RMSD) between the two sets of probabilities,

$$p_{\mathrm{RMSD}} = \sqrt{\frac{2}{N(N-1)} \sum_{1 \le i < j \le N} (p^*_{i,j} - \hat{p}_{i,j})^2}. \qquad \textbf{(15)}$$

To find the set of restraint spring constants that minimize p_{RMSD}, we perform 40 iterations of the LMS algorithm (Equation 13) during the restraint optimization procedure described above. To accelerate these iterations, we perform only 5×10^6 steps during the BD simulations from which the CPs are estimated at each iteration. In general, the conformation ensembles produced by such short simulations will depend on the initial bead-chain conformation used for the BD simulations. Therefore, to find the ensemble that minimizes p_{RMSD}, we perform several trials of the restraint optimization procedure. In each trial, we use a different initial conformation for the simulations, and we identify the set of restraint spring constants that minimize p_{RMSD} among all iterations performed. Next, these optimal spring constants and the corresponding initial conformation are used to generate a larger conformation ensemble, this time by performing 10^8 steps of BD simulation. Among the larger conformation ensembles obtained from all trials, we select the one that yields the smallest p_{RMSD}. This final ensemble is the one we deem to be optimal, i.e. most consistent with the reference CPs.

Generation of initial conformations

To obtain the different initial bead-chain conformations used for each trial of the restraint optimization procedure, one could simply generate a number of random conformations. We choose, however, a more deterministic approach aimed at generating conformations with different relative orientations of loops. Specifically, we design each initial conformation in the shape of a tight cylindrical bundle (Figure 6). To generate the bundle, all the beads connected by harmonic restraints are arranged on a circle whose circumference is just large enough to prevent overlapping those beads. Next, the intervening fragments that contain the other beads of the chain are used to connect the beads on the circle. As they join the beads on the circle, these fragments are forced to run perpendicular to the plane of the circle. Hence, there are two ways in which each fragment can connect two adjacent beads on the circle: on the same side of the plane of the circle, or on opposite sides. By connecting the beads on the circle with the intervening fragments in all possible ways, we can generate up to $2^{n_c - 1}$ distinct conformations, where n_c is the number of beads on the circle and where we omit those conformations that result from reflecting other conformations about the plane of the circle. In the present study, we selected up to 32 different bundle conformations to perform the trials of the restraint optimization procedure (Table 2).

RESULTS AND DISCUSSION

Test systems

To validate our computational method, we considered six test systems of increasing complexity. Each test system consisted of the same bead-chain model that we used to recover an optimal conformation ensemble from reference CPs. In each such system, however, we induced the formation of specific loops by connecting appropriate beads with up to eight harmonic restraints (Supplementary Table S1). To vary the complexity of these test systems, we varied the number of beads in the chain and the number of induced loops (Table 2). In particular, we simulated chains of 25, 35 and 45 beads with 2, 3 and 4 'free' loops, respectively. To mimic the effects of confinement constraints, we also simulated the same chains with additional restraints connecting the middle beads of free loops across such loops, as shown schematically in Supplementary Table S1, thus giving rise to 'tied' loops.

We used the same value of $k^*_{i,j} = k_s/200$ for the spring constants of all restrained bead pairs (i, j) in all test systems. The conformations of these test systems obtained after minimizing their potential energy are shown in Figure 3.

Reference CPs

To obtain reference sets of estimated bead CPs $p^*_{i,j}$, for $1 \le i < j \le N$, in each of the six test systems, we generated corresponding ensembles of bead-chain conformations by performing BD simulations following the same protocol described above for the ensemble recovery procedure. In particular, for each test system, we constructed an initial bead-chain conformation by threading the appropriate number of beads into the path of a 3D Hilbert curve (21). We then minimized the potential energy of the initial conformation (Figure 3), equilibrated the system with 10^6 simulation steps and performed 10^8 additional

Table 2. Validation of the conformation ensemble recovery procedure using reference CPs estimated by simulating test systems of increasing complexity

Test system[a]					Ensemble recovery[b]								
								RMSD[c]					
								2 Parameters			$n+1$ Parameters		
Figure 3[d]	N[e]	n^*[f]	Loops[g]		n[h]	N_{trial}[i]	t_c[j]	k	p	$\bar{d}$	k	p	$\bar{d}$
a	25	3	2	Free	3	4	2.9	0.050	0.0015	0.020	0.036	0.0016	0.020
b	25	4	2	Tied	4	16	2.9	0.064	0.0024	0.012	0.065	0.0048	0.018
c	35	4	3	Free	6	8	4.1	0.017	0.0015	0.016	0.056	0.0018	0.018
d	35	6	3	Tied	9	32	4.3	0.099	0.0060	0.028	0.122	0.0055	0.026
e	45	5	4	Free	10	16	5.3	0.023	0.0016	0.025	0.022	0.0019	0.025
f	45	8	4	Tied	13	32	5.5	0.175	0.0086	0.063	0.147	0.0103	0.070

[a] Characteristics of test systems used to generate conformation ensembles from which reference CPs were estimated.
[b] Results of ensemble recovery procedure applied to reference CPs.
[c] RMSD between recovered and reference values of restraint spring constants (k), CPs (p) and mean inter-bead distances ($\bar{d}$), achieved when using a general linear model (Equation 9) with the specified number of parameters per spring constant.
[d] Label used to identify test system in Figure 3.
[e] Number of beads in the chain.
[f] Number of restraints used to induce the loops in the bead chain.
[g] Number and type of induced loops.
[h] Number of restraints found by peak detection algorithm.
[i] Number of trials performed to select the optimal ensemble.
[j] Average computation time per trial in hours when performing each trial with $n+1$ parameters on one core of a 2.2-GHz AMD Opteron Processor 2427.

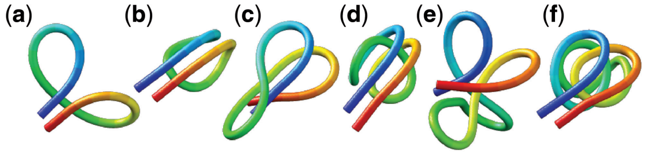

Figure 3. Energy-minimized conformations of the test systems used to generate reference CPs for validating the proposed computational method. The systems are labeled as in Table 2. Images were generated using UCSF Chimera (48).

steps, during which we collected one bead-chain conformation every 100 steps. From the collected conformations, we estimated $p^*_{i,j}$ using Equation 8. The CPs estimated for a chain of 45 beads with four free loops and four tied loops are represented as heat maps in Figure 4a. Also highlighted are the locations of bead pairs (i, j) that were connected by harmonic restraints to induce the formation of loops or to tie the loops.

These heat maps qualitatively confirm the intuition that $p^*_{i,j}$ for beads connected by restraints and for nearby beads along the chain should be greater than the background CP. These maps, however, also reveal enhanced CPs for pairs of beads that were not directly connected by harmonic restraints and that were relatively distant along the chain from other restrained beads. A similar phenomenon was observed for the chain with 35 beads, but not for the chain with 25 beads (data not shown). Thus, an enhanced probability of contact between two beads in the chain is not always due to an external force directly pulling those beads toward each other. These results underscore the complexity of interactions that can arise even for a chain of only 35 beads, when such a chain is subjected to looping constraints.

Relation between CPs and mean inter-bead distances

Simulating the above test systems to validate our computational method also provided an opportunity to investigate the behavior of chromatin assumed in previous related works. In particular, to infer the 3D conformation of chromatin from experimentally measured CPs, previous studies have assumed that the mean spatial distance d between two DNA fragments can be deduced from their CP p through a simple functional relation. For example, the power law $p \propto d^{\alpha}$ has been used with exponents $\alpha = -1.0$ (27,28) and $\alpha = -2.0$ (30). Alternatively, exponential decay (29) and logarithmic types of relations (12) have also been used. To assess whether a simple relationship between mean inter-bead distance,

$$\bar{d}^*_{i,j} = \frac{1}{N_c}\sum_{i=1}^{N_c} d_{i,j}, \tag{16}$$

and corresponding CP $p^*_{i,j}$ does hold for our test systems, we obtained $\bar{d}^*_{i,j}$ from the same conformation ensembles that were used to determine $p^*_{i,j}$.

First, however, we analyzed the results from simulations of bead chains that lacked harmonic restraints. A plot of $p^*_{i,j}$ against $\bar{d}^*_{i,j}$, for $1 \le i < j \le N$, obtained from simulating an unrestrained chain of 45 beads, shows that, in the absence of constraints inducing loop formation, the CPs follow a clear trend with a peak at $\bar{d}^*_{i,j} \approx 8\sigma$ (Figure 5a). We observed similar trends for chains with 35 and 25 beads (data not shown).

As $p^*_{i,j}$ does not vary significantly among pairs of beads separated by similar mean spatial distances $\bar{d}^*_{i,j}$ or by similar loop lengths $j - i$ (Figure 5a), it is reasonable to estimate looping probabilities by averaging $p^*_{i,j}$ over constant values of $j - i$ (Figure 5a, inset). We found that such looping probabilities for an unrestrained chain of 45 beads approximately follow the trend predicted by theory for worm-like chains with non-zero persistence length and non-zero contact distance (26), thus confirming that our simulations can reproduce the behavior of such chains. Furthermore, noting a monotonic relation between $p^*_{i,j}$ and $\bar{d}^*_{i,j}$ for $\bar{d}^*_{i,j} \geq 10\sigma$, we also fitted the power law $p^*_{i,j} \propto (\bar{d}^*_{i,j})^\alpha$ to our simulation data for the unrestrained chain, and we obtained $\alpha \approx -2.2$, which approximately agrees with the value $\alpha = -2.0$ reported in (30). These results suggest that, in the absence of harmonic restraints and for $\bar{d}^*_{i,j}$ sufficiently large, it may be appropriate to assume a simple monotonic relation between $p^*_{i,j}$ and $\bar{d}^*_{i,j}$ and to use such relation for predicting approximate values of $\bar{d}^*_{i,j}$ from known or measured values of $p^*_{i,j}$.

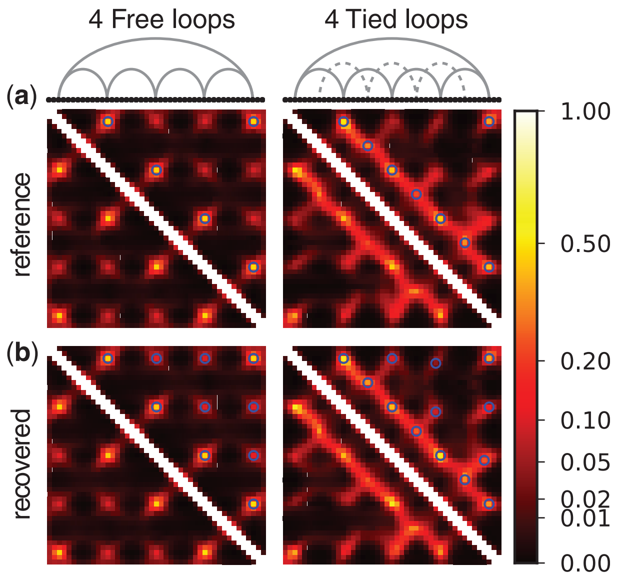

Figure 4. Heat maps representing (**a**) reference and (**b**) recovered CPs for a chain of 45 beads with (left) four free loops or (right) four tied loops. Free loops result from connecting loop end-beads with harmonic restraints (gray arcs in top-left schematic), while tied loops result from connecting middle beads across free loops (dotted arcs in top-right schematic). Blue circles on the maps identify pairs of beads that were restrained (a) when generating reference CPs and (b) when performing the ensemble recovery procedure. Test systems with two and three loops (Table 2) yielded a similarly good visual match between reference and recovered CP maps (data not shown).

We next analyzed the results from the simulations of the test systems, where specific beads were connected by restraints as described above. In this case, the plots of $p^*_{i,j}$ against $\bar{d}^*_{i,j}$ indicate that the addition of harmonic restraints complicates the relation between $\bar{d}^*_{i,j}$ and $p^*_{i,j}$ (Figure 5b and c) far beyond the clear trend obtained from the simulations of the unrestrained chains. In particular, when harmonic restraints are present, the CPs are overall greater than the corresponding values observed in the absence of restraints, and there appears to exist no simple law that relates $p^*_{i,j}$ and $\bar{d}^*_{i,j}$. In fact, different pairs of beads separated by similar mean spatial distances or by similar loop lengths yield CPs that differ significantly by up to four orders of magnitude. The observed variation of $p^*_{i,j}$ with $\bar{d}^*_{i,j}$ for the chain with four free loops is bounded by power laws with exponents as different as -3 and -8 (Figure 5b and c, dashed lines). Hence, for the test systems considered in the present study, assuming a simple functional relation and using such relation to calculate $\bar{d}^*_{i,j}$ from $p^*_{i,j}$ would introduce large fractional errors in the predicted values of $\bar{d}^*_{i,j}$, and those errors would increase with decreasing bead CPs. These results thus motivate the development of computational approaches that do not rely on calculating $\bar{d}^*_{i,j}$ from $p^*_{i,j}$ but directly compare estimated CPs with reference CPs to infer a configuration ensemble.

Method validation

Ensemble recovery from reference CPs

After obtaining the reference set of bead CPs $p^*_{i,j}$ for each test system, we applied our computational method to recover an ensemble of conformations whose estimated CPs $\hat{p}_{i,j}$ match the corresponding reference CPs. We began by selecting a few pairs of beads to be connected with harmonic restraints. This selection was performed in

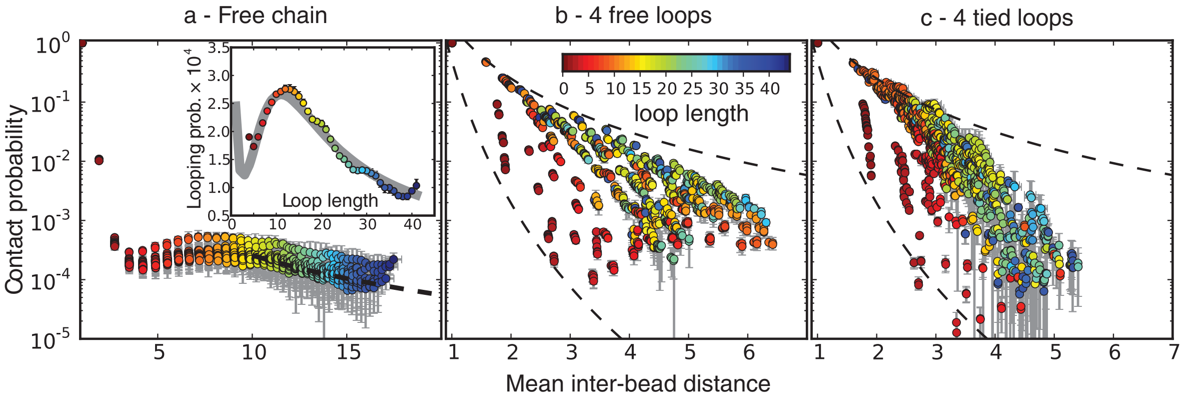

Figure 5. Variation of bead CPs with mean inter-bead distance in reference ensembles for chain (**a**) without restraints, (**b**) with four free loops and (**c**) with four tied loops. Each point represents one of the possible bead pairs in the chain. Error bars are standard deviations over 10 independent simulations. The dashed line in (a) is a fit of the power law $p \propto (\bar{d})^\alpha$, giving $\alpha = -2.23$. Inset in (a): looping probability versus loop length in ensemble for chain without restraints. The curve in this inset is a fit of Equation 3 from (26). The dashed curves in (b) and (c) are power laws with exponents -8 and -3. Distances are in units of σ.

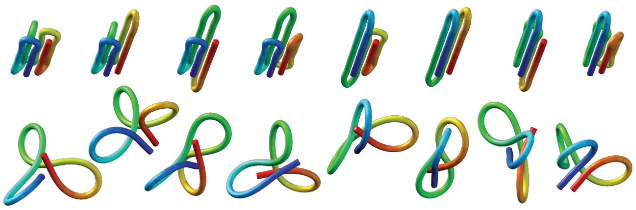

Figure 6. Initial conformations used in eight trials of the ensemble recovery procedure for a chain with 35 beads and 6 restraints (third row in Table 2), shown before (top) and after (bottom) minimization of the potential energy, Equation 1. Images were generated using UCSF Chimera (48).

an automated fashion by analyzing the reference CP maps with the peak detection algorithm described above. The algorithm successfully identified all of the bead pairs that were connected by harmonic restraints in the test systems used to generate $p^*_{i,j}$ (Supplementary Table S1). Moreover, for chains with 35 and 45 beads, the algorithm found 2, 3 or 5 additional bead pairs that were not restrained in the test systems (Supplementary Table S1) but nevertheless gave rise to CPs enhanced above the background (Figure 4).

We next adjusted the spring constants $k_{i,j}$ of the guessed restraints by performing up to 32 trials of our iterative restraint optimization procedure. Each trial used a different initial chain conformation (Figure 6) to start the BD simulations performed to estimate the CPs for the restrained bead chain.

From each trial, we obtained a different set of restraint spring constants together with the corresponding ensemble of chain conformations. For each such conformation ensemble, we used Equation 15 to calculate p_{RMSD}, the RMSD between the CPs $\hat{p}_{i,j}$ estimated for that ensemble and the corresponding reference CPs $p^*_{i,j}$ previously obtained for the test system under study. We found that p_{RMSD} varies among the trials of the ensemble recovery procedure for a given test system and that this variation increases with the complexity of the test system (Figure 7), indicating that, within the simulated time intervals, the restrained bead chain tends to get trapped into local energy minima that depend on the initial chain conformation.

For each recovered conformation ensemble, we also calculated the mean inter-bead distances $\bar{d}_{i,j}$ for $1 \le i < j \le N$. Then, to compare quantitatively these mean inter-bead distances with the corresponding reference quantities $\bar{d}^*_{i,j}$, we calculated the RMSD between the two sets of distances (30),

$$d_{\mathrm{RMSD}} = \sqrt{\frac{2}{N(N-1)} \sum_{1 \le i < j \le N} (\bar{d}^*_{i,j} - \bar{d}_{i,j})^2}. \qquad \mathbf{(17)}$$

We found that minimizing p_{RMSD} over the trials for a given test system yields the smallest, or a relatively small value for the RMSD of the mean inter-bead distances, d_{RMSD} (Figure 7). These results indicate that minimizing p_{RMSD} relative to a set of reference CPs $p^*_{i,j}$ is an effective strategy for identifying a conformation ensemble that closely matches the mean inter-bead distances of the original conformation ensemble from which the $p^*_{i,j}$ were estimated or measured.

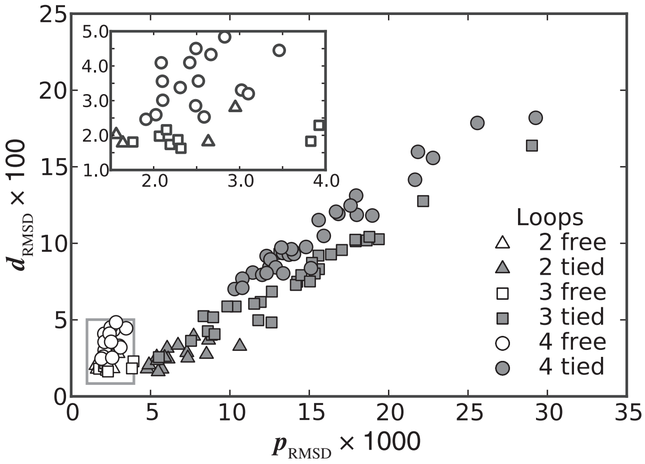

Figure 7. Plots of RMSD of mean inter-bead distances (Equation 17) against RMSD of CPs (Equation 15) for all trials of the ensemble recovery procedure and for all tested systems. Each point represents RMSD values obtained from an ensemble of 10^6 conformations at the end of a particular trial. Inset: enlarged view of boxed area. Distances are in units of σ.

Therefore, to conclude our ensemble recovery procedure, for each test system, we selected the set of restraint spring constants and the corresponding conformation ensemble that minimized p_{RMSD} among all trials. Comparing the heat maps of the CPs estimated from recovered and reference ensembles for chains with 45 beads (Figure 4a and b) shows a good qualitative agreement between the two ensembles. A similar good agreement was also observed for the simpler test systems (data not shown). This agreement is also apparent in plots of $\hat{p}_{i,j}$ against $p^*_{i,j}$ (Figure 8a).

Furthermore, the mean and standard deviation of the inter-bead distances in the recovered conformation ensembles are in excellent agreement with the corresponding quantities calculated for the reference ensembles (Figure 8b and c), confirming that our procedure successfully recovered not only the average frequency of the various inter-bead interactions but also the average inter-bead distances and the extent to which these distances fluctuate about the mean.

To visualize the reference and recovered conformation ensembles, we uniformly extracted 100 conformations from each such ensemble and aligned those conformations on the beads that were restrained in the test system used to generated the reference ensemble. The resulting 3D representations of the reference and recovered conformation ensembles reveal large fluctuations in the positions of the loops (Figure 9). The same regions of space, however, tend to be occupied by corresponding loops in the reference and recovered ensembles, thus providing a visual confirmation of the similarity between the average spatial arrangements of the two ensembles.

Simplified general linear model

The good agreement that we observed between recovered and reference ensembles is in fact a consequence of successfully optimizing the general linear model of Equation 9 with the LMS algorithm. This optimization resulted in a good prediction of restraint spring constants from the

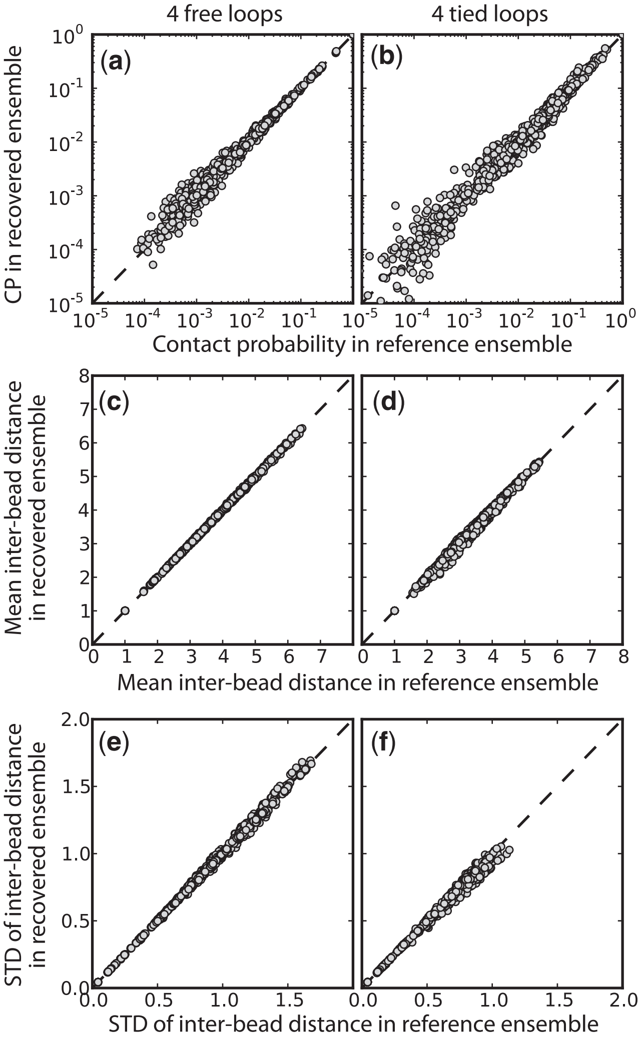

Figure 8. Comparison of (**a**,**b**) bead CPs, (**c**,**d**) mean inter-bead distances and (**e**,**f**) standard deviation of inter-bead distance determined from optimal recovered ensemble to respective quantities determined from reference ensemble for a chain with 45 beads and (a,c,e) 4 free loops or (b,d,f) 4 tied loops. Each point represents one of the possible bead pairs in the chain. The dashed lines are plots of $y = x$, not linear fits. Distances are in units of σ. Better correlations were observed for test systems with three and two loops (data not shown).

reference CPs associated with the restrained bead pairs. In fact, as noted above, some bead pairs were chosen by the peak detection algorithm for restraining, even though they were not restrained in the test systems. During the ensemble recovery procedure, however, the spring constants for the restraints on these bead pairs decreased to small values relative to the spring constants restraining the other bead pairs (Supplementary Table S1). These results indicate that the ensemble recovery procedure correctly distinguished the pairs of beads that were directly connected by harmonic restraints in the reference conformation ensemble from those pairs that were not. In particular, the RMSD,

$$k_{\mathrm{RMSD}} = \sqrt{\frac{1}{N_{\mathrm{c}}} \sum_{(i,j)\in R} (k_{i,j} - k^{*}_{i,j})^2}, \qquad \textbf{(18)}$$

between the spring constants $k_{i,j}$ predicted during the ensemble recovery procedure and the corresponding value $k^{*}_{i,j}$ used for the restraints in the test systems was <6% of $k^{*}_{i,j}$ for all such systems (Table 2). Thus, the procedure successfully deduced approximate values of the underlying spring constants using only the knowledge of reference CPs.

We asked whether a similarly good prediction of each restraint spring constant could be achieved with fewer than $n+1$ non-zero elements per row in the matrix $\mathbf{W}$ in Equation 9, i.e. with fewer than $n+1$ parameters per spring constant. To answer this question, we repeated the ensemble recovery procedure on all test systems, this time forcing all of the off-diagonal elements of $\mathbf{W}$—except those in the last column—to be zero, thus effectively using only two parameters to predict each spring constant. This choice corresponds to assuming that each restraint spring constant $k_{i,j}$ is linearly related only to the CP $\hat{p}_{i,j}$ estimated for the bead pair (i, j) restrained by that spring constant, i.e. $k_{i,j} = w_{i,j}\hat{p}_{i,j} + c_{i,j}$. We found that the resulting RMSDs of the spring constants, CPs and mean inter-bead distances did not differ appreciably from the corresponding values obtained by using $n+1$ parameters per spring constant (Table 2). Therefore, for the test systems considered in this study, it appears that the CP associated with each restrained bead pair depends primarily on the spring constant restraining that bead pair. This conclusion, however, may not hold for more complex systems, where the restraints might be less uniformly distributed among the beads and might have more variable spring constants than the restraints we used in this study to generate the reference CPs. For more complex systems, using all parameters in the general linear model of Equation 9 may be necessary to achieve adequate accuracy in the prediction of spring constants from CPs.

Computation time

The majority of the computation time required by the proposed ensemble recovery procedure is consumed by the BD simulations. These simulations are needed to estimate the CPs either for adjusting the spring constants through the LMS algorithm or for selecting the optimal conformation ensemble through a comparison of p_{RMSD} values among the ensembles obtained from different initial conformations. The simulations must be sufficiently long to ensure that the variance of the CPs estimated using Equation 8 does not outweigh the variation in CP due to differences in spring constants and initial conformations. In our work with the test systems, obtaining CPs sufficiently precise to ensure rapid convergence of the LMS algorithm in all trials of the restraint optimization procedure required 5×10^6 steps per simulation. On the other hand, ensuring that the optimal ensemble selected from several trials of the procedure matches the diversity of conformations present in the corresponding reference ensemble required 10^8 simulation steps, i.e. the same number that was used to obtain each reference ensemble. The average computation time per trial was found to increase linearly with the number of beads (Table 2).

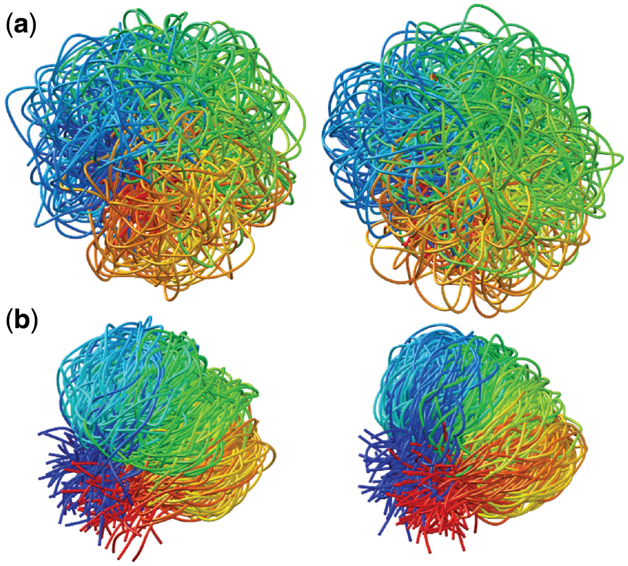

Figure 9. Spatial representation of reference (left) and recovered (right) conformation ensembles for bead chains of 45 beads with (**a**) four free loops and (**b**) four tied loops. Each depicted ensemble consists of 100 conformations extracted at equal intervals from 10^8 steps of BD simulation and aligned on the beads that were connected by harmonic restraints in the simulations used to generate the reference ensembles. The coloring order along each chain is red, yellow, green, cyan and blue. Images were generated using UCSF Chimera (48).

CONCLUSION

We have developed a computational approach to recover chromatin conformation ensembles from a set of reference CPs. The overall strategy of this approach consists of comparing the given set of reference CPs to a set of CPs obtained from simulations of a restrained bead-chain polymer model of chromatin. The results of this comparison are used iteratively to adjust the parameters of the polymer model so that, after a sufficient number of iterations and trials, an optimal conformation ensemble is obtained whose CPs closely match the corresponding reference probabilities. We have validated this procedure by using reference data sets obtained from simulations of six test systems of increasing complexity. For all such systems, the procedure yielded a conformation ensemble whose CPs, mean inter-bead distances and standard deviation of inter-bead distances all agree very closely with the corresponding reference quantities. The most complex test system that we considered was a chain of 45 beads, equivalent to roughly 135–270 kb, containing four tied loops (Figure 3f). Although this system is much smaller than the genomic loci typically investigated in 3C-based experiments, it does provide initial support to the validity of the proposed computational approach, which can already be used to investigate the spatial organization of small genomic regions.

To enable efficient and accurate analysis of experimental data sets obtained from large genomic loci, entire chromosomes or even entire genomes, the proposed computational approach will require additional improvement and validation. For example, whereas the present approach estimates CPs for beads representing fragments of equal lengths, 3C-based experiments typically provide reference CPs for fragments of various lengths. This mismatch could be overcome by mapping the experimental fragments onto the bead-chain contour and by estimating CPs for pairs of mapped fragments, rather than for pairs of beads. Another issue is computational effort. Although the procedure we described lends itself to parallelization, with each trial executing on a separate processor core, it may nevertheless become too demanding for large genomic loci. Computational effort could be lowered by improving the efficiency of conformation sampling, for example through Monte Carlo simulations, and by avoiding Equation 8 in the estimation of CPs, for example by inferring inter-bead distance distributions from sample means and higher moments of $d_{i,j}$. Finally, further validation of the procedure will require not only the simulation of larger and more complex test systems but also the availability of experimental data sets that include both 3C and FISH measurements on the same genomic region. Applications of our method to the analysis of experimental data and to the study of specific phenomena, such as gene clustering (31), are important issues that will be addressed in the future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1.

FUNDING

American Cancer Society [Instructional Research 70-002 provided to G.A. through the Moores Cancer Center, University of California, San Diego] and ARCS Foundation, San Diego Chapter [scholarship to D.M.]. Funding for open access charge: American Cancer Society.

Conflict of interest statement. None declared.

REFERENCES

1. Felsenfeld,G. and Groudine,M. (2003) Controlling the double helix. *Nature*, **421**, 448–453.
2. Blackwood,E.M. and Kadonaga,J.T. (1998) Going the distance: a current view of enhancer action. *Science*, **281**, 60–63.
3. Adhya,S. (1989) Multipartite genetic control elements: communication by DNA loop. *Annu. Rev. Genet*., **23**, 227–250.
4. Carter,D., Chakalova,L., Osborne,C.S., Dai,Y.f. and Fraser,P. (2002) Long-range chromatin regulatory interactions in vivo. *Nat. Genet*., **32**, 623–626.
5. Schleif,R. (1992) DNA looping. *Annu. Rev. Biochem*., **61**, 199–223.
6. Gheldof,N., Tabuchi,T.M. and Dekker,J. (2006) The active FMR1 promoter is associated with a large domain of altered chromatin conformation with embedded local histone modifications. *Proc. Natl Acad. Sci. USA*, **103**, 12463–12468.
7. Janicki,S.M., Tsukamoto,T., Salghetti,S.E., Tansey,W.P., Sachidanandam,R., Prasanth,K.V., Ried,T., Shav-Tal,Y., Bertrand,E., Singer,R.H. *et al*. (2004) From silencing to gene expression: real-time analysis in single cells. *Cell*, **116**, 683–698.
8. Müller,W.G., Walker,D., Hager,G.L. and McNally,J.G. (2001) Large-scale chromatin decondensation and recondensation

regulated by transcription from a natural promoter. *J. Cell Biol.*, **154**, 33–48.

9. Tumbar,T., Sudlow,G. and Belmont,A.S. (1999) Large-scale chromatin unfolding and remodeling induced by VP16 acidic activation domain. *J. Cell Biol.*, **145**, 1341–1354.
10. Cook,P.R. (1999) The organization of replication and transcription. *Science*, **284**, 1790–1795.
11. Parada,L.A. and Misteli,T. (2002) Chromosome positioning in the interphase nucleus. *Trends Cell Biol.*, **12**, 425–432.
12. Baù,D., Sanyal,A., Lajoie,B.R., Capriotti,E., Byron,M., Lawrence,J.B., Dekker,J. and Marti-Renom,M.A. (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, **18**, 107–114.
13. Lucas,J.S., Bossen,C. and Murre,C. (2011) Transcription and recombination factories: common features? *Curr. Opin. Cell Biol.*, **23**, 318–324.
14. Cremer,T. and Cremer,C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292–301.
15. Levsky,J.M. and Singer,R.H. (2003) Fluorescence in situ hybridization: past, present and future. *J. Cell Sci.*, **116**, 2833–2838.
16. Huang,B., Babcock,H. and Zhuang,X. (2010) Breaking the diffraction barrier: super-resolution imaging of cells. *Cell*, **143**, 1047–1058.
17. Zhao,Z., Tavoosidana,G., Sjolinder,M., Gondor,A., Mariano,P., Wang,S., Kanduri,C., Lezcano,M., Singh Sandhu,K., Singh,U. *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.*, **38**, 1341–1347.
18. Simonis,M., Klous,P., Splinter,E., Moshkin,Y., Willemsen,R., de Wit,E., van Steensel,B. and de Laat,W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, **38**, 1348–1354.
19. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. *et al.* (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
20. Rodley,C.D.M., Bertels,F., Jones,B. and O'Sullivan,J.M. (2009) Global identification of yeast chromosome interactions using genome conformation capture. *Fungal Genet. Biol.*, **46**, 879–886.
21. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
22. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
23. de Wit,E. and de Laat,W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev.*, **26**, 11–24.
24. Iyer,B., Kenward,M. and Arya,G. (2011) Hierarchies in eukaryotic genome organization: insights from polymer theory and simulations. *BMC Biophys.*, **4**, 8.
25. Marti-Renom,M.A. and Mirny,L.A. (2011) Bridging the resolution gap in structural modeling of 3D genome organization. *PLoS Comput. Biol.*, **7**, e1002125.
26. Rippe,K. (2001) Making contacts on a nucleic acid polymer. *Trends Biochem. Sci.*, **26**, 733–740.
27. Fraser,J., Rousseau,M., Shenker,S., Ferraiuolo,M., Hayashizaki,Y., Blanchette,M. and Dostie,J. (2009) Chromatin conformation signatures of cellular differentiation. *Genome Biol.*, **10**, R37.
28. Duan,Z., Andronescu,M., Schutz,K., McIlwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and Noble,W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
29. Tanizawa,H., Iwasaki,O., Tanaka,A., Capizzi,J.R., Wickramasinghe,P., Lee,M., Fu,Z. and Noma,K. (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.*, **38**, 8164–8177.
30. Rousseau,M., Fraser,J., Ferraiuolo,M., Dostie,J. and Blanchette,M. (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, **12**, 414.
31. Gehlen,L.R., Gruenert,G., Jones,M.B., Rodley,C.D., Langowski,J. and O'Sullivan,J.M. (2012) Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions. *Nucleus*, **3**, 370–383.
32. de Gennes,P.G. (1979) *Scaling Concepts in Polymer Physics*. Cornell University Press, New York.
33. Shimada,J. and Yamakawa,H. (1984) Ring-closure probabilities for twisted wormlike chains. Application to DNA. *Macromolecules*, **17**, 689–698.
34. Robinson,P.J.J., Fairall,L., Huynh,V.A.T. and Rhodes,D. (2006) EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proc. Natl Acad. Sci. USA*, **103**, 6506–6511.
35. Rosa,A., Becker,N.B. and Everaers,R. (2010) Looping probabilities in model interphase chromosomes. *Biophys. J.*, **98**, 2410–2419.
36. Tokuda,N., Terada,T.P. and Sasai,M. (2012) Dynamical modeling of three-dimensional genome organization in interphase budding yeast. *Biophys. J.*, **102**, 296–304.
37. Allen,M.P. and Tildesley,D.J. (1987) *Computer Simulation of Liquids*. Oxford University Press, New York.
38. Langowski,J. and Heermann,D.W. (2007) Computational modeling of the chromatin fiber. *Semin. Cell Dev. Biol.*, **18**, 659–667.
39. Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) *Numerical Recipes in C*, 2nd edn. Cambridge University Press, Cambridge, UK.
40. Iniesta,A. and de la Torre,J.G. (1990) A second-order algorithm for the simulation of the Brownian dynamics of macromolecular models. *J. Chem. Phys.*, **92**, 2015–2018.
41. Klenin,K., Merlitz,H. and Langowski,J. (1998) A Brownian dynamics program for the simulation of linear and circular DNA and other wormlike chain polyelectrolytes. *Biophys. J.*, **74**, 780–788.
42. Ghirlando,R., Litt,M.D., Prioleau,M.N., Recillas-Targa,F. and Felsenfeld,G. (2004) Physical properties of a genomic condensed chromatin fragment. *J. Mol. Biol.*, **336**, 597–605.
43. Lavelle,C. (2009) Forces and torques in the nucleus: chromatin under mechanical constraints. *Biochem. Cell Biol.*, **87**, 307–322.
44. Haynes,W.M. (ed.), (2011) *CRC Handbook of Chemistry and Physics*, 92nd edn. CRC Press, Boca Raton, FL, USA.
45. Widrow,B., Glover,J.R.J., McCool,J., Kaunitz,J., Williams,C., Hearn,R., Zeidler,J., Eugene Dong,J. and Goodlin,R. (1975) Adaptive noise cancelling: principles and applications. *Proceedings of the IEEE*, Vol 63, pp. 1692–1716.
46. Widrow,B. (1970) Adaptive filters. In: Kalman,R.E. and De Claris,N. (eds), *Aspects of Network and System Theory*. Holt, Rinehart and Winston, New York, pp. 503–587.
47. Widrow,B. and Stearns,S.D. (1985) *Adaptive Signal Processing*. Prentice Hall, Englewood Cliffs, NJ.
48. 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**, 1605–1612.

APPENDIX

THE LMS ALGORITHM

The LMS algorithm (45,46) has found numerous applications in the field of adaptive signal processing, including adaptive system identification, adaptive inverse modeling, adaptive control and adaptive interference canceling (47). This algorithm was developed to optimize iteratively and dynamically the weights of a digital filter structure, known as ALC, that performs the dot product $y_k = \mathbf{w}_k{}^{\mathrm{T}}\mathbf{x}_k$, where y_k is the ALC output, $\mathbf{w}_k$ is a vector containing $n+1$ adjustable weights, $\mathbf{x}_k$ is a vector containing $n+1$ inputs and k is the current time step for the inputs, weights and output. The choice of the inputs in $\mathbf{x}_k$ and the role of the output y_k depend on the specific application of the ALC. All applications, however, include a desired signal d_k and require the adjustment of $\mathbf{w}_k$ at each time step k so that the output y_k is, on average, as close as possible to d_k or, equivalently, so that the magnitude of the error

$$\varepsilon_k = d_k - y_k = d_k - \mathbf{x}_k^{\mathrm{T}}\mathbf{w}_k, \qquad \textbf{(A1)}$$

averaged over a long interval of k, is as small as possible. The degree to which the ALC meets this requirement can be quantified, as a function of $\mathbf{w}_k$, by defining the quadratic performance surface $\chi = E[\varepsilon_k{}^2]$, where the expected value $E[\cdot]$ is taken over the time step k. Hence, the requirement to achieve optimality of the ALC is that the weight vector $\mathbf{w}_k$ be adjusted at each time step k to minimize χ. The LMS algorithm addresses this requirement by using a variant of the steepest descent algorithm. This variant replaces the gradient $\nabla\chi$ of the quadratic performance surface with a simpler estimate obtained at time step k directly from ε_k^2, i.e.

$$\nabla\chi \approx \nabla\varepsilon_k^2 = 2\varepsilon_k\nabla\varepsilon_k = -2\varepsilon_k\mathbf{x}_k, \qquad \textbf{(A2)}$$

where $\nabla = \left[\frac{\partial}{\partial w_1}\ \frac{\partial}{\partial w_2}\ \cdots\ \frac{\partial}{\partial w_{n+1}}\right]^{\mathrm{T}}$ is the gradient operator with respect to the components of the weight vector $\mathbf{w}_k$. The gradient estimate $\nabla\varepsilon_k^2$ is then used to calculate an improved weight vector from the current one,

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mu\nabla\varepsilon_k^2 = \mathbf{w}_k + 2\mu\varepsilon_k\mathbf{x}_k, \qquad \textbf{(A3)}$$

where $\mu > 0$ is a gain factor that determines the size of the step along the negative gradient estimate. A small value of μ causes slow convergence, whereas too large a value of μ causes instability of the algorithm. It has been shown (47) that the LMS algorithm is stable for $\mu < 1/\mathrm{tr}[\mathbf{R}]$, where $\mathbf{R} = E[\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}]$ is the input correlation matrix. The strengths of the LMS algorithm are its simplicity, robustness and relatively rapid convergence despite the presence of noise in the input $\mathbf{x}_k$ and desired signal d_k.