Measuring chromosome conformation with degenerate labels
Measuring chromosome conformation with degenerate labels

Brian C. Ross
Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Paul A. Wiggins
Departments of Physics and Bioengineering, University of Washington, Seattle, Washington, USA
(Received 24 February 2012; revised manuscript received 18 June 2012; published 19 July 2012)

Although DNA conformation plays an integral role in all genetic processes from transcription to chromosome segregation, there is as yet no tractable method for capturing the \textit{in vivo} conformation of a chromosome at high resolution. Labeling and fluorescently imaging thousands of loci along the chromosome would readily yield a conformation if each locus could be uniquely distinguished in the image, but this would unrealistically require thousands of distinguishable labels and a tedious experimental process. Here we present a computational method for extracting conformations when the total number of labels far exceeds the number of distinguishable labels. We evaluate our technique using simulated conformations with lengths ranging from 10 to 100 kilobases, and discuss the prospects for an experiment.

DOI: 10.1103/PhysRevE.86.011918

PACS number(s): 87.15.hp, 36.20.Ey

I. INTRODUCTION

The positioning of genomic loci plays a central role in many cellular processes. For example, the expression of genes in eukaryotes correlates with their positioning relative to the nuclear lamina [1], nuclear pores [2], and transcription factories [3]; and the need to condense chromosomes and segregate them upon replication are challenges faced by all organisms. The mapping between the genomic position of a locus and its location in three-dimensional space is called a \textit{conformation}. In order to grasp the complexity of a chromosomal conformation, we can imagine replacing its continuous length-$L$ contour with discrete segments, where the segment length is on the order of the persistence length (bending scale) in order to resolve all the bends of the contour. (An \textit{in vivo} persistence length is a poorly defined quantity due to excluded volume, DNA binding factors, sequence dependence, etc.; to be precise we can define it as the mean distance over which the tangent correlation of the backbone drops by one exponential factor.) The persistence length is poorly known \textit{in vivo} and almost certainly varies between and within cells, but if we approximate it by its \textit{in vitro} value $l_p \approx 50 \text{ nm} \approx 150 \text{ bp}$ then the conformation of whole chromosomes is thus approximately described by a number of segments $N = L/l_p$, where $N$ ranges from ten thousand in bacteria up to a million for some eukaryotes. How to map out this rich structure \textit{in vivo} is still an unsolved problem.

A number of techniques have been developed to extract conformational information on scales from the bending persistence length up to the chromosome contour length, which can be hundreds of megabases. Electron microscopy (EM) has been used to trace the DNA contour at high resolution in \textit{ex vivo} chromosomes [4], but labeling techniques for EM are primitive and it is extremely difficult to identify individual genetic loci. Although EM is well suited to measuring some sequence-independent features of DNA, it is difficult to assess connectivity and thus hard to extract large-scale \textit{in vivo} conformations.

Cell-averaged contact maps between pairs of loci, coarse-grained on the scale of tens of kilobases, can be generated using the chromosome conformation capture (3C) technique [5] and its variants [6–11], from which inferences have been made regarding the underlying conformations [12–15]. Because 3C-based methods probe contact frequencies between pairs of loci rather than locus positions directly, conformational information inferred from 3C data contains considerable uncertainty. Furthermore, all chromosome-capture measurements are averaged over a heterogenous cell population. Another method is to use fluorescence microscopy to localize individual loci in single cells [16]; unfortunately, this technique has mostly been applied to large-scale structure, and only a handful of genomic regions have been colabeled because the labels have had to occupy distinct fluorescence channels. The advent of single-molecule super-resolution fluorescence microscopy [17–19] opens the possibility of imaging DNA contours at or below the persistence length scale [20], but unless the number of distinguishable labels can be increased into the thousands, the increased resolution will not directly allow the measurement of chromosome-scale conformations at high genomic resolution.

Here we propose a single-cell experiment in which a large number of predefined loci are labeled \textit{in vivo} using a small number of colors (so that there are many labels per distinguishable color), localized by fluorescence microscopy, and finally distinguished from their identically colored neighbors using a computational analysis which we term \textit{spatial alignment}. Multilabeling schemes have been employed previously [21] but these have involved only a few labels per color and were not used to extract explicit conformations. An example of a labeled conformation in our experiment is shown in Fig. 1(a).

The computational analysis we propose compares the known spacing and color ordering of labels along the DNA contour [Fig. 1(b)] to the positions and colors of spots in the image [Fig. 1(c)] to construct a locus-to-spot mapping. In contrast to sequence alignment, positional and spacing information is needed for spatial alignment because the ordering of spots in the image is not known. Predictive mappings require very close label spacing, ideally on the order of the bending scale of...
FIG. 1. (Color online) Mapping of simulated 10-kb conformation. (a) Simulated three-dimensional 10-kb DNA contour ($l_p = 50$ nm) decorated with colored fluorophores using the error parameters described in the text. Localization error is indicated by the lines connecting the recorded positions of the spots to their true locations on the contour. False positives have a four-pointed star shape; open circles indicate false negatives. (b), (c) The inputs to the mapping algorithm are the genomic (b) and imaged (c) positions of a set of labels along with their colors (grayscale in print). (d) A graphical representation of the mapping probabilities output by the algorithm. Each element of the probability array, which maps the genomic locus at $l$ base pairs to the spot imaged at $(x, y, z)$, is represented by a circle centered at $(l, z)$ and having an area proportional to the probability. False negative probabilities are given by the line of circles (pink online) along the top. The solid line connects the correct mappings, skipping false negatives, which are instead marked with X’s inside their respective (pink) circles. The fact that the true conformation generally passes through the largest circle of each locus located between 0 and 8 kb indicates a good mapping within this region.

DNA, and comparable resolution in the image such as provided by super-resolution fluorescence microscopy. If the analysis is fast and accurate enough, such an experiment could bridge the kilobase-scale resolution gap between the DNA-bending scale and the resolution of 3C methods, while avoiding cell averaging.

The contribution of this paper is a computational spatial alignment method that infers probabilistic conformations from (a) the genomic position and label color of each locus and (b) the physical position and color of a set of spots in an image. False positives (nonspecifically bound labels) and false negatives (undetected labels) both are taken into account, as is localization error. Our method makes an approximation that lowers the accuracy but greatly speeds up the analysis, so that reconstructions involving ~1000 labeled loci (~100 kb at bending-scale resolution) can be performed on a desktop computer in about an hour. Chromosome-scale conformational reconstructions are straightforward but require significantly more memory and computation time owing to the larger number (>10⁴) of labels needed.

II. RESULTS

A. Method overview

Our algorithm outputs a table of probabilities for mapping genomic loci to imaged spots. In order to properly construct these probabilities, one should consider only conformations in which no two genomic loci map to the same spot in the image; unfortunately, enforcing this rule exactly makes the problem intractable when more than a handful of loci are involved. Our method ignores this “no-overlap rule” except between adjacent mappings (i.e., two loci that are consecutive or separated only by false negatives), thereby making the solution much more tractable but introducing considerable error into the probability table. In order to recover some of this error, we associate a penalty with every spot in the image and iteratively recompute the probabilities while adjusting the penalty factors in order to minimize a cost function $C$. The cost function establishes two constraints on the probability table: (1) the expected number of mapped loci (excluding false negatives) should match the estimated number of spots in the image that are not false positives; and (2) the normalization condition $\sum_i p_{i\alpha} \leq 1$ must hold for each spot $\alpha$ in the image. Details of the algorithm are given in the appendix.

In order to evaluate the quality of a locus-to-spot mapping, we compute the Shannon information [22] per locus needed to specify the unique conformation, given the partial information already contained in the mapping probabilities. Shannon information is a positive number that is inversely related to the probability assigned to the correct mapping; a final information score of zero implies a perfect mapping ($p = 1$ for the correct conformation). A probabilistic mapping lacks this amount of information relative to the perfect mapping, so the objective of our algorithm is to recover as much of
this missing information as possible and thereby minimize the information measure. The information score can be measured in simulations for which the correct mapping is known, but unfortunately not in a real experiment. To estimate the mapping quality in a real experiment, we average the information metric over all possible mappings weighted by the mapping probabilities, obtaining a score we term entropy. Entropy is therefore an estimate of the information score that does not require knowledge of the true mapping; the entropy and information scores should be nearly equal if the mapping probabilities accurately reflect their likelihoods of being correct (see supplemental figures [23]). Note that accurate mapping probabilities are a necessary, but not sufficient, condition for a good mapping: A state of uniform probabilities is accurate but uninformative and consequently scores high on both information and entropy.

The computer program used for our calculations, along with source code, can be downloaded at [24].

B. Mappings of simulated conformations

To check our implementation of the algorithm, we verified that the numerical gradient of the cost function computed by the first step of the algorithm equals the analytical gradient computed by the second step. In the special case where all same-color pairs of loci are adjacent, we checked that the mapping probabilities are exact, by direct enumeration of all possible mappings.

In order to evaluate the quality of our mappings, we generated 1000 10-kb conformations for defined experimental conditions using the wormlike chain model [50-nm persistence length, three colors, 30 spots; 10% false positive and false negative rates; localization error of \( 10 + (2/15)|z| \) nm along, and \( 22 + (1/15)|z| \) perpendicular to, the focal plane defined by \( z = 0 \) ] and constructed mapping arrays for each. We then compared these locus-to-spot probabilities to their empirical likelihoods of corresponding to correct mappings (see supplemental figures at [23]). Ideally these would be equal: For example, in roughly 40% of the cases where a locus-to-spot mapping probability is \( p_{a-b} = 0.4 \), locus \( a \) should have produced spot \( b \) in the image. We also attempted several control mappings for each conformation, in order to determine the extent to which the quality of the experimental data could be blindly determined by the algorithm (see supplemental figures at [23]). Each control attempted to map the original DNA contour, with an unchanged label spacing and color ordering, to an image identical to the original but with the spot colors randomly permuted. We found several metrics that can indeed discriminate the true mapping from the controls (although a high false negative rate makes this difficult), indicating that the analysis can validate the preparation and imaging of a single experimental sample.

We next used the short 30-spot problem to create the mapping visualization shown in Fig. 1. Conformations were randomly generated (10 kb, 30 spots, three colors, error parameters as before), mapped using our algorithm, and the entropies before and after each mapping were recorded. The fifth conformation we considered produced a mapping that would look good to an experimenter, where the entropy drop exceeded one bit per labeled locus, and that had a significant number of false negatives for demonstration purposes. We plotted this conformation in Fig. 1(a), along with the experimental output in Figs. 1(b) and 1(c), and the final mapping in Fig. 1(d). The nearest distance between two distal points on this chain is about 6 nm, which implies two things: (1) excluded volume is ignorable since the radius of DNA is about 1 nm; (2) looped conformations (which occur in vivo) are effectively modeled by the conformations studied here since localization error is greater than the distance of closest approach. To visualize our mapping more concretely, we constructed the conformation that passed through the largest mapping probability at each locus [Fig. 2(c)], acknowledging that this is a simple but naive method that sometimes maps two loci to the same imaged spot. As Fig. 2 makes explicit, the uncertainty in a reconstructed conformation is a composition of the coarse grain of the contour, experimental error, and mapping error.

The quality of a mapping depends partly on the three experimental error parameters: the false positive rate, false negative rate, and localization error. We generated three sets of DNA conformations: three- and two-dimensional 10-kb contours and three-dimensional 100-kb contours. The error parameters used to label and “image” each conformation were varied randomly within the set, explaining the large spread in information recovery. Figure 3 shows the information recovery and entropy change for each of the runs for which the cost function \( C \) converged near to zero; the mapping of Fig. 1 is included on the three-dimensional (3D) 10-kb plot for comparison (green dots). A more detailed picture of

![FIG. 2. Discrete conformations. (a) The coarse-grained approximation to the contour that connects labels at their true positions, superimposed upon the full conformation (shaded line). Mismatch regions can occur where consecutive labels are widely spaced (due the random label spacing), making it impossible to resolve the intervening contour. (b) The coarse-grained contour connecting the imaged spots in their proper order, taking experimental error into account. False negative errors cause some spots to be missed, and localization error offsets the remaining spots. (c) The coarse-grained contour that connects the imaged spots (with experimental error) based on the maximum computed mapping probability for each locus (note that this heuristic does not strictly enforce no overlap).](image)
FIG. 3. (Color online) Information recovery from simulated conformations. Three sets of simulated DNA conformations were generated: 3D and two-dimensional (2D) 10-kb contours and 3D 100-kb contours. Three labeling colors were used, at a labeling density of approximately one label per 1000 base pairs per color; however, the experimental error parameters were varied randomly with each conformation. (a) Sample contours and labelings for each set. False positives, false negatives, and microscope error are all shown. (b) Change in entropy (S) and information scores (I) over the course of the mapping procedure, from an initial state of uniform probabilities that enforce the false negative rate (red online, light gray in print) to the final state (blue online, dark gray in print). There is one red (light) and blue (dark) dot for each mapping in a set that converged after 100 (400) iterations for the 10-kb (100-kb) mappings. One hundred conformations were generated per set. The 100-kb set converged considerably more slowly, so the tolerances were relaxed; squares have looser tolerances than circles and show poorer information recovery. The upper (lower) outlined dots (green online) in the 3D 10-kb plot respectively give the initial (final) states of the run from Fig. 1.

Our results gives us several checks on the systematic errors that might arise from the approximations of the algorithm. The quality of the mapping probabilities can be gauged from the supplemental figures at [23]; the S shape in the first graph indicates that our algorithm is somewhat underconfident in driving mapping probabilities from the starting mean. For a range of experimental conditions one or more of the following happens when the spot colors are put in the correct order relative to the controls (see supplemental figures at [23]): (1) the cost function converges faster to zero, (2) the final-state entropy decreases, and (3) the logarithm of the final-state partition function increases for low false negative rates. This gives a check on the quality of the experimental data. The mapping quality is fairly insensitive to the only free parameter in the analysis, which is the decay constant \( l_p \) in the propagator (see the appendix). One peculiarity of our approach is that there is a continuum of solutions—\( N \) independent free parameters are solved using \( N \) inequality constraints and one equality constraint—so the final mapping probabilities depend on the initial guess of the free parameters. This is unusual but does not alter the fact that positive information is almost always recovered in cases of low experimental error. Encouragingly, we find that entropy is generally a reasonable proxy for information, although entropy change tends to underestimate information recovery for good mappings (see Fig. 3).
Our method requires an experiment that labels DNA at defined base-pair locations and images those labels, within certain tolerances set by the limitations of our algorithm. Labeling can be accomplished in vivo using fluorescence in-situ hybridization (FISH) or a repressor-operator system (FROS) [25]; one might also apply the technique ex vivo to flat DNA spreads using fluorescent fusions to DNA-binding proteins such as restriction enzymes. Imaging methods that photoswitch and centroid-fit discrete fluorophores photoactivated localization microscopy (PALM) [17], fluorescence photoactivation localization microscopy (FPALM) [18], and stochastic optical reconstruction microscopy (STORM) [19] have already demonstrated the ability to image in three dimensions [26,27] and in multiple color channels [28,29], at a resolution suitable for mapping uncompacted conformations like the ones we tested. These imaging methods are slow, so fixation would be required to prevent movement at small scales.

It is important to have an irregular label spacing and color ordering, so that every region of DNA can be uniquely identified. For short DNA contours it is easy to engineer each individual probe or binding site to satisfy this requirement. For longer contours, various barcoding techniques might be used that exploit the random distribution of restriction sites to heterogenously label the DNA in a single step. For example, a series of restriction digests followed by adaptor ligations could be used as polymerase chain reaction (PCR) templates for generating end-labeled FISH probes or else fused to operator binding sites to construct a FROS binding array (which would need to be sequenced after assembly to determine the locus spacing). In an ex vivo experiment, fluorescently fused restriction enzymes could be used directly for imaging the restriction sites, as unbound and freely diffusing enzymes are hard to detect by PALM or STORM. If the probes can be replaced during the experiment then different colors can be imaged sequentially, obviating the need for multiple fluorescence channels.

The quality of a mapping depends on the density of spots, the number of colors, and the experimental error (localization error, false negative and false positive rates). For our simulated mappings we generated uncompacted wormlike chain DNA using the in vitro persistence length of 50 nm, and labeled it using the equivalent of three symmetric five-cutter restriction enzymes (one label per 1024 bases per color). An important variable is the distance of locus $i$ from one end of the DNA, measured along the contour; $R_i$ is the location in three-dimensional space of imaged spot $i$, and $\sigma_{i+n,i+n}$ is the statistical weight corresponding to the stretching of two loci $i$ and $i+n$ between the locations of spots $i$ and $i+n$ in the image. We have associated an unphysical free parameter $f_{i,n}$ with each mapping $i \rightarrow j$, and an unbinding penalty $w$ with each false negative $j \rightarrow i$, so each mapped locus $i$ is separated from the next mapped locus by $n_i-1$ intervening false negatives. In order to prevent two adjacent mappings from targeting the same imaged spot, we set $\sigma_{i+n,i+n} = 0$ if $\alpha_i = \alpha_{i+n}$, but our expression cannot enforce $\alpha_i \neq \alpha_j$ for nonadjacent $i$ and $j$.

Our algorithm, conformational reconstructions of up to several hundred kilobases of DNA can be performed using established experimental techniques. A host of interesting DNA structures exist at or below the 100-kb scale, from the configurations of the mitochondrial genome and putative topological domains in E. coli genomes [4] at the 10-kb scale up to the span of DNA loops in higher eukaryotes (∼100 kb). There is still room to refine the mapping algorithm, in order to allow for longer mappings, and to raise the mapping quality and thereby loosen the experimental tolerances. One obvious potential improvement is to enforce the no-overlap constraint between certain pairs of nonadjacent loci. Better ways of extracting the information recovered in the probability matrix will also be welcome, especially for cases where the mapping quality is too poor for a unique conformation to be apparent to the eye.

ACKNOWLEDGMENTS

The authors thank David Johnson for helpful discussions regarding the traveling salesman problem and Nathan Kuwada for useful comments on the paper. Funding for this work came from the National Science Foundation under Grant PHY-084845.

APPENDIX

A particular conformation is determined by a set of mapping variables $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$, where each $\alpha_i$ either indexes the spot in the image corresponding to the $i$th locus along the contour or else $\alpha_i = \emptyset$ in the case of a false negative. For example, \{3,5,2\} denotes the conformation where the first contour locus maps to spot 3 in the image, the second contour locus maps to spot 2 in the image, the second contour locus was not imaged at all, and the presumed spot 1 in the image must have been a false positive.

If we ignore correlations between nonadjacent $\alpha_i$, the statistical weight of a given mapping between two loci $i$ and $j$ is a simple product of terms connecting each pair of loci along the contour:

$$e^{-F(u)} = \sqrt{f_{i,n,i,j} f_{j,n,j,i}} \prod_{i=n+1}^{b-n_j} \sqrt{f_{i,n,i,j} e^{-F(u_{i+1,n_j} - b, R_{j}, R_{j+n_j})}} \times \sqrt{f_{i+1,n,i,j} w_{i+1,n_j}^{n_j-1}} \equiv \sqrt{f_{i,n,i,j} f_{j,n,j,i} \prod_{i=n+1}^{b-n_j} \sigma_{i+n,i+n}} . \tag{A1}$$

Here $l_i$ is the distance of locus $i$ from one end of the DNA, measured along the contour; $R_i$ is the location in three-dimensional space of imaged spot $i$, and $\sigma_{i+n,i+n}$ is the statistical weight corresponding to the stretching of two loci $i$ and $i+n$ between the locations of spots $i$ and $i+n$ in the image. We have associated an unphysical free parameter $f_{i,n}$ with each mapping $i \rightarrow j$, and an unbinding penalty $w$ with each false negative $j \rightarrow i$, so each mapped locus $i$ is separated from the next mapped locus by $n_i-1$ intervening false negatives. In order to prevent two adjacent mappings from targeting the same imaged spot, we set $\sigma_{i+n,i+n} = 0$ if $\alpha_i = \alpha_{i+n}$, but our expression cannot enforce $\alpha_i \neq \alpha_j$ for nonadjacent $i$ and $j$.

We have opted to model the statistical weighting factor $\sigma(R; L)$ using the Gaussian chain distribution [30], where $\sigma(R; L) = (2\pi s^2)^{-3/2} \exp(-R^2/2s^2)$ and $s^2 = 2l_p L/3$. (The Gaussian chain fails definitively at short contour lengths, so for $L < 2l_p$ we set $s^2 = L^2/3$.) There are several reasons to use a Gaussian model: (1) The true form of $\sigma$ is unfortunately
poorly constrained in vivo and almost certainly differs between and within organisms. Since the Gaussian chain model is a smooth diffusive distribution, it should not fail sharply if the in vivo polymer bends in unexpected ways. (2) There is only one free parameter: the decay constant \( l_p \), which sets the bending length scale of the polymer (\( l_p \approx 50 \text{ nm} \approx 150 \text{ bp} \) for in vitro DNA). (3) Convolving the localization error of the two loci (usually assumed Gaussian, having variances \( s_1^2 \) and \( s_2^2 \)) against a Gaussian \( \sigma(t) \) simply results in a new Gaussian \( \sigma'(t) \) having \( s'^2 = s_1^2 + s_2^2 \).

It turns out that nearly every quantity that depends upon a contour locus \( i \) also depends on the corresponding mapping variable \( \alpha_i \). For notational convenience, we henceforth omit the explicit dependences of most quantities on the \( \alpha_i \) variables. As a general rule, every superscripted or subscripted index attached to a quantity implies that it is also a function of the respective mapping variable \( \alpha \) corresponding to that index: \( f_a \) is shorthand for \( f_{\alpha_i} \), \( \sigma_i^{+n} \) is shorthand for \( \sigma_{\alpha_i}^{+n} \), etc.

Consider the quantity \( Z_b \) (shorthand for \( Z_{\alpha_i}^{+b} \)) that sums all possible mappings of a stretch of consecutive loci bounded by \( a \leq i \leq b \), where the endpoints are fixed by the implicit \( \alpha_a \) and \( \alpha_b \) but each intervening \( \alpha_i \) is summed over all imaged spots of the appropriate color:

\[
Z_b = \sum_{\alpha_a + \ldots + \alpha_{b-1}} \prod_{i=a}^{b-1} \sigma_i^{+\alpha_i}. \tag{A3}
\]

By setting \( a \) to the first mapped locus on the entire chain and summing over \( \alpha_x \), we obtain the half-partition function \( Z^b \), which accounts for all mappings that end at \((b, \alpha_b)\). Because we have relaxed the no-overlap condition, \( Z^b \) can be efficiently calculated using a recursive rule:

\[
Z^b = \sum_{a \prec b} \sum_{\alpha_a} w_a^{a-1} \sqrt{fa} Z_a^b \tag{A4}
\]

\[
= \sum_{m=1}^{b-1} \left( \sum_{\alpha_a} \right) \sigma_{b-m}^{a} Z^{b-m} + w_{b-1}^{b-1} \sqrt{fb}. \tag{A5}
\]

The boundary terms can be absorbed by imagining that all chains begin from some outside locus 0, having \( Z_{00} = 1 \) and \( \sigma_0^{b} = w_{b-1}^{b-1} \sqrt{fb} \):

\[
Z^b = \sum_{m=1}^{b} \left( \sum_{\alpha_a} \right) \sigma_{b-m}^{a} Z^{b-m}. \tag{A6}
\]

In like manner we obtain the second half-partition function \( Z_a \), which accounts for all mappings beginning at \((a, \alpha_a)\). We account for the boundary terms by considering all chains to end at an imaginary \( N + 1 \) locus having the properties \( Z_{N+1} = 1 \), \( \sigma_a^{N+1} = w^{N-a} \sqrt{fa} \) and \( \sigma_b^{N+1} = w^N \):

\[
Z_a = \sum_{b \succ a} \sum_{\alpha_b} Z_b^{b} \sqrt{fb} w^{N-b} \tag{A7}
\]

\[
= \sum_{n=1}^{N} \sum_{\alpha_{a+n}} \sigma_a^{a+n} Z_{a+n}. \tag{A8}
\]

Finally, the full partition function \( Z \), which contains all possible mappings of the entire chain, is the summation of \( Z_b \) over both endpoints. In practice \( Z \) is best calculated by choosing a locus \( x \), summing over all possible mappings of that locus, and adding terms arising from \( \alpha_x = \emptyset \):

\[
Z = \sum_{a=1}^{N} \sum_{b=a}^{N} \sum_{\alpha_a} w_a^{a-1} \sqrt{fa} Z_a^b \sqrt{fb} w^{N-b} \tag{A9}
\]

\[
= \sum_{a} \sum_{m=1}^{N} \sum_{n=1}^{N} \sum_{\alpha_a, \alpha_a+n} Z^{1-m} \sigma_a^{a+n} Z_{a+n}. \tag{A10}
\]

From the various partition functions we immediately obtain the mapping probabilities \( p(x \rightarrow \alpha_x) \):

\[
p_x = \frac{Z_x Z_{s \rightarrow x}}{Z}. \tag{A11}
\]

The amount of unrecovered information associated with this probability matrix is \( I = - \sum \log p_{x|\alpha_x} \), where the vector \( \alpha' \) contains the true mapping; this score is used to evaluate the algorithm on simulated inputs for which \( \alpha' \) is known. The entropy of the probability matrix, defined by \( S = - \sum \sum \log p_i \), gives an estimate of the unrecovered information when the true mapping is not known.

To enforce proper normalization of the probability array and tune the false negative rate, we perform gradient descent on the following cost function with respect to the weighting parameters \( w = \{ f_i \} \):

\[
C = \frac{1}{2} \sum_{a} \left( \max \left\{ 0, \sum_i p_{i \mid \alpha_a} - 1 \right\} \right)^2 + \frac{K}{2} \left[ \left( \sum_{i} p_{i \mid \alpha_a} - N_s \right)^2 \right], \tag{A12}
\]

where \( N_s \) is the estimated number of imaged spots that are not false positives. We break up the cost-function gradient in the following way:
Four of the six terms are straightforward:

\[
\frac{dC}{dp_i} = \max \left\{ 0, \sum_j p_{j,\alpha_i} - 1 \right\} + K \left[ \left( \sum_j \sum_{\alpha_j} p_j \right) - N_\delta \right],
\]

(A15)

\[
\frac{\partial C}{\partial Z_i} = \frac{dC}{dp_i} \frac{Z_i}{Z} + \frac{dC}{dZ_i} Z_i \sum_{n=1}^{N-i+1} \left( \sum_{i=1}^{j} \sum_{\alpha_j} \frac{dC}{dp_j} \frac{Z_j}{Z} \right) \left( \sum_{\alpha_{i+n}} \sigma_{i+n}^{a_i} \right),
\]

(A16)

\[
\frac{\partial C}{\partial Z_i} = \frac{dC}{dp_i} \frac{Z_i}{Z} + \frac{dC}{dZ_i} Z_i \sum_{n=1}^{i-1} \left( \sum_{j=1}^{m} \sum_{\alpha_j} \frac{dC}{dp_j} \frac{Z_j}{Z} \right) \sum_{n} \sigma_m^{a_i} \sigma_{i-m}^{j+n},
\]

(A17)

\[
\frac{\partial p_i}{\partial w_k} = \frac{dC}{dZ_i} \sum_{m=1}^{N-i+1} \sum_{n=1}^{\alpha_i m} \sum_{\alpha_{i+n}} Z_i^{m-n} \frac{d\sigma_{i+m}^{a_i n}}{d\sigma_{i-m}} \frac{Z_i^{n+i} \sigma_{i+n}^{a_i},}{dZ_i},
\]

(A18)

where

\[
\frac{dC}{dZ_i} \equiv -\sum_{\alpha_i} \frac{dC}{dp_i} \frac{Z_i}{Z},
\]

(A19)

\[
\frac{d\sigma_{i+j}^{a_i n}}{d\sigma_{m}^{a_i n}} = \frac{\delta_{m-j} \delta_{m-j}^{a_i n} \sigma_{i-m}^{j+n}}{2f_j},
\]

(A20)

\[
\frac{d\sigma_{i+m}^{a_i n}}{d\sigma_{i-m}} = \frac{m+n-1}{w} \sigma_{i-m}^{a_i n}.
\]

(A21)

The remaining step is to calculate \(dZ_i/dw_k\) and \(dZ_i/dw_k\), each of which involves chains of derivatives owing to the recursive structure of the half-partition functions. To calculate these derivatives efficiently, we note an analogy between our computation of half-partition functions and the propagation of signals through neural networks, and borrow the famous backpropagation algorithm [31] in order to compute our derivatives at the same order as the rest of our calculation. Consider a linear feedforward neural network with connectivity between all layers, where the output of neuron \(\alpha_i\) in layer \(i\) is \(x_i\) (shorthand for \(x_i,\alpha_i\)), and the weight connecting neurons between two layers as \(W_{ij}\). To compute the \(x_i\) we use the recursive rule that \(x_i = \sum_{m=1}^{\alpha_i m} \sum_{\alpha_{i-m}} W_{i-m}^{n+i} x_{m-n}\); this is equivalent to our calculation of \(Z_i\) or \(Z_i\). The gradient \(s_i \equiv dC/dx_i\), called a sensitivity, can be propagated backwards in a second recursive step: \(s_i = \frac{\partial C}{\partial x_i} + \sum_{m=1}^{N-i+1} \sum_{a_{i-m}} W_{i-m}^{n+i} s_{i+m}\). The gradient in the weights is then \(dC/dw_k = x_{i-m} \delta_{m}^{a_i},\) the correspondence with our situation is that \(Z_i\) or \(Z_i\) is analogous to the output \(x_i\) of a given neuron, and \(\sigma_{a_i}^{b}\) plays the role of the weight matrix \(W_{a_i}^{b}\). Using this analogy, we back propagate the sensitivities using the following rule:

\[
s_i = \frac{\partial C}{\partial Z_i} + \sum_{m=1}^{N-i+1} \sum_{a_{i-m}} \sigma_{i+m}^{a_i} s_{i+m},
\]

(A22)

\[
s_i = \frac{\partial C}{\partial Z_i} + \sum_{m=1}^{N-i+1} \sum_{a_{i-m}} \sigma_{i+m}^{a_i} s_{i+m}.
\]

(A23)

The component of \(\frac{\partial C}{\partial \sigma_{a_i}^{b}}\) that comes through the dependence of \(C\) on the half-partition functions is

\[
\frac{\partial C}{\partial \sigma_{a_i}^{b}} \bigg|_{Z, Z_i} = s_{b}^a Z_i + s_{a}^b Z_i,
\]

(A24)

so the component of \(\frac{\partial C}{\partial f_i}\) and \(\frac{\partial C}{\partial w}\) coming through the half-partition functions is

\[
\sum_j \sum_{\alpha_j} \left( \frac{\partial C}{\partial Z_i} + \frac{\partial C}{\partial Z_j} \right) = \frac{1}{2} \left( \sum_{m=1}^{N-i+1} \sum_{a_{i-m}} (\sigma_{i-m}^{a_i} Z_{i-m} + \sigma_{i-m}^{a_i} s_{i-m} Z_i) + \sum_{n=1}^{N-i+1} \sum_{a_{i+n}} (\sigma_{i+n}^{a_i} s_{i+n} Z_{i+n} + \sigma_{i+n}^{a_i} s_{i+n} Z_i) \right)
\]

(A25)

\[
\sum_j \sum_{\alpha_j} \left( \frac{\partial C}{\partial Z_i} + \frac{\partial C}{\partial Z_j} \right) = \sum_{a_{i-m}} \sum_{n=1}^{N-i+1} \sum_{a_{i+n}} (\frac{b-a-1}{w} Z_{a_{i-m}} s_{a_{i+m}} + s_{a_{i-m}} Z_{a_{i+n}} + s_{a_{i+n}} Z_{a_{i-m}}).
\]

(A27)

Our derivation assumes a separate weighting factor \(f_i\) for each mapping \(i \rightarrow \alpha_i\). In practice much better results are obtained when there is just one weighting factor per spot in the image, which all mappings to that spot use: \(f_{i,a} = f_{j,a} = f_{a}\) for all \(i\), which implies that \(dC/df_{i,a} = \sum_i dC/df_{i,a}\). The gradient optimizer actually works with the
binding energies $E = -\log w$ rather than the weighting factors themselves, in order to keep all $w = \{f_a, w\}$ positive. The cost function derivatives in the weighting energies are $dC/dE_i = -w_i (dC/dw_i)$.

We note that our problem is related to a generalized time-dependent traveling salesman problem (TDTSP) \cite{32} except that we also treat false positives and false negatives. Whereas the TDTSP asks for the unique minimum-energy conformation (optimum tour), our method estimates the partition function, which sums over all possible conformations.

The obvious way of calculating the TDTSP partition function exactly—by brute-force enumeration—unfortunately scales exponentially in the number of loci $N$. Using our approximate algorithm, memory requirements and computation time (per iteration) are both roughly proportional to $N^4$: There are $N^3 \times (N_c \times N_f)$ elements of each array, each of which couples to all $\sim N$ spots of all $\sim N$ preceding layers (due to false negatives). We greatly shorten the calculation by truncating the latter two sums: We ignore both long runs of consecutive false negatives and all $a_b$ factors between distant $a$ and $b$, subject to some probability threshold in either case. As a result each iteration of our computation is roughly of order $N^2 \rho$, where $N$ is the number of loci and $\rho$ is the density (mean number of neighboring spots).

---

[1] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talbaut, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat et al., Nature (London) 453, 948 (2008).

[2] J. M. Casolari, C. R. Brown, S. Komili, J. West, H. Hieronymus, and P. A. Silver, Cell 117, 427 (2004).

[3] C. S. Osborne, L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik et al., Nat. Genet. 36, 1065 (2004).

[4] L. Postow, C. D. Hardy, J. Arsuaga, and N. R. Cozzarelli, Genes Dev. 18, 1766 (2004).

[5] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, Science 295, 1306 (2002).

[6] Z. Zhao, G. Tavoosidana, M. Sjölinde, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh et al., Nat. Genet. 38, 1341 (2006).

[7] M. Simonis, P. Klous, E. Splinter, Y. Moskvin, R. Willemsen, E. De Wit, B. Van Steensel, and W. De Laat, Nat. Genet. 38, 1348 (2006).

[8] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum et al., Genome Res. 16, 1299 (2006).

[9] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner et al., Science 326, 289 (2009).

[10] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei et al., Nature (London) 462, 58 (2009).

[11] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang et al., Cell 148, 84 (2012).

[12] D. Bäü, A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom, Nat. Struct. Mol. Biol. 18, 107 (2011).

[13] M. A. Umbarger, E. Toro, M. A. Wright, G. J. Porreca, D. Bäü, S.-H. Hong, M. J. Fero, L. J. Zhu, M. A. Marti-Renom, H. H. McAdams et al., Mol. Cell 44, 252 (2011).

[14] T. Sexton, E. Yaffe, E. Kenisberg, F. Bantignies, B. Leblanc, M. Hottman, H. Pannellino, A. Tanay, and G. Cavalli, Cell 148, 458 (2012).

[15] Y. Zhang, R. P. McCord, Y. J. Ho, B. R. Lajoie, D. G. Hildebrand, A. C. Simon, M. S. Becker, F. W. Alt, and J. Dekker, Cell 148, 908 (2012).

[16] P. A. Wiggins, K. C. Cheveralls, J. S. Martin, R. Lintner, and J. Kondev, Proc. Natl. Acad. Sci. USA 107, 4991 (2010).

[17] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, Science 313, 1642 (2006).

[18] S. T. Hess, T. P. K. Girirajan, and M. D. Mason, Biophys. J. 91, 4258 (2006).

[19] M. J. Rust, M. Bates, and X. Zhuang, Nat. Methods 3, 793 (2006).

[20] A. Matsuda, L. Shao, J. Boulanger, C. Kervrann, P. M. Carlton, P. Kner, D. Agard, and J. W. Sedat, PloS One 5, e12768 (2010).

[21] L. S. Shopland, C. R. Lynch, K. A. Peterson, K. Thornton, N. Kepper, J. von Hase, S. Stein, S. Vincent, K. R. Molloy, G. Kreth et al., J. Cell Biol. 174, 27 (2006).

[22] C. E. Shannon, ACM SIGMOBILE Mobile Comput. Commun. Rev. 5, 3 (2001).

[23] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.86.011918 for additional figures.

[24] www.phys.washington.edu/~pwiggins/align3d/.

[25] C. C. Robinett, A. Straight, G. Li, C. Willhelm, G. Sudlow, A. Murray, and A. S. Belmont, J. Cell Biol. 135, 1685 (1996).

[26] B. Huang, W. Wang, M. Bates, and X. Zhuang, Science 319, 810 (2008).

[27] G. Shtengel, J. A. Galbraith, C. G. Galbraith, J. Lippincott-Schwartz, J. M. Gillette, S. Manley, R. Sougrat, C. M. Waterman, P. Kanchanawong, M. W. Davidson et al., Proc. Natl. Acad. Sci. USA 106, 3125 (2009).

[28] M. Bates, B. Huang, G. T. Dempsey, and X. Zhuang, Science 317, 1749 (2007).

[29] H. Shroff, C. G. Galbraith, J. A. Galbraith, H. White, J. Gillette, S. Olenych, M. W. Davidson, and E. Betzig, Proc. Natl. Acad. Sci. USA 104, 20308 (2007).

[30] H. Yamakawa, Helical Wormlike Chains in Polymer Solutions (Springer, Berlin, 1997).

[31] D. E. Rumelhart, G. E. Hintont, and R. J. Williams, Nature (London) 323, 533 (1986).

[32] L. Gouweia et al., Eur. J. Operat. Res. 83, 69 (1995).