METHOD ARTICLE

Improved inference of chromosome conformation from images of labeled loci [version 1; referees: 1 approved with reservations]

Brian C. Ross, James C. Costello
Computational Bioscience Program, Department of Pharmacology, University of Colorado, Anschutz Medical Campus, Aurora, CO, 80045, USA

Abstract

We previously published a method that infers chromosome conformation from images of fluorescently-tagged genomic loci, for the case when there are many loci labeled with each distinguishable color. Here we build on our previous work and improve the reconstruction algorithm to address previous limitations. We show that these improvements 1) increase the reconstruction accuracy and 2) allow the method to be used on large-scale problems involving several hundred labeled loci. Simulations indicate that full-chromosome reconstructions at 1/2 Mb resolution are possible using existing labeling and imaging technologies. The updated reconstruction code and the script files used for this paper are available at: https://github.com/heltilda/align3d.

Keywords

chromosome, conformation, reconstruction, fluorescence, genetic, loci

This article is included in the International Society for Computational Biology Community Journal gateway.

Corresponding authors: Brian C. Ross (brian.ross@ucdenver.edu), James C. Costello (JAMES.COSTELLO@ucdenver.edu)

Author roles: Ross BC: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Costello JC: Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Funding was provided by the Boettcher Foundation (J.C.C.), National Institutes of Health [2T15LM009451 to B.C.R.], and a Cancer League of Colorado grant (J.C.C. and B.C.R.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Ross BC and Costello JC. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ross BC and Costello JC. Improved inference of chromosome conformation from images of labeled loci [version 1; referees: 1 approved with reservations] F1000Research 2018, 7(ISCB Comm J):1521 (https://doi.org/10.12688/f1000research.16252.1)

Introduction

Measurement of in vivo chromosome conformation is a major unsolved problem in structural biology despite its known biological importance. The present state-of-art is either to obtain indirect information about conformations using 3C-derived methods which measure DNA-DNA contacts (typically in a cell-averaged population), or else to directly measure the cellular locations of individual chromosomal loci in single cells by microscopy. The major limitation of direct localization is one of throughput: only ~3 – 5 labeled loci can be uniquely identified ‘by color’ in a standard microscope image, whereas a whole-chromosome reconstruction would involve labeling and identifying hundreds or thousands of loci. Several research efforts aim to remove this limitation using either experimental or computational approaches. The experimental approaches aim to allow an increased number of labels that can be distinguished in an image. Alternatively, computational methods are being developed to infer the identity of each label if it cannot be uniquely identified in an image, by using the known label positions along the DNA contour. Here we focus on the computational inference method called align3d, and present improvements to allow high-quality, chromosome-scale conformational reconstructions.

First, we briefly describe the align3d algorithm. Using a) the genomic locations and colors of labeled loci and b) the spatial locations and colors of spots in a microscope image, together with a relation tying the genomic distance between two loci to their average spatial displacement, this method constructs a table of ‘mapping probabilities’ \( p(L \rightarrow s) \) for a given labeled genomic locus \( L \) having produced spot \( s \) in the microscope image. Each mapping probability \( p(L \rightarrow s) \) is calculated by dividing the summed statistical weights of conformations where locus \( L \) maps to spot \( s \), which we term a mapping partition function and denote \( Z_{L,s} \), by the full partition function \( Z \) that is the summed weight of all conformations. A proper calculation of \( Z_{L,s} \) and \( Z \) would consider all conformations having no more than one locus at any given spot in the image, similar to a traveling salesman tour, but this exact calculation is intractable for large problems. Instead, align3d counts all conformations for which adjacent loci do not overlap at the same spot (see Figure 1), using a variant of the forward-backward algorithm that can propagate between non-adjacent layers. This is a major source of error as the vast majority of conformations contributing to the partition function overlap at non-adjacent loci, and one consequence is that the normalization of mapping probabilities makes no sense for a non-overlapping conformation, as \( \sum_s p(L \rightarrow s) \) can exceed 100% for certain spots. To recover from this error, align3d assigns a penalty to each spot and iteratively adjusts these penalties until the spot penalties regardless of the number of labeled loci. This rule guarantees that a) the rate of missing spots is as expected, and b) the mapping probabilities are properly normalized. Let \( q_s \) denote the penalty attached to spot \( s \); then the update rule for that spot penalty is:

\[
q_s = \frac{1}{1 - p_m(c)} \cdot \frac{P(s)/N}{\min(1, P(s)/N)} \cdot q_s^{-1}
\]

where \( N \) is the number of loci, \( P(s) = \sum_s p(L \rightarrow s) \) is the total probability of mapping any locus to spot \( s \), and \( p_m(c) \) is the estimated rate of missing spots having color \( c \). The justification for this rule is given in Appendix I (Supplementary File 1).
We can also update a penalty $q$ that is associated with missing spots of color $c$. This gives a faster way to enforce a desired missing spot rate because there are fewer $q$ penalties than $q$ parameters. An update to $q_c$ is equivalent to a reverse update to all $q_c$ for spots $s$ of color $c$, so the update rule is:

$$q'_c = \frac{1}{1 - \frac{1}{P(s)/N}} q_c. \tag{2}$$

Typically, we first optimize the $q$ parameters to achieve a target missing spot rate, then optimize the $q$ parameters to enforce $P(s) \leq 1$ while maintaining the missing spot rate. In either case, we apply Equation 1 or Equation 2 to bring the $q$ or $q$ parameters close to their final values. When the cost function stops improving, we switch to the steepest-descent algorithms used in Ross and Wiggins, 2012 to polish $q$ or $q$.

Next, we give two exact formulas for the partition functions $Z_{<n>}$ and the full partition function $Z$ that determine our locus-to-spot mapping probabilities. We focus on the full partition function $Z$ since the formulas for $Z_{<n>}$ are identical. The largest term in each formula, which we denote $Z_0$ (or $Z_0^n$ when spot penalty optimization is used), is the original estimate from Ross and Wiggins, 2012 calculated using a variant of the forward-backward algorithm. Additional terms are computed in the same way, except that certain loci are constrained to map to certain spots. All of the constraints we will apply are illegal constraints, in that they force multiple loci to overlap at some spot in the image; therefore these terms only count illegal conformations that we would like to remove from the baseline calculation. By computing these terms and subtracting them from $Z_0$, we eliminate the overlapping conformations and improve the calculation. It turns out that this process erroneously subtracts conformations with multiple overlaps more than once and thus we have to add back in higher-order corrections (i.e. partition functions having multiple constrained spots). Repeating this logic leads to exact formulas for $Z$ taking the form of series expansions, which are dominated by the lowest-order terms as those have the fewest restrictions on conformational overlaps. Figure 2A illustrates an example of such a series expansion, where each parenthesis subscript ($X, Y, \ldots$), on a term label denotes an illegal constraint forcing loci $X, Y, \ldots$ to overlap at spot $s$ when that term is calculated. We use this notation throughout.

There are two ways we might remove conformations containing overlapping loci, leading us to two different series expansions for the true partition function $Z$. Suppose that we are calculating the term $Z_{<n>}$ whose single illegal constraint forces loci $A, C, \ldots$ to overlap at spot $s$. One option is to forbid any of the other unconstrained loci from also mapping to spot $s$, since spot $s$ is already overused. This leads to series expansion 1. Alternatively, allowing further overlaps with spot $s$ from the unconstrained loci gives us series expansion 2. Figure 2B illustrates the differences between the two series.

Each of the two series expansions is a weighted sum over all possible illegally-constrained terms having two properties:

1) each locus and each spot appear at most once in the indices, and 2) two or more loci map to each constrained spot. To be formal, we use $\Omega$ to represent the set of all possible illegal constraints: each element of $\Omega$ consists of a set of two or more non-adjacent loci and a single spot where they are forced to overlap. Each expansion thus takes the form

$$Z = \sum_{\phi \in \Omega} w_\phi Z_\phi$$

where $Z_\phi$ is zero if any two constraints share a locus or spot. We will choose the integer weights $w_\phi$ so as to cancel out the overlapping conformations. By symmetry arguments, the weighting factor should not depend on the identities of the loci or spots, but rather only by the number of constrained spots $n_\phi$, and the number of loci $n^2_\phi$ involved in each $k^2$ constraint. For example, $w_{(ACE)_{BD}}$ is determined by $n_\phi = 2$, $n^2_\phi = 3$ and $n^2_\phi = 2$. 

![Figure 2. Series expansions.](image-url)
Here we specify each series expansion by giving a formula for the weights \( w_\phi \) in terms of \( n_\phi \) and the various \( N_\psi \). We also explain how to select an appropriate set of terms \( \psi \) when there are too many terms to evaluate. Our selection prohibits any legal or overlapping conformation from contributing a negative weight to the partition function estimate, thereby guaranteeing positive mapping probabilities and allowing use of the reconstruction-quality metrics given in Ross and Wiggins, 2012. Derivations of the coefficient formulas and the term-selection criteria for each series expansion appear in Appendix 2 (Supplementary File 1).

**Series expansion 1** For series expansion 1, we do not allow the unconstrained loci to map to spots that were used in constraints. Then the weights \( w_\phi \) are given by:

\[
w_\phi = (-1)^{n_\phi}
\]  
(3)

To select terms for a series approximation, we first choose a set of illegal constraints \( \psi \) to disallow, then include all series terms \( Z_\psi \) containing only those constraints: i.e. \( \phi \subset \psi \). This guarantees non-negative mapping probabilities. In order to efficiently evaluate the largest terms, we recommend selecting the \( N_\psi \) constraints having the highest product of mapping probabilities in the baseline calculation \( Z_\psi \) (or \( Z_\psi^{\text{opt}} \) if spot penalties will be used). For example, we would include \( (AC) \) if \( p(A \rightarrow s) \cdot p(C \rightarrow s) \) is sufficiently large.

**Series expansion 2** For series expansion 2, the unconstrained loci are allowed to map to spots that were used in constraints. Then the weights \( w_\phi \) are given by:

\[
w_\phi = \prod_{k=1}^{n_\phi} (-1)^{n_\psi} (n_\phi - 1).
\]  
(4)

To select terms for a series approximation, we first choose a set of \( N_\psi \) single-locus-to-spot mappings \( \Psi \), then include all terms \( Z_\phi \) whose illegal constraints use only mappings in \( \Psi \). For example, the constraint \( (AC) \) would be included if \( \Psi \subseteq \{ A \rightarrow s, C \rightarrow s \} \). In order to select the largest terms, we recommend building \( \Psi \) from the \( N_\psi \) largest mapping probabilities calculated from \( Z_\psi \) or \( Z_\psi^{\text{opt}} \).

**Results**

We tested the improved align3d method by generating random chromosome conformations using our software tool wormulator (version 1.1), and simulating the process of error-prone labeling, imaging and finally producing the locus-to-spot mapping probabilities. We considered three scenarios for our simulations. 1) The ‘Toy’ scenario involves 10 genomic loci, where each locus is labeled using one of 3 colors. For these simple problems the partition function can be calculated exactly. 2) Our simulated Experiment 1 uses standard DNA labeling methods and traditional 3-color microscopy to label 30 loci with 3 colors, thus interrogating a significant fraction of a chromosome contour. 3) Our simulated Experiment 2 labels 300 loci across a chromosome-length contour. The reconstruction of Experiment 2 is made possible by using the Oligopaints labeling technique to label in 20 different colors.

For each scenario, we randomly generated 100 conformations using a wormlike chain model (packing density = 0.3 kb/nm, persistence length = 300 kb, as suggested by the measurements of Trask, Pinkel and van den Engh, 1989); applied a random labeling at a mean density of 1 locus per megabase; and simulated experimental error: 100/200-nm Gaussian localization error in xy/z, a 10% rate of missing labels, and a 10% rate of nonspecifically-bound labels. A typical simulated experiment from the Toy scenario is shown in Figure 3A.

![Figure 3. Example reconstruction. A. Randomly generated and labeled chromosome contour with simulated experimental error: localization error (lines offsetting spots from the labeled genomic loci) and missing labels (open circles). This example lacks nonspecifically-bound labels (floating spots). B. Spot mapping probabilities calculated using both the largest series term \( Z_\psi \) (grey circles), and the exact \( Z \) that can be computed using 2210 series terms (blue circles). The dotted red line connects the true locus-to-spot mappings, which are used to calculate the unrecovered information. \( I(Z_\psi) = 0.835 \text{ bits/locus} \) and \( I(Z) = 0.54 \text{ bits/locus} \). C. Unrecovered information \( l \) and entropy \( S \) (left panel) and log \( Z \) (right panel) versus the number of terms used in the series expansions.](image-url)
For each simulated conformation, we fed the label positions and colors together with the simulated 3D images into the align3d algorithm to produce locus-to-spot mapping probabilities. For example, the experiment shown in Figure 3A produced the mapping probabilities shown graphically in Figure 3B using circles, where the size of each circle indicates probability magnitude. Here grey circles show the mapping probabilities computed from \( Z_i \) with no use of spot penalties, and blue circles show those same probabilities computed using the exact \( Z \). This example shows how excluding high-weight and heavily-overlapping conformations reduces and improves the partition function estimate (see Figure 3C) and concentrates the probability mass into the ‘true’ locus-to-spot mappings (shown connected by the dotted red line in Figure 3B).

Our reconstruction quality metric is the amount of unrecovered information from the mapping probabilities, defined as \( I = -\left( \log p(L \rightarrow s) \right) \), where the average \( \langle \cdot \rangle \) is taken over the set of true locus-to-spot mappings \( (L_i, s_i) \). The ideal case of \( I \rightarrow 0 \) implies a perfect reconstruction with no mistakes and zero uncertainty, but in practice \( I \) is always positive. In a real experiment where the true mappings are not known, we use a proxy for unrecovered information that we term entropy, defined as \( S = -\left( \log p(L \rightarrow s) \log p(L \rightarrow s_j) \right) \), whose average is taken over all locus-to-spot mappings, not just the correct mappings. The goal is to have \( S = I \) so that a real experiment will have an accurate estimate of the reconstruction performance. The left-hand panel of Figure 3C shows how \( I \) and \( S \) depend on the accuracy of the calculation for the simple example shown, using either of the two series expansions and varying the number of terms from 1 (simply \( Z \)) to 2210 which is the full set of terms for either series and thus computes \( Z \) exactly. Entropy generally overestimates the amount of unrecovered information (see Supplementary Figure S1 and Supplementary Figure S2, Supplementary File 1), because the large mapping probabilities should be even larger, and the small ones even smaller, than their assigned values (see Supplementary Figure S3, Supplementary File 1). We believe this miscalibration is mostly caused by align3d overestimating the missing spot rate.

Validation of Equation 1–Equation 4. We first validated each of the two series expansions by comparing them against exact partition function calculations for the simulated Toy experiments. In all cases, both series expansions, when taken to their maximum number of terms, exactly reproduced the partition function calculations obtained by direct enumeration over all possible non-overlapping conformations. This test validates Equation 3 and Equation 4. We also verified that both series expansions could be used in conjunction with spot penalty optimization (Equation 1 and Equation 2), both by numerically validating the cost function gradient calculation and by testing for convergence on these small problems.

Improved optimization allows large-scale reconstructions. First, we tested whether the iterative spot-penalty optimization rules given by Equation 1 and Equation 2 could work on large-scale problems such as those of Experiment 2, where the old gradient descent optimizer in align3d had difficulty. The results are shown in Figure 4, which compares the number of iterative steps required to converge the \( q \) (missing-spot penalty) and \( q \) (spot penalty) parameters without/with use of our improved optimization rules (labeled ‘old’/‘new’ respectively in the legend). Since the spot penalties \( q \) are optimized for probability normalization only after \( \tilde{q} \) parameters have been optimized to achieve a desired missing spot frequency, we only attempted to optimize the \( q \) parameters for simulations where \( \tilde{q} \) converged. There were two results from this experiment. First, more attempts to optimize the \( \tilde{q} \) and \( q \) parameters successfully converged when using the new optimization rules in conjunction with gradient descent, as indicated by the greater volume of the ‘new’ histogram and the correspondingly larger numbers shown in the legends. Secondly, of the trials that did converge, our new method required significantly fewer iterations and thus less computation time than the old method, as indicated by the relative skews of the distributions.

Use of more colors dramatically improves reconstructions. Our most striking result is that simulations of the ambitious Experiment 2 produce far better results than even the Toy scenario, despite the fact that these simulations have more loci per color than either the Toy scenario or Experiment 1. This can be seen in the amount of unrecovered information \( I \) shown in the simulation-averaged plots of Figure 5A. Thus a push to 20-color labeling could prove critical for genomic reconstruction at the chromosome scale and beyond. At the end of this section we revisit Experiment 2, in order to assess the reconstruction quality when analyzing more realistic DNA contours having tighter confinement.

Series expansion 2 outperforms series expansion 1. Next, we compared the convergence properties of our two expansions on the three scenarios of simulated experiments. Figure 5A gives a sense of how the amount of unrecovered information varies with the number of terms taken in each series, without (solid lines)
Figure 5. Comparison of the convergence rates of series expansion 1 and series expansion 2. A. Median unrecovered information \( I \) as a function of the number of terms used in each series expansion, without using spot penalty optimization (solid lines) versus with optimization (dotted lines), and over the three simulation scenarios (panels left-to-right). Each curve was derived from the 100 individual curves corresponding to the 100 simulations in each scenario using a simple point-by-point median average. B. Percentile distribution of the difference between the unrecovered information using series 2 minus the unrecovered information using series 1; the fact that this difference quickly drops below zero in nearly all individual simulations shows that series 2 recovers more information in the first few terms than does series 1.

and with (dotted lines) the use of spot penalties. Each of the 3 panels summarizes all 100 simulated experiments of that scenario, and each experiment in that scenario shows a unique relationship between information recovery and number of series terms computed. Representative curves of individual experiments in each scenario are shown in Supplementary Figure S1 (Supplementary File 1). In order to summarize these very dissimilar curves, Figure 5A shows a median average of all 100 individual experimental curves taken at each data point. Note that this averaging process does not necessarily preserve the shape of the curves from typical individual simulations.

In order to directly compare the two series expansions, we plotted their difference in unrecovered information \( I_2 - I_1 \) versus the number of series terms in Figure 5B. In this case, we plotted the full distribution showing the median (50th percentile) as well as the 10th, 25th, 75th and 90th percentile curves. These plots show directly that series 2 almost always outperforms series 1 when only a few terms can be evaluated. The reason is that the terms in series 2 are larger in magnitude owing to their looser constraints, and thus remove the extraneous part of the partition function more quickly than the terms of series 1 (see Supplementary Figure S1 and Supplementary Figure S4, Supplementary File 1). Based on these results, we recommend using series expansion 2 in all situations where the partition function cannot be evaluated exactly.

**Spot penalty optimization is the most efficient way to recover information.** Spot penalty optimization is an iterative process where each iterative step requires the evaluation of some number of series terms. An optimization requiring \( t \) iterations thus multiplies computation time by a factor of \( t \) relative to the simple evaluation of the series. Alternatively, one could spend the extra computation time on taking the series to a higher order without spot penalty optimization. Figure 6A plots the unrecovered information when a) taking series 2 to a certain order without optimization, versus b) using spot penalty optimization on only the first term yielding \( \hat{Z}_2 \). The dotted line in each panel shows the median number of terms requiring the same computation time as \( \hat{Z}_2 \). The Toy scenario shows that, if the series expansion is carried deep enough, it becomes more accurate than \( \hat{Z}_2 \): in other words the difference \( I - I_2 \) becomes negative. However, for the practical scenarios of Experiments 1 and 2 this crossover point requires taking more terms than would be needed to match the computational cost of calculating \( \hat{Z}_2 \) (the dotted line). Based on these results, we recommend always performing spot penalty optimization, especially for larger reconstructions.
Series expansions can improve optimization information recovery. Although spot penalty optimization is the most efficient way to recover information, that process alone can only extract a certain fraction of the recoverable information: once the cost function is zero, optimization can proceed no further despite the problem not having been solved exactly. At this point, the only way forward is to go higher in the order of series terms used; we can still apply spot penalties to this sum of terms and iteratively optimize them as before using Equation 1 and Equation 2. Figure 6B plots the difference in unrecovered information when applying spot penalty optimization between a) a variable number of terms in series expansion 2, and b) only $Z_0$ (the first series term). This figure shows that including additional series terms in the optimization improves the information recovery, albeit at a slow rate (especially for large problems).

### 20-color labeling leads to near-perfect reconstructions

As shown in Figure 5A, the unrecovered information for the whole-chromosome Experiment 2 averages around 0.2 bits per locus, implying near perfect mapping probabilities. However, because these results were based on randomly-generated unconfined conformations, they do not establish whether such good information recovery is possible with real chromosomes which are likely to be more compact. To test Experiment 2 on realistic chromosome conformations, we generated four plausible conformations of human chromosome 4 by running the GEM software package\(^1\) on the smoothed human Hi-C data set provided by Yaffe and Tanay, 2011\(^1\) and using a 3D spline interpolation to increase the resolution from 1 Mb to 100 kb. These conformations were then virtually labeled at 300 randomly-selected loci with simulated experimental error as before (100/200 nm localization error in xy/z, 10% missing- and extra-spot rates). Mapping probabilities were reconstructed by taking series expansion 2 to the lowest order that included at least 1000 terms, then applying and optimizing spot penalties. Compared with the random-walk conformations used to test the Experiment 2 scenario, these reconstructions did somewhat worse (\(~0.25\) versus \(~0.2\) bits of unrecovered information per locus) owing to fact that physical confinement of chromosomes increases the density of competing spots in the image.

Despite the drop in performance, 0.25 bits of unrecovered information per locus is still an extremely strong reconstruction, implying that the correct locus-to-spot mappings are assigned $p$-values averaging around $2^{-0.25} \approx 84\%$. Starting from such accurate and confident mapping probabilities, one can infer a

---

### Table

<table>
<thead>
<tr>
<th># terms</th>
<th># terms</th>
<th># terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

### Figure 6. Optimization in conjunction with series expansions.

A. Comparison of unrecovered information using series expansions without iteration, denoted $I$, to the unrecovered information obtained by optimizing spot penalties using only the first series term, denoted $I_{opt}^{1st}$, over three experimental situations. Vertical dotted lines indicate the median number of series terms computable with the same computational time as was required to obtain $I_{opt}^{1st}$. For Experiments 1 and 2 the difference $I - I_{opt}^{1st}$ is typically positive at the intersection of the dotted line, indicating that spot penalty optimization method is the more efficient way of recovering information.

B. Comparison of unrecovered information using spot-penalty optimization in conjunction with multiple series terms versus optimization of $Z_0$ alone, showing the added benefit of including more terms in the series.
reasonable conformation simply by assigning each locus to the spot to which it maps with the highest probability (or calling a missing spot if $1 - \sum_{i} p_{I \rightarrow \omega} > \text{any} p_{l \rightarrow \omega}$), and drawing a line in the image that connects these spots in genomic order. The conformations produced by this simple rule are shown in Figure 7: the correct conformation is shown with a blue line and errors in the inferred conformation are shown in red. The reconstructed conformations are $\sim 90\%$ accurate as determined by an alignment between the true and inferred spot sequences traveling along the DNA contour (40, 33, 26, 33 mismatches + indels for the 4 respective experiments). Most mistakes are of a sort that does not change the large-scale structure. For example, one common error is to erroneously skip one or more spots in the image (since at present align3d overestimates the missing-spot rate), thus ‘looping out’ a small part of the conformation and effectively lowering the resolution.

**Discussion**

We have developed and evaluated two improvements to the align3d method for reconstructing chromosome structure. Both of these improve the partition function estimates that determine the locus-to-spot mapping probabilities, which can provide the basis for (probabilistic) reconstructed conformations. The first improvement is a more robust spot-penalty optimizer that allows for large-scale reconstructions involving hundreds of labeled loci, such as will be needed to uncover whole-chromosome conformations. The second improvement is two series expansion formulas for the partition functions, which in principle allow the mapping probabilities to be solved to arbitrary accuracy within the limitations of the experiment and the underlying DNA model. In practice, the series approach is difficult for two reasons: 1) there are a huge number of terms in each series expansion, and 2) the lowest-order approximation $Z_{n}$ overestimates $Z$ by many orders of magnitude, unlike other series expansions where the initial approximation is close to the final answer. Despite the difficulties, the series formulas that we give offer some way forward to improve on the original estimate $Z_{n}$. Of the two formulas, we recommend using series expansion 2, which has the larger-magnitude terms and thus recovers the most information when only a few terms can be evaluated.

Our problem of finding likely (i.e. low-free-energy) DNA conformations passing through a set of imaged spots is similar to the well-known traveling salesman problem (TSP), in which a salesman must find the shortest route connecting a set of cities. Somewhat more closely related is a generalization of the TSP called the time-dependent traveling salesman problem (TDTSP)\(^{10}\), where the intercity distances change every step on the tour; this is analogous to our situation where the free energy needed to thread DNA between two spots depends not only on their separation but also on the length of DNA used to connect them. In our case, the presence of missing and extra spots generalizes our problem still further: in the TDTSP analogy the salesman would be allowed to skip stops and cities for a penalty. Our main finding is that the partition function of this generalized TDTSP (which encompasses traditional TSP and TDTSP problems) can be expressed as a sum of terms computable using a (modified) forward-backward algorithm. This result should also apply to other route-finding applications, and might provide a fresh approach since research has historically focused on estimating the single optimal solution rather than the partition function.

From a genomic standpoint, our most exciting result is that the combination of our computational improvements together with 20-color labeling technology gives almost perfectly-accurate reconstructions. Out of $\sim 4$ bits per locus of uncertainty inherent in the reconstruction problem, our method recovers about 3.75 bits, despite somewhat overestimating the missing-spot rate in these simulations. Such confident mapping probabilities allow for the direct construction of individual conformations that are about 90% accurate. We want to emphasize that our
reconstructions used model parameters that were correctly calibrated, but not necessarily ‘correct’ in terms of details that an experimenter would not know. For example, because the conformations were generated using a wormlike chain DNA model but align3d assumes a Gaussian chain model, the underlying DNA model used in the reconstruction is wrong in a sense, just as it will be using in vivo data where the underlying DNA model is unknown. However, we did set the Gaussian decay constant to correctly match the persistence length as that can be measured experimentally. Additionally, the correct average rates of missing and extra spots over all experiments were provided to the analysis, but using those align3d still had to estimate the actual number of missing spots per color in each experiment. The robustness of the analysis to experimental unknowns gives evidence that reconstructions using real-world experimental data will be of similar quality to those in our simulations, and if so then direct measurement of chromosome conformations is possible today with current technology.

Data availability
The simulated conformations and labelings used to generate the plots in this paper, together with the output of the align3d analysis, can be found at: https://github.com/heltilda/align3d/tree/master/seriesExpansions/a3dRawData.zip

Software availability
Results in this paper were generated using version 1.1 of align3d, built using version 1.1 of Cicada scripting language. Simulated conformations and labelings were generated using version 1.1 of wormulator.

Supplementary material
Supplementary File 1: align3dSupplement.pdf. File containing three appendices giving the derivations of the equations used in this text (Appendix 1 and Appendix 2), and the supplemental figures (Appendix 3).

Click here to access the data

All source files used in preparing this paper are available from the GitHub page for this paper: https://github.com/heltilda/align3d/tree/master/seriesExpansions

License: GPL 3.0
Archived code at time of publication:
align3d: http://doi.org/10.5281/zenodo.1411501
License: GPL 3.0
wormulator: http://doi.org/10.5281/zenodo.1411503
License: GPL 3.0
Cicada scripting language: http://doi.org/10.5281/zenodo.1411505
License: MIT License

Grant information
Funding was provided by the Boettcher Foundation (J.C.C.), National Institutes of Health [2T15LM009451 to B.C.R.], and a Cancer League of Colorado grant (J.C.C. and B.C.R.).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
The authors want to thank Rani Powers and Jenny Mae Samson for helping review the manuscript.

References

Reference Source


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


Open Peer Review

Current Referee Status:  

Version 1

Referee Report 07 November 2018

https://doi.org/10.5256/f1000research.17750.r38643

Ewan Birney 1, Carl Barton 2

1 Global Alliance for Genomics and Health, Boston, MA, USA
2 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

In this paper the authors update and build on a method they have previously published known as 'align3d'. This method attempts to infer the chromosome conformation based on images of fluorescently tagged genomic loci. The authors claim that this updated method increases the accuracy of the inferred conformation as well as allowing the method to run on larger instances of the problem. They then go on to demonstrate where the method allows for the near perfect reconstruction of larger scale, simulated, labelled images. We believe that the article is worthy of indexing on the condition that some minor issues, outlined below, are addressed.

In the introduction the authors mention a couple of other methods attempting to resolve similar problems. I think that this section should be expanded as there is no critical comparison of how this method compares to each of those mentioned. In particular the computational methods should be compared and contrasted so it is clear to the reader how this method differs from others.

Whilst reviewing this paper we were unable to access the supplementary data. This must be made available before the paper can be indexed. Some things the authors could include in the supplementary section that would be useful from the computational perspective would be the type of series expansion being used and any information on how quickly the series expansion converge to the original formula. The authors have experimentally checked on the convergence properties but sometimes it's quite simple to determine theoretically how quickly some approximation converges. This information could be useful in determining better expansions and would explain more concretely why they get some of the results they see.

In the experimental section of the paper the authors generate three different types of simulation that they denote 'Toy', 'Experiment 1' and 'Experiment 2'. In the discussion of the results the authors make the following comment:

‘Use of more colors dramatically improves reconstructions. Our most striking result is that simulations of the ambitious Experiment 2 produce far better results than even the Toy scenario, despite the fact that these simulations have more loci per color than either the Toy scenario or Experiment 1. This can be seen in the amount of unrecovered information shown in the simulation-averaged plots of Figure 5A. Thus a push to 20-color labeling could prove critical for genomic reconstruction at the chromosome scale and beyond. At the end of this section we revisit Experiment 2, in order to assess the reconstruction quality.
when analyzing more realistic DNA contours having tighter confinement.’

The authors should make some attempt to explain this situation. Actually if you increase the number of colours and also increase the number of loci with the same colour then you would not obviously assume that the problem should be harder. It very much depends on how each is increased within proportion to each other.

An increase in the number of unique colours available should lead to the problem being exponentially easier as you are effectively exponentially decreasing the ambiguity in the data set. Should you also increase the number of loci labelled with the same colour then you wouldn’t expect the problem to become harder unless that increase was large enough to outweigh the effects of the increase in the number of available colours. In this sense it could be argued that many instances of the ‘Toy’ example are fundamentally more challenging than the (on the face of it) more complicated ‘Experiment 2’. This should somehow be addressed by the authors.

Also in the discussion of the experimental results the authors note that ‘20-color labelling leads to near-perfect reconstructions.’ This result is consistent with our results reported by Barton et al. (2017)\(^1\). It would be good to mention this as although the computational methods are different, the simulations are generated in different ways and the resolution simulated is different, both methods suggest that if ~20 colours are available then near perfect reconstruction is possible. The authors should also point out the similarity of the number of colours needed in both their method and ours.

The differences in computational methodology yet similarity in the numbers of colours needed for near perfect reconstruction perhaps suggests to me that both methods are in some sense ‘naive’. There must exist a minimum number of colours required for a certain average reconstruction performance (with the appropriate caveats) but we would be surprised if it was as high as 20. It could be interesting to see the authors add some discussion about this connection and any insight they might have into it.

Finally there have been a number of different attempts to simulate super resolved images of the type used in this and other computational methods. If the authors can use this data as input or the data can easily be coerced into an appropriate format for this method then the paper would be much stronger with the addition of results of using the method against these datasets. In this way the authors can clearly demonstrate that the method they propose is not simply good on their own simulated data, but also performs robustly on other independently generated simulations.

References

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Partly
If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

**Competing Interests:** No competing interests were disclosed.

**Referee Expertise:** genomics

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

---

**Discuss this Article**

**Version 1**

Reader Comment 11 Nov 2018

Peter Rogan, University of Western Ontario, Canada

Having had the opportunity to recently see Oligopaint data and speak with a number of practitioners of this technique, reproducibility remains a significant issue. The simulations described in this paper cannot be practically applied for chromosome conformation analysis until convincing data are obtained:

"The robustness of the analysis to experimental unknowns gives evidence that reconstructions using real-world experimental data will be of similar quality to those in our simulations, and if so then direct measurement of chromosome conformations is possible today with current technology."

FISH using short single copy DNA probes is the underlying basis of this technique (PMID 11381034). We successfully labeled single copy oligonucleotides by ligation or hybridization of fluorescent detection reagents, but it was technically challenging (PMID 16460913). Considerable effort is required to ensure low fluorescence background for clinical applications of single copy probes (PMID 12923866), because Cot-1 repeat blocking DNA can produce artifacts in short probe hybridization (PMID 23376933). Expertise in human cytogenetic identification of hybridized chromosomes (e.g. reverse DAPI banding patterns) would also to convincingly demonstrate accuracy of Oligopaint results.

**Competing Interests:** Founder of CytoGnomix. The company holds patents and markets single copy probes.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com