# Using contact statistics to characterize structure transformation of biopolymer ensembles

Priyojit Das,[1] Rosela Golloshi,[2] Rachel Patton McCord,[2] and Tongye Shen[2]
[1]*UT-ORNL Graduate School of Genome Science and Technology, Knoxville, Tennessee 37996, USA*
[2]*Department of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, Tennessee 37996, USA*

(Received 28 August 2019; published 31 January 2020; corrected 15 May 2020)

As a unique subset of functional polymers, many biopolymers have a set of well-defined three-dimensional (3D) structural characteristics that can be described by spatial contacts between monomers. Statistical analysis of the contacts has been extremely productive in characterizing the biopolymer structural ensemble, such as for 3D chromosome structures. Often, native contacts and compartment structures are the focus of the studies, while the generic polymer aspect, such as the overall decaying of contacts with increasing sequence distance, is analyzed separately or preemptively removed. Here, we explore insights that can be gained by performing "compartment analysis" that keeps the distance decay, which we believe is particularly useful for characterizing the structure transformation of biopolymers. We tested contact analysis on several such transformations under physical perturbation or biological processes, including (1) unfolding of proteins induced by thermal denaturation, (2) chromosome conformation transition during the cell cycle, and (3) chromosome unpacking by physicochemical perturbations. Useful score functions were developed to further quantitatively characterize the transformation judging from the contact analysis. We also find that the sinusoidal undertone of eigenvector patterns (the "unwanted," low frequency signal, in contrast to the detailed A/B compartment) that had previously been attributed to biological effects of centromere proximal and distal interactions may in fact reflect a universal feature of polymers that have relatively weaker long-range contacts.

DOI: 10.1103/PhysRevE.101.012419

## I. INTRODUCTION

Many linear biopolymers, including well-understood examples of proteins and nucleic acids, are structured or semistructured polymers, which means that they have defined structure(s) or at least well-defined structural components [1,2]. Often, their three-dimensional (3D) folded conformations are essential to fulfill their biological functions, and their structural features separate them from generic polymers. How these biopolymers are able to fold and sustain specific shapes has been an important question in biological physics research. The essential features are encoded by short-ranged favorable contact interactions between monomers that are not sequence neighbors. At the simplest level, even a model of "two-colored" beads on a string reproduces basic features of structure formation for biopolymers [3,4]. In protein folding, this type of model reflects a basic hydrophobic-polar (HP) dichotomy where H-H contact interactions are favorable compared to that of H-P [5]. In chromosome folding, this two-color model reflects a dichotomy of gene-rich/gene-poor regions (A/B compartments [6]) where gene poor or gene rich regions tend to aggregate with like types (e.g., favorable B-B and A-A interaction compared to A-B interaction) [7]. With a more refined biopolymer model containing bead types that accurately reflect the specific physical interactions between different parts of the polymers, other complicated and vital-for-function structures emerge.

How one should characterize the conformations of these structural ensembles is an essential question for the physical description of the biopolymers and the interpretation of their biological functions. For the structural ensemble of a highly structured biopolymer, the structural fluctuation can be characterized similar to a "vibrating" solid [8,9], where a set of Cartesian coordinates is conventionally used as the degrees of freedom (DOFs). At the other limit, a generic polymer that does not have specific non-neighboring interactions are best characterized by internal coordinates of bonding terms (bond, angle, and torsion angle) [10]. In between, when one wants to study a biopolymer's transformation, for example, from a structural to a disordered state, it is more suitable to use contact DOFs [11,12]. The concept of contacts (spatial proximity formed between different parts of the polymer) is frequently used to characterize such (semi)structured conformations, and these contact interactions are often collectively displayed in the form of a matrix. In the case of chromosomes, such contact matrices often are derived from genome-wide chromosome conformation capture (Hi-C) data.

Statistical analysis methods, especially principal component analysis (PCA), have been used for studying the collective "modes" of biopolymer contacts. For example, a popular contact analysis method, known as the "compartment analysis" in the chromosome field, was designed to characterize and even amplify the "signature" of the well-folded structures [6,13]. On the other hand, studying the more flexible or even somewhat disordered structural ensembles can be important and have practical biophysical applications. Researchers often perturb the well-folded biopolymers with various environmental perturbations to examine how their structural stability and dynamics alter, eventually moving these biopolymers out of their native conformations and into an unfolded state [14–17]. Also, many biopolymers, from intrinsically disordered proteins to whole chromosomes at certain points in the cell cycle, may be viewed as weakly interacting polymers (with significantly less specific long-range contacts) that approach a more generic polymer state. Thus, it is desirable to

 

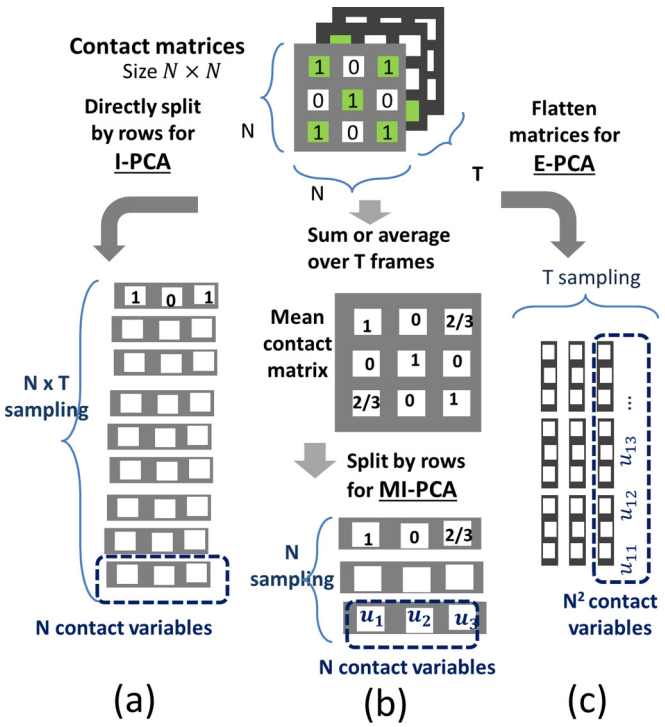


FIG. 1. A cartoon illustration of three different analysis methods, (a) I-PCA, (b) MI-PCA, and (c) E-PCA, on contact matrices.

use contact analyses not only for characterizing the ultimate structure aspect (the "enthalpic" aspect) of these biopolymers, but also for accessing the generic polymer aspect of the ensemble.

The schemes of using principal component analysis for studying contacts discussed in this work, E-PCA, I-PCA, and MI-PCA, are recapitulated in Fig. 1. As we have described previously, contact fluctuation analysis, E-PCA, characterizes the dominant fluctuation of the structural ensemble, while contact structural analysis, I-PCA (and its slight variation version MI-PCA) describes the consensus features of the ensemble such as domain and compartment classification [18]. Specifically, the matrix form of the contact information is used for these analyses. The column and row indices of these matrices are naturally derived from the linear indexing of the polymer. Contact interaction strength $u_{ij}$ has a value of 1 if contact formed between monomers $i$ and $j$, and 0 otherwise. A simple definition based on distance $R_{ij}$ provides $u_{ij} = \Theta(R^C - R_{ij})$, where $R^C$ is the cutoff. One needs further define $U_{ij} = \langle u_{ij} \rangle$ as the ensemble average of the contact interaction strength, also known as interaction frequency.

E-PCA treats each contact *explicitly* as an independent variable $u_{ij}$ and, for a matrix of $N \times N$, it largely contains an order of $N^2$ variables varying from one matrix to another. Here, $N$ represents the dimension of the system, i.e., the number of basic building blocks of that biopolymer. For proteins, these building blocks can be amino acids [19] or coarse-grained segments that may even contain secondary structural elements [20], and, for chromosomes, the building blocks are usually genomic regions: bin size ranges from $10^3$ to $10^7$ base pairs (bp), with typically used values from $10^4$ to $10^6$ bp. In general, E-PCA has a covariance matrix ${C^E}_{ijkl} = \langle (u_{ij} - \langle u_{ij} \rangle)(u_{kl} - \langle u_{kl} \rangle) \rangle$. In contrast, for I-PCA and MI-PCA, which are methods frequently used for studying chromosome contact matrices, one treats each row of contact matrix as a set of independent data and one has total N *implicit* contact variables $u_i$ that vary from row to row, ${C^I}_{ij} = \langle (u_i - \langle u_i \rangle)(u_j - \langle u_j \rangle) \rangle$. Specifically for MI-PCA, we first obtain contact average over rows $\langle U_j \rangle = \sum_i U_{ij}/N$ and a matrix $\delta U$ defined by its elements $\delta U_{ij} = U_{ij} - \langle U_j \rangle$. Then, the covariance matrix for MI-PCA $C^I = \delta U^T \delta U$ is obtained. Once the covariance matrix $C^I$ or $C^E$ is obtained, principle component analysis (PCA) is applied to obtain the collective modes of fluctuation from the corresponding mean contacts $\langle u_i \rangle$ or $\langle u_{ij} \rangle$. These modes of variation for $\langle u_i \rangle$ reveal the domain information of the ensemble while variation modes of $\langle u_{ij} \rangle$ display dynamics.

## II. THE GENERIC POLYMER LIMIT AND THE PROBLEM OF REMOVING DISTANCE DECAY IN CONTACT MATRICES

For chromosome contact data and particularly for the detection of A/B spatial compartments, instead of direct analysis of the "raw" contact interactions $u_{ij}$, often researchers manipulate the matrix to enhance the weight of the off-diagonal components (long-range contacts) and obtain an alternative definition of contact interaction, $v_{ij} = u_{ij}/b(|i-j|)$. Here, scaling factor $b$ is a function of genomic distance and scales the raw contact data $u_{ij}$. It was initially designed to preemptively "subtract" the random, generic polymer nature of the contact matrix and amplify the structured polymer aspect of these biopolymers [21–23]. Even though some early analysis recommended eigenvector decomposition of contact matrices without such modification [24], most applications of PCA to chromosome contact matrices (especially intrachromosomal contacts) have continued to include such an operation (some of the recent examples can be found in [25–27]) and the differences between the results obtained with and without this normalization have not been thoroughly explored, to our knowledge, in any previous work.

It is nontrivial to model factor $b(|i-j|)$, and various function forms have been tested to best serve the signal enhancing purpose. There are largely two types of approaches. One is using a derived or empirical function form, and previous studies on generic polymers have been quite extensive [28–30]. For example, factor $b$ be can a power law form $|i-j|^{-D/2}$ or an exponential form $\exp(-\alpha|i-j|)$. The other enhancing approach uses the statistics of raw data $u_{ij}$ itself. The power law form was derived from the generic polymer limit, in which case the distribution of the end-to-end displacement $\vec{r}$, according to the random walk theory [29], is $p(\vec{r}; L)d\vec{r} = (2\pi La^2)^{-3/2} \exp[-3r^2/(2La^2)]d\vec{r}$, where $L$ is the number of basic building blocks and $a$ is the linear length of the building block unit. This equation can be generalized in a general $D$-dimensional polymer as $p(\vec{r}; L) = (2\pi La^2)^{-D/2} \exp[-Dr^2/(2La^2)]$. Thus, for a 3D random polymer ($D = 3$), the chance of residues $i$ and $j$ forming a contact can be viewed as a loop or a polymer cyclization problem, where the chance of a random walker revisiting the same location $r = 0$ after distance $L = |i-j|$ steps is $p(0; i-j)$, i.e., $p_{ij} \sim |i-j|^{-3/2}$. This random walker model can be used to describe the simplest case: an ideal generic

polymer in a theta solvent. The practical situations can be more complicated. For example, a poor solvent condition packs the polymer, while on the other hand polymers in good solvent are more extended. Researchers found that a fractal dimensional biopolymer [30] might be better for the hierarchical structure of chromosome fold [21]. Using the generalized form of random walk models, one obtains a power law decay $p(0;L) \sim L^{-D/2}$. Practical fitting for chromosome data even provides varying $D$ depending on $L$, such as $D \approx 2$ for longer $L$ and as small as $D \approx 1.5$ for shorter $L$ [31]. Other variety of empirical function forms with a gentler decay have been used as well, such as exponential form $p(L) \sim \exp(-\alpha L)$, which has the advantage of smoother short length scale. However, such exponential formula can also be limited by its applicable range [31].

As shown in Fig. 2(a), we calculated the top eigenvectors of MI-PCA using a directly generated mean contact matrix for exponential and power law decay and a computer simulated random polymer. One can observe that the results are always a symmetric odd function with respect to the middle point. The components of MI-PCA eigenvectors always have mixed signs, i.e., some are positive, while others are negative, if the diagonal elements of the matrix are larger than the off-diagonal elements. Practically, this is always true for a polymer's contact matrix since the contact frequencies of sequence neighbors are much higher. The location of sign change can be defined as a node or a boundary. Often these nodes for a structured polymer are important and can be used to separate biopolymers into distinct structural domains, often called A/B compartments for chromosomes [32]. In contrast to MI-PCA, the elements of an E-PCA eigenvector can potentially have the same sign, i.e., a global "breathing" motion that have contacts forming and breaking simultaneously.

Besides mean contact matrices constructed using exact mathematical formulas of the exponential or power law decay forms, we also obtain the mean contact matrix using computer simulation data of a 65-bead generic polymer that lacks specific contact interaction. For a fair comparison of polymers with different number of beads, we can project the result on the ideal mathematical formulation-based results of a 100-bead polymer using a scaling transformation: $x^* = x \times (100/65)$ and $y^* = y \times (100/65)^{-1/2}$. This random polymer simulation is a limiting case of a protein polymer simulation where the native contact interactions are removed. The rationale for scaling to account for the difference between different sizes of the mean contact matrix will also be given in the next section, where we discuss the conformations of a 65-residue protein polymer under thermal perturbation. Our simulation results matched quite well with the ones generated by mathematical formula, particularly, the exponential decay ones [Fig. 2(a)].

From Fig. 2(b), we also can observe that high order eigenvectors (those with a smaller eigenvalue) have more nodes and in general, the $n$th mode, MI-PC $n$, contains $n$ nodes at the generic polymer limit. It is interesting to point out that the eigenvectors MI-PC1 to 5 resemble an expansion of cosine series, which is not accidental. In fact, one can locate a family of matrices having their eigenvectors being exactly a series of trigonometric expansion. The eigenvectors of Rouse polymer matrix $A$ provide such an instance, where matrix $A$ is essentially a second-order differentiation operator [33,34].

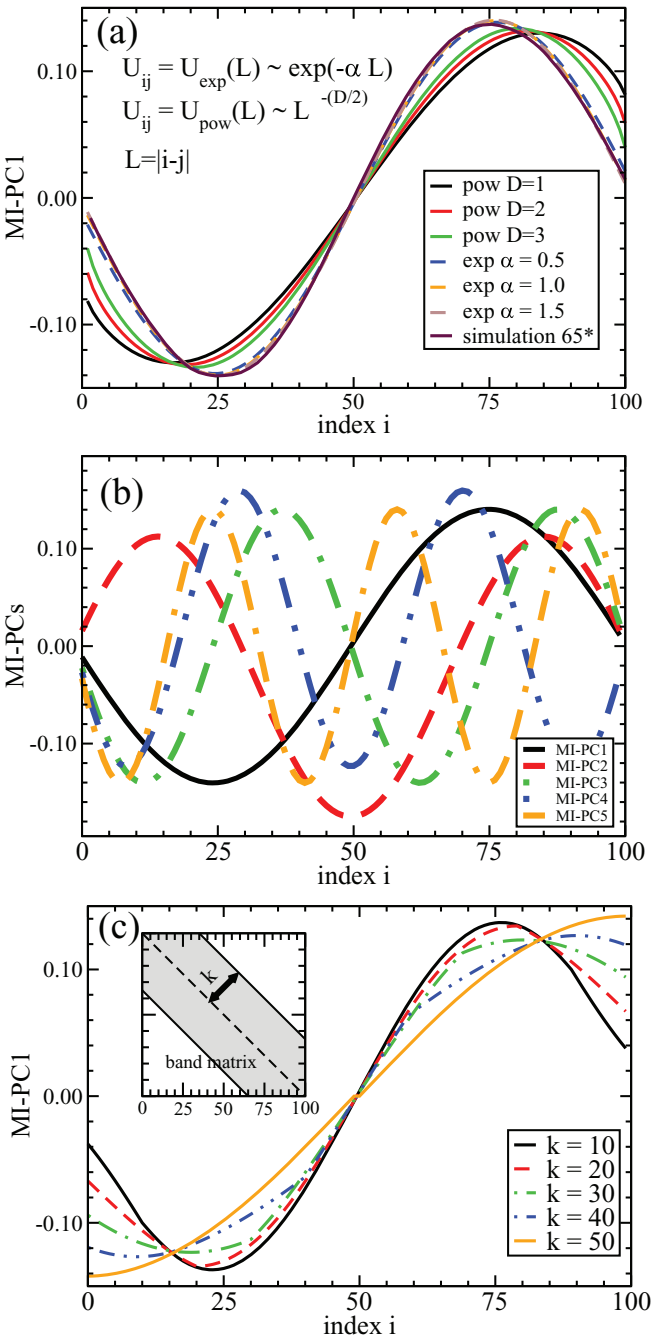


FIG. 2. (a) The dominant mode of MI-PCA for a random polymer, MI-PC1, including power law form of mean matrix, exponential decay form of mean matrix, and a result of computer simulation of the polymer. (b) The top five eigenvectors of MI-PCA for the case of the exponential decay with $\alpha = 1.5$. Eigenvectors with smaller eigenvalues have increasing number of nodes. (c) The top eigenvector of a band matrix whose nonzero elements are uniform varies with changing bandwidth $k$.

In general, matrix $B = A + cxI$ ($I$ being identity matrix and $c$ a constant) shares the same eigenvectors with matrix $A$. Practically, all mean contact matrices with much stronger diagonal component and a very fast off-diagonal decay will have a covariance matrix $C^I$ with similar characteristics.

We also examined the MI-PC1 signature of another type of contact matrix, band-diagonal matrices [Fig. 2(c)]. One can

see that with increasing band width, the signature of MI-PC1 deviates from the eigenvectors of the "ideal" non-interacting polymers displayed in Fig. 2(a). As we will demonstrate below, some of the features of the deviation resemble certain observations of mitotic chromosome structures.

Because it is not clear which function form is best at removing the random polymer "background feature" in practical situations, the alternate common approach of treatment is normalization by an empirical average or median contact count at each length scale [21–23,35]. The common approach is directly normalizing by the average of the off-diagonal, more precisely, the average of k-diagonal elements, where k is a measurement of its relation to the main diagonal, i.e., $v_{ij} = u_{ij}/b_k$ with $b_k = \sum_{ij} u_{ij} \times \delta_{|i-j|-k} / \sum_{ij} \delta_{|i-j|-k} = \sum_i u_{i,i+k}/(N-k)$. Here Kronecker delta selects $k = |i-j|$. This way, the strength of matrix elements $v_{ij}$ will not drop off as that of $u_{ij}$ with increasing $k$. One drawback for this practical approach is that, when $k$ is large, that is for those extreme super-diagonals and sub-diagonals, there are few matrix elements to contribute to the average.

While pre-treating a contact matrix can be useful to emphasize off-diagonal interactions, it has a few potential disadvantages. First, as we will demonstrate below, such treatment will remove or at least reduce the aspect of polymer (thermal) fluctuation, and thus make the examination of polymer conformations subjected to physical and chemical perturbations less direct. Secondly, the treatment will create false positive information on domain structure when the polymer is near the generic polymer limit. The reason is straightforward. At the generic polymer limit, a truly fully sampled $u_{ij}$ will provide all elements $v_{ij} \equiv 1$ and thus, return a zero valued covariance matrix and thus a uniform domain. However, in any practical situation, there will be at least numerical noise that make $v_{ij}$ less than perfect and thus create false-positive domains in the now nonzero covariance matrix. Thus, we find that treated matrices can be useful to reveal changes in folding patterns when the "signal" is strong enough, but they may not be best when one wants to compare a broad spectrum of conformation ensembles, as we will demonstrate below using specific examples.

It is also important to point out that, throughout this work, the pretreatment by scaling ($u_{ij} \rightarrow v_{ij}$) refers to off-diagonal distance normalization. This operation must not be confused with many correction schemes to remove biases from raw Hi-C data [24,36–39]. For example, the correction to remove experimental artifacts (which arise from Hi-C steps involving sequence-specific DNA cleavage and differential PCR amplification of sequences with different GC content) is achieved by iterative correction and eigenvector decomposition (ICE) normalization. This ICE approach balances the matrix row and column sums by transformation $u_{ij} = \frac{u'_{ij}}{\beta_i \beta_j}$, where $\beta_i = \sum_k u'_{ik}$ and $u'_{ik}$ is the raw number of contact hits [24,38].

## III. CONTACT ANALYSIS FOR PARTIALLY STRUCTURED PROTEINS AND CHROMOSOMES

Here we provide examples of biopolymer data that may exhibit a large range of dynamic (folding) motions along a spectrum of structure order, from relatively structured to the effective generic polymer limit, without specific contacts. The first example is a model protein system subject to increasing thermal noise which undergoes an unfolding transition. The second example is chromosome changes during the cell cycle, where chromosome structure dramatically reorganizes from interphase to metaphase. Last but not the least, we examined the less structured ensemble resulting from a physicochemical perturbation of chromosome structures, in which chemicals such as salt and histone deacetylase inhibitor drugs act to decondense chromosomes, which might destabilize chromosome structures and push them towards the direction of losing specific contacts.

### A. Proteins approach the generic polymer limit with rising temperature

The first example is a structural contact analysis of a small protein (65 amino acid residues), chymotrypsin inhibitor CI-2 (PDB ID: 2CI2) [40] subjected to thermal perturbations. The advantage of using simulation data is that they provide information about individual structural fluctuation and we can perform E-PCA for a comparison with MI-PCA. They also provide direct 3D coordinate information that is not available for chromosome structures at high resolution. Here, we use MI-PCA for this pedagogical system to demonstrate how partially folded protein conformations can be characterized by contact statistics and how the results change with increasing temperature. CI-2 has a simple fold of several secondary structure elements (four $\beta$ strands and one $\alpha$ helix), as shown in Fig. 3(a). We adopted a simple coarse-grained Hamiltonian (C$\alpha$ Go model) of the protein that has frequently been used to study protein contact dynamics and folding [41,42]. The configurations were sampled from the molecular dynamics simulation package GROMACS [43] and setup program SMOG [44]. The simulation timestep $dt$ is 0.0005 and total $2 \times 10^8$ steps were used for each simulation. Snapshots were taken at every 10 000 steps. Finally, 20 000 snapshots were converted to contact matrices. Here a simple distance scheme is used, i.e., a contact is being formed and $u_{ij} = 1$ when the distance between C$\alpha$ atoms of a pair of amino acid residues $r_{ij}$ is less than a distance cutoff 6.5 Å, and $u_{ij} = 0$ otherwise. The contact matrix of the crystal structure using this definition is shown in Fig. 3(a).

We display the results of I-PCA (particularly, MI-PCA) and E-PCA for this system at various temperatures ($T =$ 0.5, 1.0, 4.0, and 10.0 $T_f$) in Figs. 3(b) and 3(c), respectively. Here, temperature $T$ is expressed in the unit of the folding temperature $T_f$ at which point the protein spends roughly 50% of time in the native, folded state and the other 50% unfolded. For a lower temperature $T = 0.5\,T_f$, we observed that MI-PC1 characterizes the mean contact map with clear domain features, while the corresponding E-PC1 describes the internal contact breaking and forming dynamics of this folded structure. At $T = T_f$, we see that the structural features revealed by MI-PCA are weakened. At even higher temperatures, the structural features disappear and the results approach that of the generic polymer limit shown in the previous section. Note that these high temperatures are not meant to be taken literally. Rather, they are used to demonstrate the effect of weakening native interactions. An interesting observation is that, even at

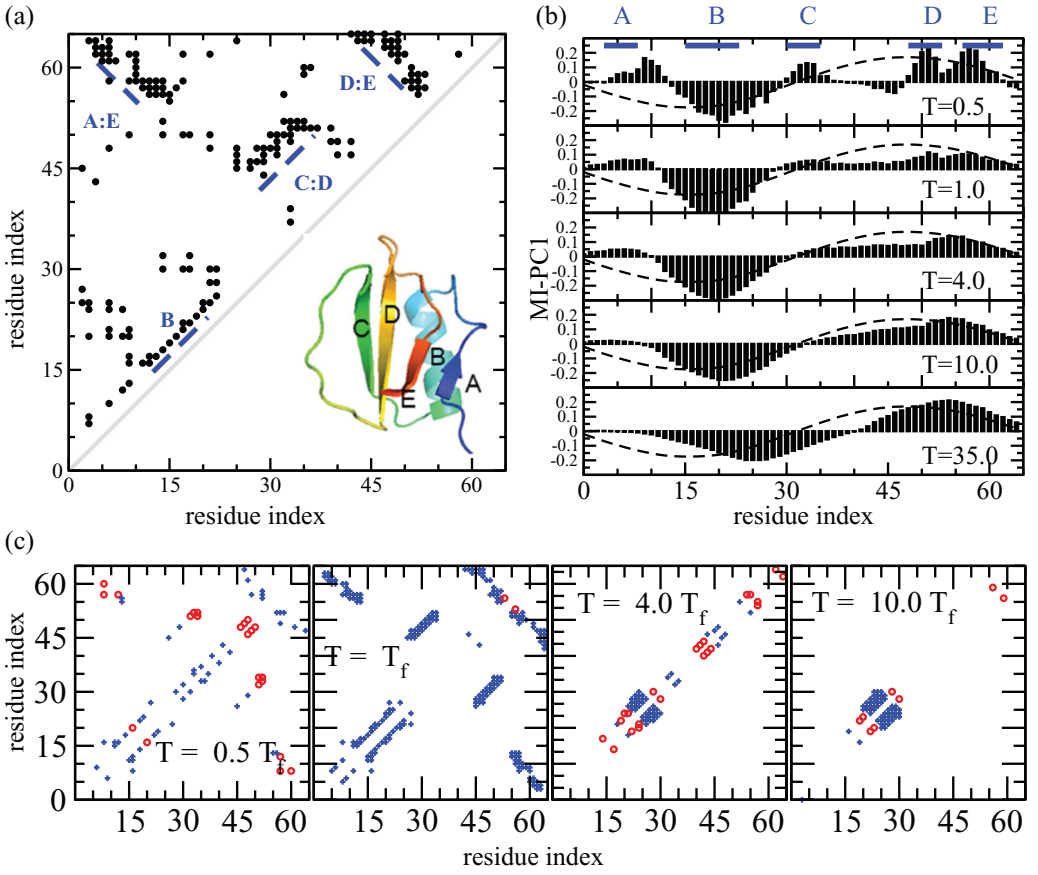


FIG. 3. (a) Structure and contact map of protein CI2. (b) MI-PC1 of protein CI2 at different temperatures. The dashed line is the reference system of backbone interaction only. (c) The corresponding E-PC1 of protein CI2 at different temperatures are shown using its significant elements (a cutoff of $|d_{ij}| > 0.025$).

high $T = 35\ T_f$, IPCA still reveals noticeable deviation from the ideal polymer behavior. Meanwhile, one could not easily distinguish such ensemble from a random unfolded ensemble by direct visualization, which indicates the sensitivity of the I-PCA method.

For a comparison and a demonstration of what E-PCA reveals, we show the top dynamic modes at the corresponding temperatures. Note the drastic differences of contact structural analysis I-PCA and contact dynamics analysis E-PCA's eigenvectors. One is expressed essentially by a linear plot, i.e., 1D function $a_i$ while the other is often expressed in the form of the matrix $a_{ij}$ that symbolizes a particular fluctuation mode of contact dynamics. For example, $T = 4.0\ T_f$ E-PC1, the dominant mode of fluctuation, displays the forming/breaking of the $\alpha$ helix, which further nucleates with the rest of the secondary structures around $T_f$, which is consistent with experimental and theoretical studies [45,46].

### B. Mitotic chromosomes approach the effective generic polymer limit during the cell cycle

Chromosomes are semistructured biopolymers which change their conformations and compactness depending on the cell cycle stage [47]. The transition from interphase chromosome structure through prophase and into metaphase was recently captured in detailed Hi-C contact maps for synchronized DT-40 chicken cells [48]. In G2 ($t = 0$ min), before the cell enters mitosis, chromosomes show all the typical hallmarks of highly organized interphase structure, including topologically associating domains (TADs) [49,50] and A/B compartments [6]). As the cell progresses through mitosis, these conformations start to disappear. Upon entering prophase [$t = 2$ min after cyclin-dependent kinase 1 (CDK1) checkpoint release], compartments and TADs are lost, and the chromosomes start to become linearly organized structures. The chromatids become more shortened and thicker as the cell enters prometaphase ($t = 10$ min) and ultimately lead to the formation of fully condensed metaphase chromosomes. Here, we examine the contact matrices of the 39 chicken chromosomes at seven time points ($t = 0$, 2, 5, 7, 10, 30, and 60 min) starting from G2 phase to late prometaphase, binned at 40 kb and preprocessed with iterative correction [48].

We display the MI-PC results of chicken chromosome 21 (chr21) in Fig. 4. As shown, chr21 is quite structured at the G2 phase ($t = 0$ min). But as the prophase progresses, the MI-PC1 signals start to resemble that of a generic noninteracting polymer. From these observations, it can be concluded that chr21 transitions from well folded (highly ordered structure with specific contacts) to a linearly organized "effective polymer" (fiber) that lacks specific long-range contacts during the mitotic phase. It is important to discuss the structural nature of the mitotic chromosome here. As demonstrated by the

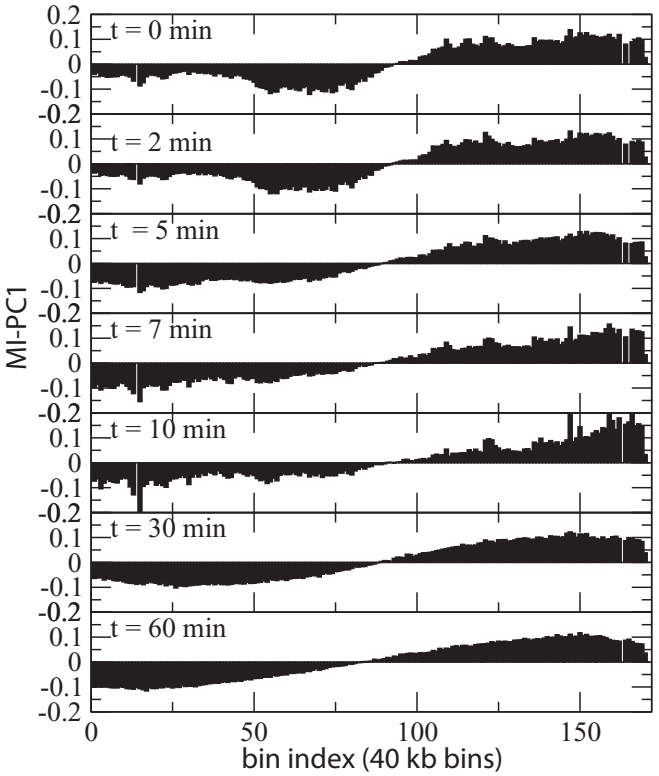


FIG. 4. MI-PC1 of chicken chromosome 21 (chr 21) at time-points progressing from G2 to late prometaphase.

imaging results at $t = 60$ min in Ref. [48], chromosomes are shown as thick fibers with a thickness of 1 $\mu$m. It is important to point out that the close match to the MI-PCA curve of a generic noninteracting polymer during metaphase does not mean that the structure behaves as a linear DNA polymer at a much finer resolution. Instead, the compacted metaphase fiber as a whole acts as a coarse-grained polymer without specific distal contacts, while a sophisticated local structural model of regular loops being packed along a spiral "vine" to form such fiber was shown [48]. Also, it is noteworthy that one can observe the underlying sinusoidal undertone throughout the time points of the cell cycle monitored here, even at $t = 0$ min where the chromosomes are still quite structured. Such undertone indicates the relatively weaker long-range contacts compared to the near diagonal ones.

### C. Chromosomes maintain detailed structure under salt and drug treatments

We next asked whether chemical perturbations to chromosomes, analogous to increased temperature in protein simulations, would cause their structure to approach a more generic polymer state. We have examined two types of perturbations with Hi-C: changes in chromatin compaction via histone modifications and changes in cation concentration [51] (high and low salt). We find overall that these perturbations did not unpack the chromosome to the level of an effective generic polymer. However, we note that this approach can distinguish the degrees of partial chromosome folding that otherwise would look nearly identical in a distance decay normalized compartment analysis.

Histone deacetylases (HDACs) directly affect local chromatin structure by removing acetyl groups from histone tails and in turn making chromosomes more packed [52]. Their inhibitors, HDACi, such as trichostatin A (TSA) allow the chromosome to exist in a more acetylated state, an open conformation which is more easily accessible to transcription factors [53]. The decondensation of chromosomes by TSA has been shown to affect the overall physical stiffness of nuclei [54], but it is not clear how it affects the polymer properties of chromosomes as measured by Hi-C contacts. We have analyzed the Hi-C chromosome contacts of cells treated with 0.5 $\mu$M TSA for 2 hours, a time point at which local histone modifications can be detected [53]. As demonstrated in Fig. 5 using human chr21, MI-PC1 is heavily structured under these perturbations and does not resemble the signature of a generic polymer. However, one can clearly identify interesting regions (such as 21q22.2-21q22.3) that show significant differences in MI-PCA profiles that is not noticeable with a decay-normalized analysis. A full discussion of the important biological implications of these specific changes are beyond the scope of the current study, whereas here we emphasize that applying MI-PCA without matrix pretreatment reveals variations between two structures that are not apparent in traditional "compartment analysis."

Perturbations such as altering concentration of salts may also affect the level of packing [51]. For example, depleting salt perturbs the electrostatic interaction, leading to stronger electrostatic repulsion and swelling of the chromosomes. On the other hand, stronger electrostatic shielding at high salt concentration will compact chromosomes [55]. However, exploratory study of MI-PC1 analysis of Hi-C contacts under these perturbations shows that higher order topology is maintained, and these chromosomes do not behave more like a generic noninteracting polymer (data not shown).

### D. Cross comparison using score functions

We demonstrate here that score functions based on the MI-PC1 can further characterize the generic polymer nature (or lack thereof) of biopolymers such as proteins and chromosomes. Before we present the application with concrete data, we first introduce two empirical score functions (based on MI-PC1 as the input) to measure how close a particular conformational ensemble is to the two limiting cases: at one end, the native, folded structure and at the other end, the generic, noninteracting polymer.

For comparing whether a structure is close to a specific folded structure, we must use information from that specific structure as a reference state, as there is not a universal structural signature for all structured polymers. Specifically, we define a structure deviation score $R_{\rm str} = \sum_i (a_i - a_{i,\rm ref})^2$. Here the reference state is chosen as the MI-PC1 of a highly structured ensemble. For example, for protein ensemble, we can use the low temperature limit as a reference state. A high value of $R_{\rm str}$ may indicate a large deviation from that structure. On the other hand, we would like to define how an ensemble is deviated from to the presumed generic polymer limit with no specific structures. Here we define an

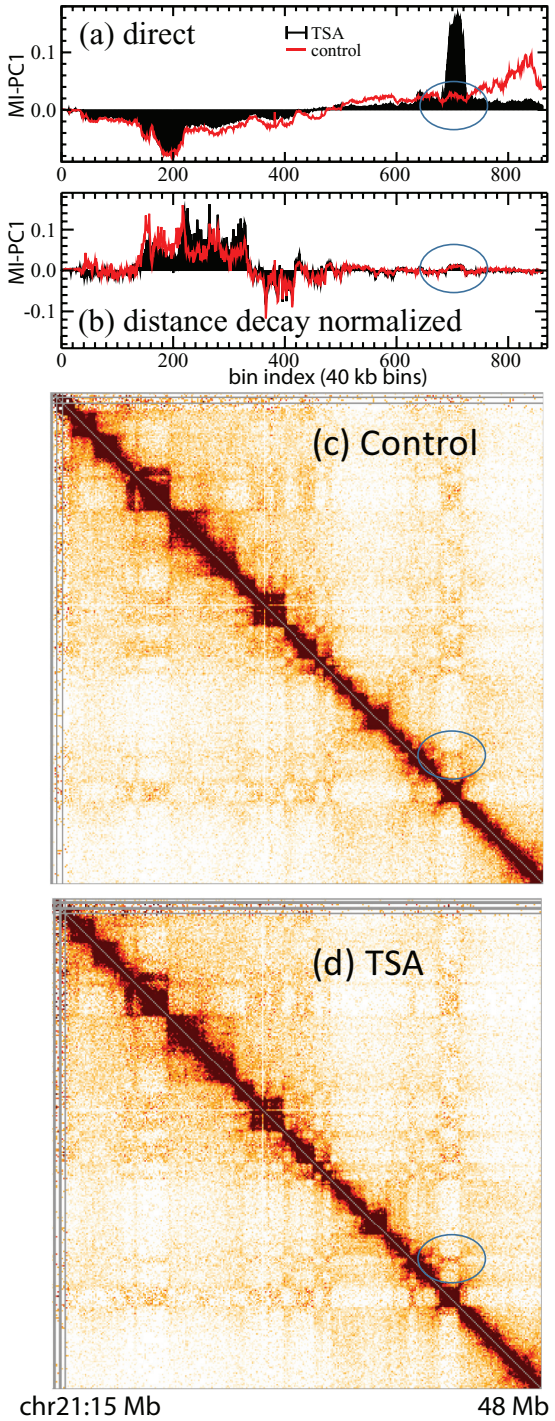


FIG. 5. The comparison of drug (TSA) treated human chromosome 21 and the corresponding control using direct contact analysis, i.e., MI-PCA on $u_{ij}$ (a) and the conventional approach with the distance-decay scaled (b), i.e., MI-PCA on $v_{ij}$. Here the bin size used for the analysis is 40 kb. The corresponding contact interaction matrices $u_{ij}$ (ICE normalized, but no distance normalization applied, bin size 100 kb) for the control (c) and TSA treated (d) chr21. The region where direct contact analysis detected a difference between TSA and control is highlighted with ellipses.

*ad hoc* function, termed the interacting polymer score, $R_{ip} = \sum_{i=1}^{N-1} (a_{i+1} - a_i)^2 \times \frac{N^2}{4\pi^2}$, which works out well practically. Here, the linear size of the chromosome contact matrix, N is defined as the ratio of the total size of chromosome over the bin size. The rationale behind the resolution-free definitions of $R_{ip}$ and $R_{\rm str}$ is as follows. Ideally, we want our score functions to be independent of how we choose the bin size. When we perform the MI-PCA, we can describe the normalized top eigenvector with its components $a_i$ with $i = 1, \ldots, N$. However, this definition of $a_i$ is not resolution free and depends on bin size. By the definition of a normalized eigen vector, we have $\sum_{i=1}^{N} a_i^2 \equiv 1$, so with increasing $N$ the value of element $a_i$ decreases as $1/\sqrt{N}$. This point was well demonstrated in the transformation of $y^* = y \times \sqrt{65/100}$ in Fig. 2(a), so we can project I-PCA results of a hypothetical 100-bead ideal polymer using a 65-bead polymer simulation. Thus, a resolution-independent definition of normalized eigenvector is necessary. Thus, we define a transformation $a_i \to \tilde{a}(x)$ which makes $\int_{-\pi}^{\pi} \tilde{a}(x)^2 dx \equiv 1$, where we define

$$x \equiv 2\pi \times (i/N) - \pi. \tag{1}$$

Here, we map the discrete index range $i \in [1, N]$ to the normalized continuous range $x \in (-\pi, +\pi]$. We choose to use $\pi$ in our definition of a normalized range because, as demonstrated above, the IPCA eigenvectors of polymers approaching the noninteracting limit have the nature of a cosine function series. Comparing the normalization condition in the integral form and in the discrete sum form and since $dx \to \Delta x = 2\pi/N$, we can define a new resolution-free function

$$\tilde{a}(x) \equiv a_i \times (2\pi/N)^{-1/2}. \tag{2}$$

Finally,

$$\begin{aligned} R_{ip} &\equiv \int \left(\frac{d\tilde{a}(x)}{dx}\right)^2 dx = \int \left(\frac{\tilde{a}(x+\Delta x) - \tilde{a}(x)}{\Delta x}\right)^2 dx \\ &= \sum_i \left(\frac{a_{i+1} - a_i}{\Delta x}\right)^2 \times (2\pi/N)^{-1} \times \Delta x \\ &= \sum_i (a_{i+1} - a_i)^2 \times (2\pi/N)^{-2}. \end{aligned} \tag{3}$$

Thus, we have a factor $(N/2\pi)^2$ in the definition of $R_{ip}$. This factor essentially accounts for the slope being steep in the normalized function $\tilde{a}(x)$ since the length is $2\pi$ instead of $N$. On the other hand, for $R_{\rm str}$, we do not need to explicitly compensate when comparing different resolutions $N$, since

$$R_{\rm str} \equiv \int (\tilde{a}(x) - \tilde{a}_{\rm ref}(x))^2 dx = \sum_i (a_i - a_{i,\rm ref})^2. \tag{4}$$

We first apply our scoring functions to the protein folding of CI-2 at different temperatures. As shown in Fig. 6(a), with rising temperature the polymer gradually loses long-range contacts (measured by a smaller interacting polymer score) and gradually loses its nativeness (measured by an increasing of deviation from the native reference MI-PC1). Here the native structure reference state is defined by $T = 0.5\, T_f$. For the second example of chromosomes during the cell cycle, we can measure the corresponding properties for chr21, and we obtain an interesting curve with the evolution of time [Fig. 6(b)]. It first decreases the structural order by losing native contacts

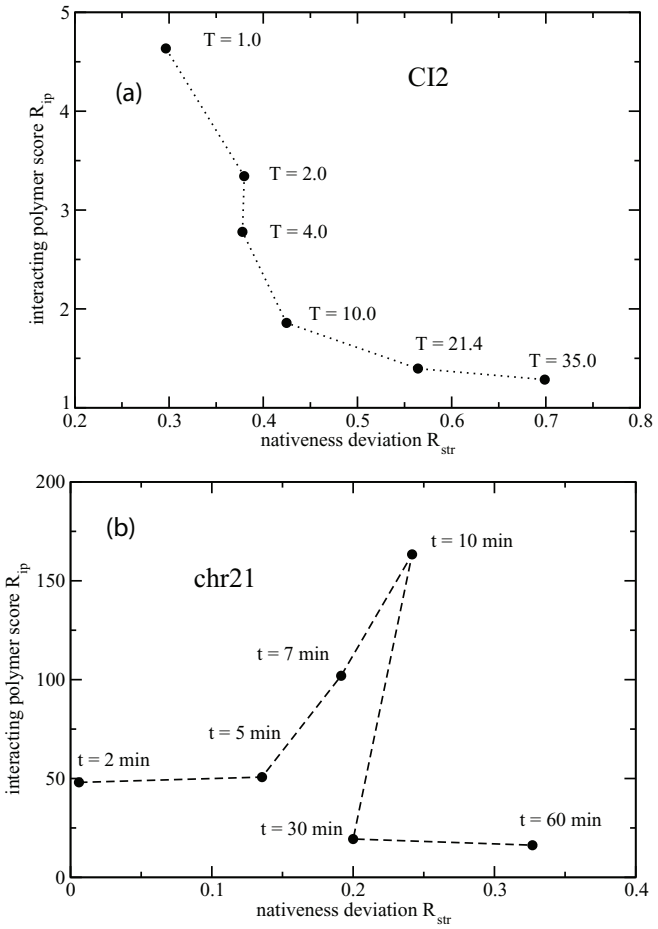


FIG. 6. (a) Protein CI-2 at different temperatures characterized by two parameters $R_{\rm str}$ and $R_{ip}$, where the $x$ axis measures against a specific folded structure at the low temperature limit and the $y$ axis represents how well the structure resembles the features of the corresponding generic polymer. (b) Chicken chromosome 21 at different time points during the cell cycle characterized by two parameters $R_{\rm str}$ and $R_{ip}$.

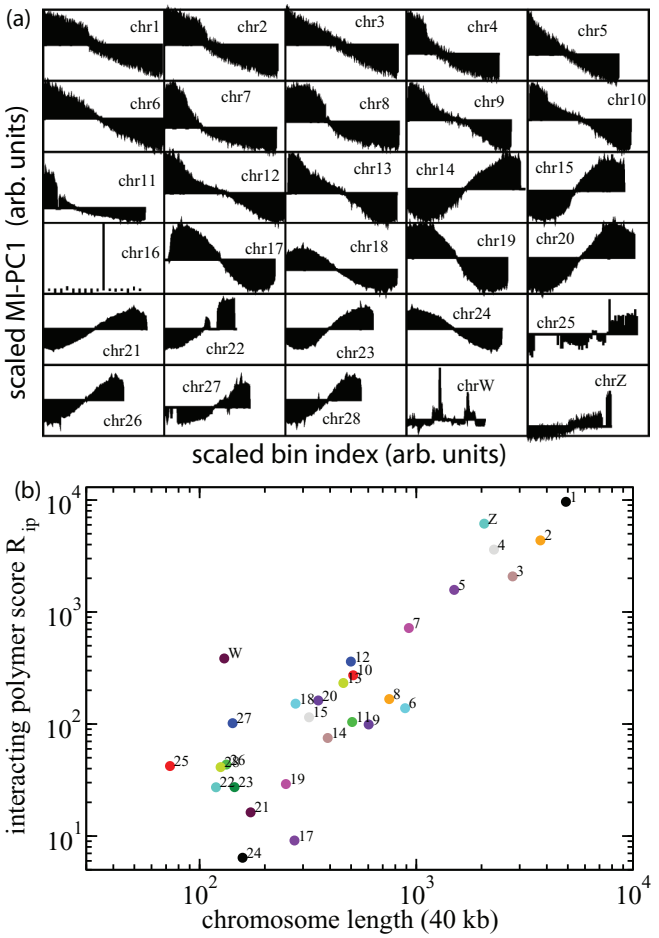


FIG. 7. (a) The structural features of chicken chromosomes at $t = 60$ min expressed by the MI-PC1s of the contact matrices. (b) The $R_{ip}$ versus the corresponding chromosome length.

and forming new contacts then increases nativeness again at $t = 10 - 30$ min before finally approaching the noninteracting limit at $t = 60$ min. Other chicken chromosomes show a similar dynamic signature as well, suggesting that this multiphase progression towards the mitotic chromosome is a fundamental biological feature. The detection of this multiphase structure transition feature showcases the usefulness of the contact analysis as this feature is not immediately apparent in the original Hi-C contact matrices.

Also, the conclusion that the nearly noninteracting, generic effective polymer state of prometaphase chromosomes at $t =$ 60 min holds throughout the chicken genome, not only for chr21. In Fig. 7, we display the MI-PC1 [Fig. 7(a)] and the corresponding interacting polymer score [Fig. 7(b)] across chicken chromosomes at prometaphase. We observed an interesting feature: the longer chromosomes have a larger $R_{ip}$ than the shorter ones. A few chicken chromosomes, chr16, 22, 25, and Z, display gaps in their contact maps which may affect their scores, but the differences observed between others cannot be simply explained by data artifacts and curve roughness. Further, this trend cannot be explained by the more trivial aspect of polymer chain length dependency, as the score function $R_{ip}$ has been properly normalized in its definition. It is interesting to speculate why larger chromosomes have larger $R_{ip}$ at $t = 60$ min and are visually different from the sinusoidal curves. One possible explanation is that larger chromosomes lag behind shorter ones in term of cell cycle dynamics, taking longer to reach their final mitotic structure. Another scenario could be that, with a fixed persistent length, larger chromosomes make more long-range contacts and thus effectively have a wide band-diagonal like matrix, as illustrated in Fig. 2(c).

Another interesting observation is that allosome chrW appears to be the only chromosome highly structured at $t =$ 60 min, both from a direct observation of MI-PC1 and from its position as a clear outlier on the score function plot. Upon closer inspection of the contact map, we speculated that these "structures" in fact represented errors in the sequencing assembly of the chicken genome chrW as represented in genome database galgal5 [56]. This misassembly, confirmed by galgal6, would have been much less noticeable without this current method of examination.

## IV. CONCLUDING REMARKS

Statistical analysis of contact matrices has been used to characterize the structure and dynamics of biopolymers. Particularly, I-PCA type of contact structural analysis is used to describe chromosome structures. Often, the matrix is pretreated to scale up the long-range contacts and enhance the signature of the structure. We explore the advantages of performing similar analyses with untreated contacts, especially for those structural ensembles of biopolymers with little specific structure and thus few long-range contacts. We also further suggest secondary tools to characterize the transformation from structured polymers to a generic polymer limit based on the first eigenvector of I-PCA. We found that certain operations on chromosomes will make them approach the noninteracting polymer limit (such as those that occur during certain phases of the cell cycle) while other perturbations are not dramatic enough. The methods and score functions developed can be useful to quantitatively study a wide range of largely unstructured biopolymer systems, from intrinsically disordered protein domains to mitotic chromosomes.

## ACKNOWLEDGMENTS

We specially thank R. J. Lindsay for helpful discussions on coarse-grained protein simulations using GROMACS. A portion of the research reported in this publication was supported by NIGMS of the National Institutes of Health under Award No. R35GM133557 to R.P.M. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[1] A. Fersht and U. A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman, New York, 1999).
[2] T. A. Brown, *Genomes 4* (Garland Science, New York, 2017).
[3] M. Doi and H. See, *Introduction to Polymer Physics* (Clarendon, Oxford, 1996).
[4] D. Michieletto, E. Orlandini, and D. Marenduzzo, Phys. Rev. X **6**, 041047 (2016).
[5] K. A. Dill, Biochemistry **24**, 1501 (1985).
[6] E. Lieberman-Aiden *et al.*, Science **326**, 289 (2009).
[7] G. Tiana and L. Giorgetti, *Modeling the 3D Conformation of Genomes* (CRC, Boca Raton, FL, 2019).
[8] M. Karplus and J. N. Kushick, Macromolecules **14**, 325 (1981).
[9] N. Gō and H. A. Scheraga, J. Chem. Phys. **51**, 4751 (1969).
[10] P. J. Flory and J. G. Jackson, *Statistical Mechanics of Chain Molecules* (Hanser, Munich, 1989).
[11] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[12] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).
[13] A. D. Schmitt, M. Hu, and B. Ren, Nat. Rev. Mol. Cell Biol. **17**, 743 (2016).
[14] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman, New York, 1999).
[15] T. Takahashi, K. O. Okeyo, J. Ueda, K. Yamagata, M. Washizu, and H. Oana, Sci. Rep. **8**, 13684 (2018).
[16] M. G. Poirier, A. Nemani, P. Gupta, S. Eroglu, and J. F. Marko, Phys. Rev. Lett. **86**, 360 (2001).
[17] M. Sun, R. Biggs, J. Hornick, and J. F. Marko, Chromosome Res. **26**, 277 (2018).
[18] R. J. Lindsay, B. Pham, T. Shen, and R. P. McCord, Nucleic Acids Res. **46**, 8143 (2018).
[19] Q. R. Johnson, R. J. Lindsay, and T. Shen, J. Comput. Chem. **39**, 1568 (2018).
[20] R. J. Lindsay, J. Siess, D. P. Lohry, T. S. McGee, J. S. Ritchie, Q. R. Johnson, and T. Shen, J. Chem. Phys. **148**, 025101 (2018).
[21] L. A. Mirny, Chromosome Res. **19**, 37 (2011).
[22] B. R. Lajoie, J. Dekker, and N. Kaplan, Methods **72**, 65 (2015).
[23] J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, Methods **58**, 268 (2012).
[24] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, Nat. Methods **9**, 999 (2012).
[25] Y. Zhou *et al.*, Nat. Commun. **10**, 1522 (2019).
[26] H. Miura, S. Takahashi, R. Poonperm, A. Tanigawa, S.-i. Takebayashi, and I. Hiratani, Nat. Genet. **51**, 1356 (2019).
[27] J. Ray, P. R. Munn, A. Vihervaara, J. J. Lewis, A. Ozer, C. G. Danko, and J. T. Lis, Proc. Natl. Acad. Sci. USA **116**, 19431 (2019).
[28] D. S. McKenzie, Phys. Rep. **27**, 35 (1976).
[29] P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).
[30] T. G. Dewey, *Fractals in Molecular Biophysics* (Oxford University Press, Oxford, 1997).
[31] A. L. Sanborn *et al.*, Proc. Natl. Acad. Sci. USA **112**, E6456 (2015).
[32] R. P. McCord, A. Nazario-Toole, H. Zhang, P. S. Chines, Y. Zhan, M. R. Erdos, F. S. Collins, J. Dekker, and K. Cao, Genome Res. **23**, 260 (2013).
[33] P. E. Rouse, J. Chem. Phys. **21**, 1272 (1953).
[34] B. H. Zimm, J. Chem. Phys. **24**, 269 (1956).
[35] K. Pal, M. Forcato, and F. Ferrari, Biophys. Rev. **11**, 67 (2019).
[36] C. A. Meyer and X. S. Liu, Nat. Rev. Genet. **15**, 709 (2014).
[37] W. Li, K. Gong, Q. Li, F. Alber, and X. J. Zhou, Bioinformatics **31**, 960 (2014).
[38] E. Yaffe and A. Tanay, Nat. Genet. **43**, 1059 (2011).
[39] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu, Bioinformatics **28**, 3131 (2012).
[40] C. A. McPhalen and M. N. G. James, Biochemistry **26**, 261 (1987).
[41] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, Annu. Rev. Biophys. **37**, 289 (2008).
[42] J. J. Portman, S. Takada, and P. G. Wolynes, Phys. Rev. Lett. **81**, 5237 (1998).
[43] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, Comput. Phys. Commun. **91**, 43 (1995).
[44] J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford, PLoS Comput. Biol. **12**, e1004794 (2016).

[45] S. E. Jackson and A. R. Fersht, Biochemistry **30**, 10428 (1991).
[46] A. G. Ladurner, L. S. Itzhaki, and A. R. Fersht, Fold Des. **2**, 363 (1997).
[47] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker, Science **342**, 948 (2013).
[48] J. H. Gibcus *et al.*, Science **359**, eaao6135 (2018).
[49] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, Nature **485**, 376 (2012).
[50] E. P. Nora *et al.*, Nature (London) **485**, 381 (2012).
[51] R. Amat, R. Böttcher, F. Le Dily, E. Vidal, J. Quilez, Y. Cuartero, M. Beato, E. de Nadal, and F. Posas, Genome Res. **29**, 18 (2019).
[52] E. Seto and M. Yoshida, Cold Spring Harbor Perspect. Biol. **6**, a018713 (2014).
[53] Z. Wang, C. Zang, K. Cui, D. E. Schones, A. Barski, W. Peng, and K. Zhao, Cell **138**, 1019 (2009).
[54] A. D. Stephens, P. Z. Liu, E. J. Banigan, L. M. Almassalha, V. Backman, S. A. Adam, R. D. Goldman, and J. F. Marko, Mol. Biol. Cell **29**, 220 (2018).
[55] J. Widom, J. Mol. Biol. **190**, 411 (1986).
[56] W. C. Warren *et al.*, G3: Genes Genomes Genetics **7**, 109 (2017).

*Correction:* A support statement for the third author was missing from the Acknowledgments section and has been inserted.