Genome dashboards: Framework and Examples

Zilong Li¹, Ran Sun¹ Thomas C. Bishop^{1*}

¹College of Engineering and Science, Louisiana Tech University, P.O. Box 10157 Ruston, LA 71272, USA

ABSTRACT

Genomics is a sequence-based informatics science and a 3D structure-based material science. Here we describe a framework for developing genome dashboards specifically designed to unify informatics with studies of chromatin structure and dynamics. The framework is based on the mathematical representation of geometrically exact rod models and the generalization of DNA base pair step parameters. A Model-View-Controller software design approach is proposed to implement genome dashboards as finite state machines either as desktop or web based applications. Two examples are demonstrated using our minimal genome dashboard called G-Dash-min. The data unification achieved with a genome dashboard supports the bi-directional exchange of data between informatics and structure. Thus any experimentally or theoretically determined sequence based informatics track can inform DNA, nucleosome or chromatin modeling (e.g. nucleosome positions) and structure features can be analyzed as informatics tracks in a genome browser (e.g. DNA base pair step parameters: Roll, Tilt, Twist). Here the framework is applied to chromatin, but genome dashboards are more broadly applicable. Genome dashboards are a novel means of investigating structure-function relationships for regions of the genome ranging from base pairs to entire chromosomes and for generating, validating, and testing mechanistic hypotheses.

INTRODUCTION

Chromatin is the biomaterial that contains the genome in all higher organisms. There is no consensus on the structure of chromatin (1), but there is a wealth of informatics and structure data available. From an informatics point of view, efforts such as the 1000 Genomes (2), ENCODE (3) and the 4D Nucleome (4) projects provide sequence based reference data. Coupling the reference data with Next Generation Sequencing (NGS) and informatics analysis pipelines enables individual labs to conduct Genome Wide Association (GWA) studies that link chromatin reprogramming with disease and altered gene expression as described, for example, in (6, 7).

Hi-C (8, 9), Micro-C (10) and other chromosome conformation capture methods (11, 12) provide distance constraints as a measure of the large scale organization of chromatin structure. Super resolution microscopy (13, 14) provides optical visualization of 3D structure at nanometer scale resolution, while electron microscopy (15), NMR (16),

and X-ray crystallography (17) provide ångström scale resolution of individual nucleosomes and nucleosome arrays. Strategies for modeling chromatin structures are rapidly maturing (18, 19, 20). The desire to merge computational and experimental approaches is recognized (21, 22), but a significant challenge in chromatin structural biology is unifying these diverse data sets to advance our understanding of structure-function relationships and to validate genomic mechanisms of action.

Here we label sequence based informatics as a 1D representation, contact and distant constraints as 2D representations, X-ray, NMR and super-resolution microscopy as 3D representations and dynamics as a 4D representation of chromatin. There exists a growing collection of computational tools that convert 2D data to 3D structures of chromatin (<https://www.4dnucleome.org/software.html>), and computational models can promote 3D structures to dynamics or sampling data (4D). But, there remain few tools, other than ICM-Web (23), that directly link sequence (1D) with chromatin structure (3D) or dynamics (4D). Thus researchers utilizing 1D sequence based methods are missing 3D structural data including steric and geometric constraints in their analyses, and researchers utilizing 3D and 4D computational and experimental methods are missing the wealth of informatics data available in sequence based data sets.

A genome dashboard, like an automobile dashboard or airplane cockpit, integrates a console for managing information with controllers for navigating a physical world that appears in a window. A “genome dashboard” unifies informatics (1D), contact and two-angle representations (2D) (24), structure (3D), and dynamics (4D) data describing DNA, nucleosomes and chromatin. For the purpose of developing such genome dashboards, we have identified a framework that unifies 1D and 3D representations and a general method for implementing it that supports data visualization and manipulation. The framework is bi-directional, i.e. can map 3D representations to 1D and 1D representations to 3D.

The framework is based on mathematical representations of geometrically exact rods (25), presented below as “The Model”, followed by “Design Considerations” for implementing it using Model-View-Controller software development principles. This approach enables a commodity-off-the-shelf (COTS) approach for assembling genome dashboards that is both extensible and portable. Two examples

*To whom correspondence should be addressed. Tel: +1 318 257 5209; Fax: +1 318 257 4000; Email: bishop@latech.edu

are then presented to demonstrate that G-Dash-min, a minimal web based implementation of a genome dashboard, can function in real time in both directions to convert informatics data into a physical model and a physical model into informatics data.

FRAMEWORK

The Model

From the genome dashboard perspective, informatics (1D) is any data that maps to DNA sequence. Generally speaking, NGS relies on aligning experimental data with chromosome coordinates and information theory for analysis. Physical models are 3D/4D descriptions that employ energy functions and physical laws for analysis. The energy functions are typically grouped into external and internal energies. $U = U_{ext} + U_{int}$, where U_{ext} captures through space interactions and U_{int} captures local conformation and dynamical properties. Both require knowledge of material properties and geometry. Material properties such as van der Waals radii, dielectric properties, partial charge distributions, moments of inertia, mass, stiffness parameters, bond angles, etc.. are all parameters associated with a specific physical model or force field. The model may employ atomic, coarse-grained or continuous medium approximations. Geometry is the structure of the physical model, it may be expressed in a laboratory (external) or material (internal) reference frame and may be obtained from theoretical or experimental techniques.

Our strategy for unification (merging data from different sources) is based on the idea that DNA is the common thread in chromatin structural biology. Unification is achieved through laboratory (Cartesian coordinate) and material (internal coordinate) representations of DNA as an oriented space curve or ribbon, Figure 1. Since unification is based on geometric considerations, our strategy is independent of the parameters associated with a specific physical model. Of course, associating an energy landscape with a physical model requires one to choose a physical model. Our framework does not provide this model but it does not restrict the user's choice of model. Implementations of the genome dashboard concept may support only one or many models.

In a laboratory reference frame, a continuous ribbon has centerline $\vec{r}(s)$ and unit length directors $\hat{d}_i$ embedded in the ribbon that capture the local orientation of the ribbon. The directors can be represented by a director frame matrix $\mathbf{D} = \{\hat{d}_1(s), \hat{d}_2(s), \hat{d}_3(s)\}$ (26). This matrix also serves to transform a representation in the material(internal) frame to a representation in the laboratory (Cartesian) frame.

An equivalent description of the ribbon is based on the director frames themselves. This description is a material reference frame description that captures the translations and rotations connecting one director frame to the next, represented here by $[\vec{\Gamma}(s), \vec{\Omega}(s)]$. The two representations $[\vec{r}(s), \mathbf{D}(s)]$ and $[\vec{\Gamma}(s), \vec{\Omega}(s)]$ are equivalent descriptions of the conformation of an oriented space curve, denoted simply as $C(s)$ for the Cartesian Coordinate representation and $T(s)$ when expressed in the material frame and interpreted as informatics tracks.

Figure 1. Unification is the process of merging data from different sources. Physical models and informatics data are unified by mathematical representations of an oriented space curve in laboratory $[\vec{r}(s), \mathbf{D}(s)]$ and material $[\vec{\Gamma}(s), \vec{\Omega}(s)]$ reference frames. The conformation of a physical model $C(s)$ is associated with the laboratory frame, and informatics data $T(s)$ is associated with the material frame. Masks $M(s)$ alter the material properties of DNA. Exchanging data between laboratory and material frames unifies the physical model and informatics.

Converting between the $[\vec{\Gamma}(s), \vec{\Omega}(s)]$ and $[\vec{r}(s), \mathbf{D}(s)]$ representations requires either a differentiation or an integration as expressed by the following equations.

$$\frac{d\vec{r}}{ds} \equiv \vec{t} = \mathbf{D}\vec{\Gamma} \quad \frac{d\hat{d}_i}{ds} = \mathbf{D}\vec{\Omega} \times \hat{d}_i \quad (1)$$

$$\vec{r}(s_1, s_2) = \int_{s_1}^{s_2} \mathbf{D}\vec{\Gamma} ds \quad \hat{d}_i(s_1, s_2) = \int_{s_1}^{s_2} \mathbf{D}\vec{\Omega} \times \hat{d}_i ds \quad (2)$$

Here $\vec{t}(s)$ is the unnormalized tangent to the ribbon expressed in the laboratory frame. $\vec{\Gamma}(s)$ is the same vector expressed in the internal frame. $\mathbf{D}\vec{\Omega}(s)$ is recognized as the vector corresponding to the instantaneous axis of rotation of the director frames located along the ribbon at position s , as represented in the laboratory coordinate frame. Discrete approximations to Equations 1 and piecewise integration as expressed in Equations 2 can be employed to obtain a collection of discrete director frames. Different models may require different numerical algorithms to achieve the required discretization. Numerical methods suitable for DNA are discussed below. They should be reading strand invariant. Together, the two representations provide a basis for unifying informatics ($[\vec{\Gamma}(s), \vec{\Omega}(s)]$) and structure ($[\vec{r}(s), \mathbf{D}(s)]$) data.

DNA conformation $C(s)$ is at best a base pair discrete approximation to a continuous oriented space curve (27, 28). Base pair step parameters (29, 30) and associated algorithms provide established methods for describing double and single stranded DNA as a discrete oriented space curve at atomic resolution. A sequence specific di-nucleotide accurate model of dsDNA in the $[\vec{\Gamma}(s), \vec{\Omega}(s)]$ (base pair step parameter) representation can be obtained from x-ray (31) or molecular dynamics (32, 33) studies. A $[\vec{r}(s), \mathbf{D}(s)]$ description is obtained by integrating $[\vec{\Gamma}(s), \vec{\Omega}(s)]$. There are two widely used tools for base pair step parameter analysis. 3DNA (34) uses a Euler Angle (E-A) based method and employs a “RollTilt” approximation (35). Curves+ (36) uses a Euler-Rodrigues (E-R) based method (37). Both methods utilize

a mid-step plane construction to ensure that the computed parameter values are not affected by the choice of reading strand. Mathematically, one must invert the sign of Tilt and Shift upon strand reversal to preserve the alignment of director frames with the DNA major and minor groove (38).

Helical parameter values obtained from the 3DNA and Curves+ methods are known to differ (39). Differences may arise from two sources. The assignment of director frames to base pairs may differ. However, the methods for assigning director frames are well-defined for ideal pairing (30), so these differences typically occur only for significant deviations from ideal geometries. The other source of differences is method dependent. These differences have not been well studied. Helical parameter values obtained from 3DNA and from Curves+ differ even when the director frames used for the calculations are identical, i.e. even when the first problem is eliminated. Thus, values obtained from one method should not in general be interchanged with the other. Nonetheless these two methods work well for all-atom representations of the pairing and stacking of base pairs in double stranded DNA with the caveat that neither provides information about the DNA backbone. Recent efforts now support proper reconstruction of the DNA backbone (40).

Chromatin, for our purposes, is a biomolecular structure composed of DNA and external agents, such as histones, that alter the material properties (conformation, dynamics, flexibility, energetics, chemical properties) of a contiguous length of DNA from s to $s+n$. We label any such external agent, including histones, as a Mask, $M(s, s+n)$. Any number of identical or unique Masks ($M_i(s_i, s_i+n_i)$) may be associated with a sequence of DNA (s_i to s_i+n_i). With this approach, chromatin folding is the process of activating Masks. There are two strategies for activating Masks. The first is achieved in the material reference frame with the conformation of the masked DNA denoted by a list of internal coordinates $M(s, s+n) = [[\vec{r}_M(s), \vec{D}_M(s)], \dots, [\vec{r}_M(s+n), \vec{D}_M(s+n)]]$ that spans n base pairs. As the name suggests, the Mask replaces the $[[\vec{r}(s), \vec{D}(s)], \dots, [\vec{r}(s+n), \vec{D}(s+n)]]$ values associated with DNA. We can calculate $C(s)$ from Equations 2, as discussed above. The second approach is achieved in the laboratory reference frame with the conformation of DNA described as a rigid entity with $M(s, s+n) = [[\vec{r}_M(s), \mathbf{D}_M(s)], \dots, [\vec{r}_M(s+n), \mathbf{D}_M(s+n)]]$. The Mask consists of Cartesian coordinates and director frames which can be converted to $T(s)$ using Equations 1 as discussed above. The Cartesian coordinate representation of $M(s, s+n)$ requires only a single translation and rotation to position each rigidly masked element in the laboratory reference frame.

In terms of Masks, a nucleosome is a DNA superhelix and histones. The histones can be represented independently of DNA as a single entity (sphere, cylinder, ellipsoid), a collection of beads, or an all-atom model. Alternatively the DNA and histones can be included in the nucleosome Mask as a single entity. Docking individual histones or the complete histone octamer to a superhelix or placing entire nucleosomes between linkers can be achieved with the same methods and tools used for describing the relative rotations and translations of base pairs. However the “RollTilt” small angle approximation is no longer valid. Linker DNA connecting

Figure 2. Model-View-Controller (MVC) Design. Model: Laboratory frame $[\vec{r}(s), \mathbf{D}(s)]$ and material frame $[\vec{r}(s), \vec{Q}(s)]$ descriptions of DNA as the common thread, an inventory of Masks $M(s)$, and procedures for converting between representations. View: a Molecular Visualization (MV) displays $C(s)$, a Genome Browser (GB) displays $T(s)$ and a Control Panel (CP) provides a graphical interface to the controller. G-Dash-min uses JSmol and Biodalliance for the MV and GB components, respectively. A commodity-off-the-shelf (COTS) approach enables a genome dashboard to use any desired MVs and GBs. Controller: manages the exchange of data between Model and Views.

Masks is often assumed to be free DNA, but in general, even linker DNA may be described by Masks, e.g. bent or less flexible linkers. Likewise chemical modification of the DNA, e.g. methylation, which does not change the sequence but the physical-chemical properties of DNA, is a Mask.

In the context of a genome dashboard, chromatin folding is an informatics problem of describing all the unique Masks and tracking their locations along a sequence of DNA. These Masks can be manipulated based on informatics or physical analyses. In this manner, genome dashboards are designed to enable users to efficiently define and navigate chromatin folding landscapes.

Design Considerations

A genome dashboard is a finite state machine that can be efficiently developed using Model-View-Controller (MVC) design principles (41), Figure 2. This approach ensures that a dashboard’s components are independent, replaceable, and extensible.

The “Model” in the MVC schema is the data and related logic. For a genome dashboard, the Model includes the $[\vec{r}, \mathbf{D}]$ and $[\vec{r}, \vec{Q}]$ representations of DNA as a discrete (or even continuous (42, 43)) oriented space curve, the inventory of Masks, $M_i(s_i, s_i+n_i)$, any associated track data, $T(s)$, and procedures for converting between representations. In general, the rotations and translations associated with a Mask may be large. If a Mask is a rigid entity, this information can be leveraged to improve performance. For example, representing all 147 base pairs of DNA and eight histones in a nucleosome as a single director frame along with a large deformation of the path of DNA reduces computational and data costs by approximately $n \times 146$, where n is the number of nucleosomes containing 147 base pairs.

The “View” in the MVC schema provides the user interface and renders data. A genome dashboard includes a 3D Molecular Visualization (MV) for rendering $[\vec{r}, \mathbf{D}]$, a Genome Browser (GB) for rendering $[\vec{\Gamma}, \vec{\Omega}]$, and a Control Panel (CP) as a graphical user interface to the Controller. A genome dashboard can be designed as a web application or a stand alone application. For web applications, javascript based MVs such as JSmol (44) and NGL Viewer (45) are optimal. For stand alone applications, MVs such as VMD (46) and PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.) are optimal. Likewise, for web applications, the GB should be javascript based, like Biodalliance (47). For stand alone applications, JBrowse (48) or other modern genome browsers may provide advantages.

The “Controller” in the MVC schema manages the exchange of data between the View and the Model. For a given genome dashboard, the MV and GB can be COTS elements, but the Control Panel and Controller are application specific. We expect that different instances of the genome dashboard concept will target different users and utilize different physical models. Different strategies for managing the MV, GB, and CP may emerge, but the underlying Model remains as described above.

EXAMPLES

Based on the framework proposed above, we have developed a minimal genome dashboard named “G-Dash-min”. Here we demonstrate two examples using G-Dash-min to show how genome dashboards contribute to our understanding of biological function by the unification of informatics and physical models.

Informatics to physical model

A hormone response element (HRE) is a specific sequence of DNA representing 15 base pairs. Selective binding of an activated hormone receptor (HR) to the HRE is a critical component of the hormone response mechanism, see Fig 1-41 of (49). A variant of this gene regulatory mechanism is employed to control numerous physiologic functions in all higher organisms. To demonstrate the power of data unification achieved with our G-Dash-min application, we have identified estrogen response elements (ERE) using ERE-Finder (50). This informatics data is displayed as the ERE Track in Fig 3. All experimentally determined nucleosome positions for the human genome (51) are also displayed in Fig 3 as the Nuc-Pos Track. These Tracks provide locations for EREs and nucleosomes; however, the 1D representations are insufficient to determine whether or not the locations are physically realizable. These tracks are sufficient to identify several regions of interest. One of them is associated with chromosome 6 and coordinate location 168,131,722 to 168,132,130. Here we find three overlapping nucleosomes and an ERE that appears to function as a classic switching mechanism. We explore this hypothesis with G-Dash-min by generating physical models.

We first selected the single nucleosome shown in the bottom of the Nuc-Pos Track in Fig 3 and generated a coarse-grained model. Mapping the ERE location onto the physical model provided its physical location, but without knowledge of major

groove orientation one can still not determine the accessibility of the ERE site for ER binding. We then constructed an all-atom model. With the all-atom model, we see that the major groove is actually facing towards the histones. This prevents the estrogen receptor DNA-binding domain (1HCQ) from binding to this region of the DNA major groove. To bind 1HCQ to the all-atom model we downloaded the all-atom model from G-Dash-min then loaded it and 1HCQ into VMD. A simple VMD script fit the DNA in 1HCQ to the DNA in the G-Dash-min all-atom model. (This script and the models are provided as supplementary data.)

We then used the same approach to model the two nucleosomes shown in the top of the Nuc-Pos Track in Fig 3. Mapping the ERE location to the coarse-grained model suggests the ERE may be accessible to ER. Using the same procedure and script as before, we find that 1HCQ can physically access this ERE with the nucleosome present. We point out that 1HCQ is only the DNA binding domain of the estrogen receptor. There exist steric conflicts between the ER-DBD and the histones so this is not the complete story; but it strongly suggests this site as a candidate for a genetic switching mechanism.

With this example we demonstrate with G-Dash-min how informatics is used to construct a physical model that extends and validates the interpretation of the informatics data. We emphasize that any informatics track or combination of tracks can be used to inform the physical model.

Physical model to informatics

Models of chromatin are rapidly maturing. As the models develop, there is increasing demand to capture biologic realism, including DNA sequence, experimentally determined nucleosome positions, states of chemical modification etc... Manually curating informatics data to build a “biologically inspired” model is a time-consuming and tedious task that informs the initial model but does not necessarily support the interpretation of modeling results. There exists a growing collection of published physical models to which such interpretations could be applied. Here we import a HOXC Mesoscale model generated by the Schlick lab (52) into G-Dash-min to demonstrate how informatics can be overlaid onto physical models to extend the interpretation.

In G-Dash-min we provide an upload function that is specific for DiscoTech based models (53). We upload and convert the HOXC Mesoscale model into a G-Dash-min model, Figure 4. The HOXC model utilizes a 9 base pair per bead model that includes both the location and orientation of each bead and each DiscoTech based nucleosome, i.e. $\vec{r}(s)$ and $\mathbf{D}(s)$ are provided in the model. We treat the DiscoTech nucleosomes as Masks and calculate $[\vec{\Gamma}(s), \vec{\Omega}(s)]$ based on the $[\vec{r}(s), \mathbf{D}(s)]$ data provided for the beads and Masks. The $[\vec{\Gamma}(s), \vec{\Omega}(s)]$ values computed are no longer DNA base pair step parameters; however, they still represent an oriented space curve or ribbon. We refer to them as generalized step parameters (GSP) and utilize the same naming conventions as for the DNA parameters: Tilt, Roll, Twist, Shift, Slide and Rise, and display them as informatics tracks in the genome browser. Our generalized parameter values differ significantly from those associated with DNA. Small angle approximations are no longer valid, see data in Figure 4-d.

Figure 3. Black boxes: $C(s)$ and $T(s)$ representations of two allowed states. Upper boxes are $T(s)$ representations of nucleosome positions (blue bars) and an ERE (red bar). Lower boxes are $C(s)$ representations (small beads represent 5 base pairs large beads represent histone octamers). Red ellipses: The corresponding all-atom structures with the estrogen receptor DNA-binding domain docked to the DNA as in PDB entry 1HCQ. a) The ERE is located within a nucleosome with the major groove facing inward. The receptor is prohibited from binding. b). The ERE is located in a nucleosome free region. Docking 1HCQ indicates that the ERE is physically accessible.

To determine if either of the E-R or E-A approaches is suitable for our GSPs, we have converted all six HOXC models reported in (52) from the $C(s)$ to the $T(s)$ and back to the $C(s)$ representations. Each G-Dash-min converted model contains approximately 1779 beads, represents 55,000 base pairs and has a unique conformation. The RMSD between the initial and final $C(s)$ structures are computed using VMD's RMSD functions. The E-R approach yields RMSD values ranging from 0.03Å to 0.15Å after alignment, and the E-A approach yields RMSD values ranging from 0.002Å to 0.003Å after alignment. Our experience is that both the E-A and E-R methods are acceptable algorithms for implementing "The Model" even when thermal fluctuations or deformations associated with nucleosomes and chromatin, or even generalized step parameters, are modeled. (All of these models are provided as supplementary data.)

The HOXC model was constructed for a specific sequence of DNA: HOXC10 of annotated human genome assembly 38 that begins at chr12:53,985,065. However, the data provided does not contain sequence information because the model itself is sequence independent. Thus, whenever a DiscoTech model is uploaded into G-Dash-min it must be aligned with sequence using the yellow bar sequence selection bar, Figure 4. If sequence information is included in the model, the model can be automatically aligned to

data in the genome browser. Once aligned, the structural informatics tracks enable us to compare the nucleosome positions used in the model (Nucleosomes Track in Figure 4) to experimentally determined nucleosome positions (Nuc-Pos Track in Figure 4). It is clear that the two tracks differ. Resolving these differences promises to advance our understanding of both the experimental and modeling data.

To complete our multi-dimensional representation of chromatin folding, we have incorporated 2D representations into G-Dash-min. Figure 4-b is a two angle plot and Figure 4-c is a distance-distance matrix plot. As with the structure tracks, these are automatically generated for models containing sufficiently many nucleosomes. The two angle plot is a Woodcock Equivalent (WE) Plot (24, 54). On these plots, α is the angle between the centers of three adjacent nucleosomes, and β is the dihedral rotation angle obtained from the centers of four adjacent nucleosomes. For this analysis, the reported α and β values are computed as if the linkers were straight, even if the linkers are not. For this reason, we have adopted the label "Woodcock Equivalent" Plot. The distance-distance matrix reports the center-to-center distance between all nucleosomes. Unlike the exchange of data between the informatics (1D) and physical models (3D), the WE Plot data and distance-distance maps are one directional. The 2D representations are obtained from the 3D model but cannot be used to generate 3D

Figure 4. a) HOXC coarse-grained model of chromatin containing approximately 55k base pairs of DNA and 284 nucleosomes. Uploading the HOXC model to G-Dash-min generates: b) a two angle representation of the HOXC model, c) a distance-distance matrix based on nucleosome centers of mass, and d) structural informatics data. “Generalized Helical Parameter” (“Twist” and “Rise”) and nucleosome position (“Nucleosomes”) data are displayed along side experimentally determined nucleosome positions (“Nuc-Pos”) and other informatics data (“GenCode”).

models. Methods exist for this purpose but have not yet been implemented in G-Dash-min.

DISCUSSION

As a working example of a genome dashboard, G-Dash-min demonstrates that informatics and physical models can be unified in a web based application in real time. Our tests demonstrate that interactive usage can be achieved for systems containing 10,000 to 50,000 base pairs or coarse-grained beads. The algorithms for converting informatics to physical models and physical models to informatics work in both directions for base pair resolution models using either the Euler-Rodrigues (Curves+) or Euler-Angle (3DNA) based method. We believe our application to the HOXC model is the first demonstration that both the Euler-Rodrigues and the Euler-Angle algorithms can be applied to coarse-grained models discretized well beyond the base pair level. We label the internal coordinates “generalized step parameters” in this case.

G-Dash-min also demonstrates that a commodity-off-the-shelf approach coupled with Model-View-Controller principles can be employed to efficiently develop genome dashboards that are customizable, extensible and portable. G-Dash-min generates atomic and coarse-grained models of DNA, nucleosomes, and chromatin using various experimental and theoretical data. Coupling G-Dash-min’s atomic modeling capabilities with high-performance, high-throughput workflows, and our TMB Library of nucleosome simulations (55) provides a software ecosystem for overnight comparative molecular dynamics simulations of nucleosomes (56). Such models are necessary for developing designer nucleosomes and assessing receptor-DNA interactions in their native context (57).

The genome dashboard framework achieves unification of informatics and physical models; the challenge is data representation. Genomics data has well-defined data formats, but there are numerous data formats for Cartesian coordinate data and no established conventions for representing director frame data or step parameters. We have demonstrated that the

E-R and E-A algorithms can support multi-scale modeling of DNA as the common thread in chromatin using a geometrically exact rod model. However, we have deliberately avoided energetic considerations. The genome dashboard user or developer must decide which energy model is most appropriate for their particular application. We thus expect many genome dashboards to be developed, each tailored for a specific use.

As described here, genome dashboards are designed to work with chromatin folding, but the concept and framework are not limited to chromatin or even to eukaryotes. The informatics data can be any data associated with an indexing system, e.g. protein sequence. The structure data can be any slender polymer or potentially even an ordered collection of points in space that are linked to form an oriented space curve. The latter includes observations of nucleosomes with super-resolution microscopy. If the sequence ordering of nucleosomes in an image can be determined and director frames can be assigned to each nucleosome, it will be possible to unify 1D, 2D, 3D, and 4D representations of chromatin as in Figure 4.

ACKNOWLEDGEMENTS

We thank Joohyun Kim and Jinghua Ge of the Center for Computation and Technology, Louisiana State University, and John Gentle of Science Gateways Community Institute (SGCI) for assistance and advice. We thank the Schlick Lab at NYU for sharing coordinate data for the HOXC models. Ran Sun's contributions were related mostly to all-atom modeling and helix parameter analysis. Zilong Li developed G-Dash-min as a web-based tool and prepared the manuscript.

FUNDING

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health [P20 GM103424-17]; National Science Foundation [OIA-1541079]; and the Louisiana Board of Regents.

Conflict of interest statement. None declared.

REFERENCES

- Fussner,E., Ching,R.W. and Bazett-Jones,D.P. (2011) Living without 30 nm chromatin fibers. *Trends Biochem Sci.*, **36**, 1-6. [PubMed:20926298] [doi:10.1016/j.tibs.2010.09.002].
- Auton,A., Brooks,L.D. and et al. (2015) A global reference for human genetic variation. *Nature.*, **526**, 68-74. [PubMed:26432245] [PubMed Central:PMC4750478] [doi:10.1038/nature15393].
- Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature.*, **489**, 57-74. [PubMed:22955616] [PubMed Central:PMC3439153] [doi:10.1038/nature11247].
- Dekker,J., Belmont,A.S. and et al. (2017) The 4D nucleome project. *Nature.*, **549**, 219-226. [PubMed:28905911] [PubMed Central:PMC5617335] [doi:10.1038/nature23884].
- Voong,L.N., Xi,L. and et al. (2017) Genome-wide Mapping of the Nucleosome Landscape by Micrococcal Nuclease and Chemical Mapping. *Trends Genet.*, **33**(8), 495-507. [PubMed:28693826] [PubMed Central:PMC5536840] [doi:10.1016/j.tig.2017.05.007].
- Cowper-Sallari,R., Zhang,X. and et al. (2012) Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.*, **44**, 1191-1198. [PubMed:23001124] [PubMed Central:PMC3483423] [doi:10.1038/ng.2416].
- Sadlon,T., Brown,C.Y. and et al. (2018) Unravelling the molecular basis for regulatory T-cell plasticity and loss of function in disease. *Clin Transl Immunology.*, **7**, e1011. [PubMed:29497530] [PubMed Central:PMC5827651] [doi:10.1002/cti2.1011].
- Belton,J.M., McCord,R.P. and et al. (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods.*, **58**, 268-276. [PubMed:22652625] [PubMed Central:PMC3874846] [doi:10.1016/j.ymeth.2012.05.001].
- Belaghal,H., Dekker,J. and Gibcus,J.H. 2017 Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods.*, **123**, 56-65. [PubMed:28435001] [PubMed Central:PMC5522765] [doi:10.1016/j.ymeth.2017.04.004].
- Hsieh,T.S., Fudenberg,G. and et al. (2016) Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. *Nat Methods.*, **13**, 1009-1011. [PubMed:27723753] [doi:10.1038/nmeth.4025].
- Dekker,J., Rippe,K. and et al. (2002) Capturing chromosome conformation. *Science.*, **295**, 1306-1311. [PubMed:11847345] [doi:10.1126/science.1067799].
- Sati,S. and Cavalli,G. (2017) Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma*, **126**, 33-44. [PubMed:27130552] [doi:10.1007/s00412-016-0593-6].
- Duim,W.C., Jiang,Y. and et al. (2014) Super-resolution fluorescence of huntingtin reveals growth of globular species into short fibers and coexistence of distinct aggregates. *ACS Chem. Biol.*, **9**, 2767-2778. [PubMed:25330023] [PubMed Central:PMC4273975] [doi:10.1021/cb500335w].
- Ricci,M.A., Cosma,M.P. and Lakadamyali,M. (2017) Super resolution imaging of chromatin in pluripotency, differentiation, and reprogramming. *Current opinion in genetics & development*, **46**, 186-193. [PubMed:28843811] [doi:10.1016/j.gde.2017.07.010].
- Wilson,M.D. and Costa, A. (2017) Cryo-electron microscopy of chromatin biology. *Acta crystallographica. Section D, Structural biology.*, **73**, 541-548. [PubMed:28580916] [PubMed Central:PMC5458496] [doi:10.1107/S2059798317004430].
- Mlynárik,V. (2017) Introduction to nuclear magnetic resonance. *Anal. Biochem.*, **529**, 4-9. [PubMed:27210513] [doi:10.1016/j.ab.2016.05.006].
- Tan,S. and Davey,C.A. (2011) Nucleosome structural studies. *Curr. Opin. Struct. Biol.*, **21**, 128-136. [PubMed:21176878] [PubMed Central:PMC3052702] [doi:10.1016/j.sbi.2010.11.006].
- Schlick,T., Hayes,J. and Grigoryev,S. (2012) Toward convergence of experimental studies and theoretical modeling of the chromatin fiber. *J Biol Chem.*, **287**, 5183-5191. [PubMed:22157002] [PubMed Central:PMC3285299] [doi:10.1074/jbc.R111.305763].
- Perišić,O. and Schlick,T. (2016) Computational strategies to address chromatin structure problems. *Phys Biol.*, **13**, 035006. [PubMed:27345617] [doi:10.1088/1478-3975/13/3/035006].
- Portillo-Ledesma,S. and Schlick,T. (2019) Bridging chromatin structure and function over a range of experimental spatial and temporal scales by molecular modeling. *WIREs Comput. Mol. Sci.*, e1434. [doi:10.1002/wcms.1434].

21. Ozer, G., Luque, A. and Schlick, T. (2015) The chromatin fiber: multiscale problems and approaches. *Curr Opin Struct Biol.*, **31**, 124-39. [PubMed:26057099] [PubMed Central:PMC4476956] [doi:10.1016/j.sbi.2015.04.002].
22. Perkel, J.M. (2017) Plot a course through the genome. *Nature.*, **549**, 117-118. [PubMed:28880304] [doi:10.1038/549117a].
23. Stolz, R.C. and Bishop, T.C. (2010) ICM web: the interactive chromatin modeling web server. *Nucleic Acids Res.*, **38**, W254-61. [PubMed:20542915] [doi:10.1093/nar/gkq496].
24. Woodcock, C.L., Grigoryev, S.A. and et al. (2015) A chromatin folding model that incorporates linker variability generates fibers resembling the native structures. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 9021-9025. [PubMed:8415647] [PubMed Central:PMC47493] [doi:10.1073/pnas.90.19.9021].
25. Simo, J.C., Marsden, J.E. and Krishnaprasad, P.S. (1988) The Hamiltonian structure of nonlinear elasticity: the material and convective representations of solids, rods, and plates. *Arch. Ration. Mech. Anal.*, **104**(2), 125-183. [doi:10.1007/bf00251673].
26. Simo, J.C. and Vu-Quoc, L. (1991) A geometrically-exact rod model incorporating shear and torsion-warping deformation. *Int J Solids Struct.*, **27**, 371-393. [doi:10.1016/0020-7683(91)90089-X].
27. Calladine, C.R. and Drew, H.R. (1986) Principles of sequence-dependent flexure of DNA. *J Mol Biol.*, **192**, 907-918. [PubMed:3586013] [doi:10.1016/0022-2836(86)90036-7].
28. Fathizadeh, A., Eslami-Mossallam, B. and Ejtehadi, M.R. (2012) Definition of the persistence length in the coarse-grained models of DNA elasticity. *Phys Rev E Stat Nonlin Soft Matter Phys.*, **86**, 051907. [PubMed:23214814] [doi:10.1103/PhysRevE.86.051907].
29. Dickerson, R.E. (1989) Definitions and nomenclature of nucleic acid structure component. *Nucleic Acids Res.*, **17**, 1797-1803. [PubMed:2928107] [PubMed Central:PMC317523] [doi:10.1093/nar/17.5.1797].
30. Olson, W.K., Bansal, M. and et al. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J Mol Biol.*, **313**, 229-37. [PubMed:11601858] [doi:10.1006/jmbi.2001.4987].
31. Olson, W.K., Gorin, A.A. and et al. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A.*, **95**, 11163-11168. [PubMed:9736707] [PubMed Central:PMC21613] [doi:10.1073/pnas.95.19.11163].
32. Lavery, R., Zakrzewska, K. and et al. (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, **38**, 299-313. [PubMed:19850719] [PubMed Central:PMC2800215] [doi:10.1093/nar/gkp834].
33. Pasi, M., Maddocks, H. and et al. (2014) μ ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272-83. [PubMed:25260586] [PubMed Central:PMC4231739] [doi:10.1093/nar/gku855].
34. Lu, X.J. and Olson, W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108-5121. [PubMed:12930962] [PubMed Central:PMC212791] [doi:10.1093/nar/gkg680].
35. el Hassan, M.A. and Calladine, C.R. (1995) The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. *J Mol Biol.*, **251**, 648-64. [PubMed:7666417] [doi:10.1006/jmbi.1995.0462].
36. Lavery, R., Moakher, M. and et al. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917-5929. [PubMed:19625494] [PubMed Central:PMC2761274] [doi:10.1093/nar/gkp608].
37. Gonzalez, O., Petkeviciūtė, D. and Maddocks, J.H. (2013) A sequence-dependent rigid-base model of DNA. *J Chem Phys.*, **138**, 055102. [PubMed:23406150] [doi:10.1063/1.4789411].
38. Diekmann, S. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J Mol Biol.*, **205**(4), 787-791. [doi:10.1016/0022-2836(89)90324-0].
39. Babcock, M.S., Pednault, E.P. and Olson, W.K. (1994) Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *J Mol Biol.*, **237**, 125-156. [PubMed:8133513] [doi:10.1006/jmbi.1994.1213].
40. Petkeviciūtė, D., Pasi, M., Gonzalez, O. and Maddocks, J.H. (2014) cgDNA: a software package for the prediction of sequence-dependent coarse-grain free energies of B-form DNA. *Nucleic Acids Res.*, **42**, e153. [PubMed:25228467] [PubMed Central:PMC4227758] [doi:10.1093/nar/gku825].
41. Dijkstra, E.W. (1974) Programming as a discipline of mathematical nature. *Am. Math. Monthly*, **81**(6), 608-612. [doi:10.1080/00029890.1974.11993624].
42. Bishop, T.C. (2009) VDNA: the virtual DNA plug-in for VMD. *Bioinformatics*, **25**, 3187-3188. [PubMed:19789266] [PubMed Central:PMC3167696] [doi:10.1093/bioinformatics/btp566].
43. Bishop, T.C. and Hearst, J.E. (1998) Potential Function Describing the Folding of the 30 nm Fiber. *The Journal of Physical Chemistry B*, **102**(33), 6433-6439. [doi:10.1021/jp980909s].
44. Angel Herráez. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**(4), 255-261. [PubMed:21638687] [doi:10.1002/bmb.2006.494034042644].
45. Alexander, S.R. and Peter, W.H. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**(W1), W576-W579. [PubMed:25925569] [PubMed Central:PMC4489237] [doi:10.1093/nar/gkv402].
46. Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD - Visual Molecular Dynamics. *J. Molec. Graphics*, **14**, 33-38. [doi:10.1016/0263-7855(96)00018-5].
47. Down, T.A., Matias, P. and Tim, J.P.H. (2011) Dalliance: Interactive Genome Viewing on the Web. *Bioinformatics*, **27**(6), 889-890. [PubMed:21252075] [PubMed Central:PMC3051325] [doi:10.1093/bioinformatics/btr020].
48. Skinner, M.E., Uzilov, A.V. and et al. (2009) JBrowse: A next-generation genome browser *Genome Res.*, **19**(9), 1630-1638. [PubMed:19570905] [PubMed Central:PMC2752129] [doi:10.1101/gr.094607.109].
49. Anthony, W. N. and Gerald, L. (1997) *Hormones*. Academic Press.
50. Andrew, P.A. and Adam, G.J. (2019) erefinder: Genome-wide detection of oestrogen response elements. *Mol Ecol Resour.*, [PubMed:31177626] [doi:10.1111/1755-0998.13046].
51. Yongbing, Z., Jinyue, W. and et al. (2018) NucMap: a database of genome-wide nucleosome positioning map across species. *Nucleic Acids Res.*, **47**(D1), D163-D169. [PubMed:30335176] [PubMed Central:PMC6323900] [doi:10.1093/nar/gky980].
52. Bascom, G.D., Myers, C.G. and Schlick, T. (2019) Mesoscale modeling reveals formation of an epigenetically driven HOXC gene hub. *Proc Natl Acad Sci U S A.*, **116**(11), 4955-4962. [PubMed:30718394] [PubMed Central:PMC6421463] [doi:10.1073/pnas.1816424116].
53. Zhang, Q., Beard, D.A. and Schlick, T. (2003) Constructing irregular surfaces to enclose macromolecular complexes for mesoscale modeling using the discrete surface charge optimization (DISCO) algorithm. *J Comput Chem.*, **24**, 2063-2074. [PubMed:14531059] [doi:10.1002/jcc.10337].
54. Helmut, S., William, M. G. and Robijn, B. (2001) DNA Folding: Structural and Mechanical Properties of the Two-Angle Model for Chromatin. *Biophys. J.*, **80**(4), 1940-1956. [PubMed:11259307] [PubMed Central:PMC1301383] [doi:10.1016/S0006-3495(01)76164-4].
55. Sun, R., Li, Z. and Bishop, T. C. (2019) TMB Library of Nucleosome Simulations. *J. Chem. Inf. Model.*, [PubMed:31490686] [doi:10.1021/acs.jcim.9b00252].
56. Smith, J.A., Romanus, M. and et al. (2013) Scalable online comparative genomics of mononucleosomes: a BigJob. *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, **13**, 23:1-23:8. [doi:10.1145/2484762.2484819].
57. Bishop, T.C., Kosztin, D. and Schulten, K. (1997) How hormone receptor-DNA binding affects nucleosomal DNA: the role of symmetry. *Biophys J.*, **72**, 2056-2067. [PubMed:9129808] [PubMed Central:PMC1184400] [doi:10.1016/S0006-3495(97)78849-0].

Physical Model
 $C(s)$

unification

Informatics
 $T(s)$

Laboratory frame
 $[\vec{r}(s), \underline{D}(s)]$

$[M_i(s_i, s_i + n_i)], \dots$

Material frame
 $[\vec{\Gamma}(s), \vec{\Omega}(s)]$

$[M_i(s_i, s_i + n_i)], \dots$