GSO: A Semantic Web Ontology to Capture 3D Genome Structure

W. Jim Zheng[1], Jingcheng Du[1], Jijun Tang[2], Cui Tao[1]

[1]School of Biomedical Informatics
University of Texas Health Science Center at Houston
{wenjin.j.zheng,jingcheng.du,cui.tao}@uth.tmc.edu

[1]Department of Computer Science
University of South Carolina
{jtang}@cse.sc.edu

Abstract. More and more evidences indicate that the 3D conformation of eukaryotic genome plays important functional role in the cell. While extensive experimental investigations have been performed to study such structure-function relationships using techniques such as Hi-C methods, there is no data standard to capture the 3D conformation of the genome. In previous work, we have developed an object-oriented framework, Genome3D, to integrate, model and visualize human genome in 3-dimension. We report here about a high level ontology, Genome Structure Ontology (GSO), to capture the 3D conformation of eukaryotic genome. GSO captures the structural organization of eukaryotic genome at levels ranging from chromosome, to 30nm fiber, to nucleosome and then to the atomic level of DNA. We believe such ontology could play a significant role in annotating, analyzing and describing molecular data for the 3D conformation of eukaryotic genome.

1 Introduction

The latest research indicates that spatial conformation and interaction of chromatin (1, 2) play a fundamental role in important cellular functions (such as gene regulation) and cell state determination (e.g., stem cell pluripotency). The influx of new details about the higher-level structure and dynamics of the genome enabled many recent efforts measuring and modeling genome structures at various resolutions and levels of detail (3-10). While incorporating this spatial information can yield a 3D genome model, the size and complexity of these data dictate the design and development of new algorithms and methods in data integration and model construction.

The 3D conformation of eukaryotic genome is complex: at high level, chromosomes go through different states through the cell cycle and take different conformations; each of these states will have different structures; at low level, the conformation from 30nm fiber to nucleosome, and to histone and DNA are also affected by the cell cycle. Precise modeling of these chromosomes' conformation is critical for us to understand the structure-function relationships for the 3D conformation of the genome. As the exploration of the genome 3D structure becomes more and more

comprehensive, developing a Genome Structure Ontology (GSO) to capture the structural information in formal specifications becomes imperative. While there are extensive ontologies, such as gene ontology and sequence ontology that have been developed to model genome information, none is aimed at capturing the 3D genome structure.

We created the first model-view framework of eukaryotic genomes, *Genome3D*, to enable integration and visualization of genomic and epigenomic data in a three-dimensional space (11). Our model of the physical genome implicitly contains all levels of structure and hierarchy, and provides an underlying platform for integrating multi-scale genomic information within three dimensions. Our viewer uses a hierarchical model of the relative positions of all nucleotide atoms in the cell nucleus. Through this work, we have gained extensive experience in capturing the multi-scale information about the 3D structure of human genome. Here we report our initial attempt to transform this knowledge into Genomic Structure Ontology at high level.

2 Methods and Results

We created the GSO in Web Ontology Language (OWL) (12) in the Protégé ontology editing environment. OWL is built on formalisms that use Description Logic (DL)(13) to allow reasoning and inference. The Rule Interchange Format (RIF)(14) can be used to add rules to OWL and can be used to infer new knowledge from an OWL based ontology and reason about OWL individuals. Moreover, the Semantic Web community has developed open source tools for editing, storing, and reasoning over information represented in OWL. Building the GSO in OWL will allow us to directly OWL's formalism and tools for semantic defining the knowledge in the domain and perform reasoning in the future.

Figure 1 shows the hierarchy defined in GSO. The hierarchy reflects the conceptual structure of eukaryotic chromosome, together with high level, critical information about the state of the chromosome and the cell cycle phases, which are critical in determining the 3D conformation of the chromosome. The details for the anaphase chromosome are captured with details all the way down to the DNA. The design of this hierarchy is accommodative such that the details for other phases or other types of genomes can be incorporate in the future.

GSO provides a way for us to semantically define important genome structure entities such as Nucleosome Core Particles (NCP). Each NCP has a 146 bps long DNA wrapping around the histone core, with a radius and rise, and has a unique position and orientation define by a radial vector and an axis. The histone core is an octamer with two copies of Histone H2A, H2B, H3 and H4. Figure 2 shows the ontological representation of the NCP. Each NCP has a base pair (BP) index, which is composed of three vectors, axis of radiation, center position, and the radius to the starting BP. We created a new class called *BPIndex* and three object properties (*axisofRotation*, *centerPosition*, and *radiusToStartingBP*) to represent the three vectors. Each vector is

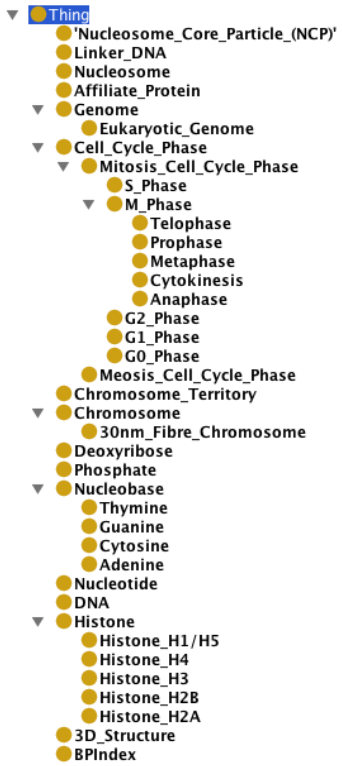

Figure 1: Hierarchy of genome structure for eukaryotic genome defined in GSO.

represented by a class called *3D_Stucture*. The domain of the object properties is *BPIndex* and the range is *3D_Stucture*. Three data properties *x_coord*, *y_coord*, and *z_coord* have been defined to represent the x, y, and z coordinator values in a 3D structure. In addition, we also created properties to represent other important features of an NCP. The Data property *rise_nm* is defined to represent the rise per turn of the NCP in nm; the Data property *rot_rad* is defined to represent the length of DNA spiral around the NCP in radians; the Data property *ncp_bps* is defined to represent the number of bps in the DNA NCP wrap; and the Data property is defined to represent the number of NCPs in file. After these classes and properties have been defined, we can represent individual NCPs in RDF with respective of the GSO definition. Figure 3 shows an example RDF representation of an NCP (partial).

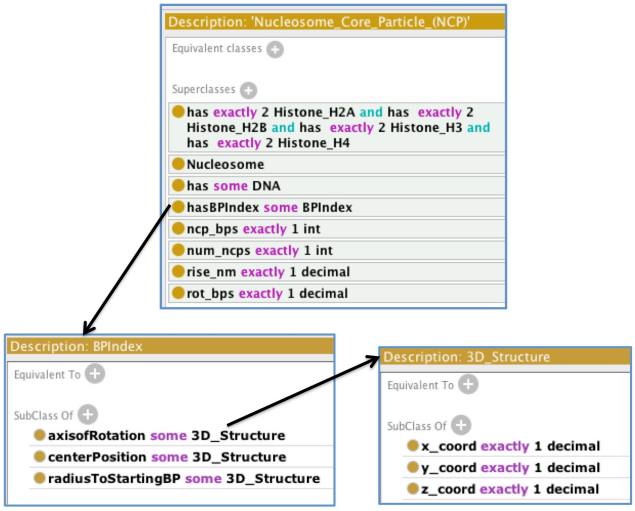

Figure 2: Ontological definition of NCP

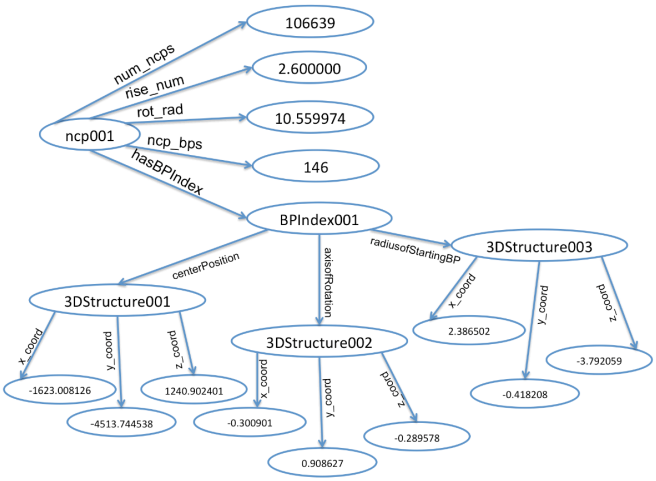

Figure 3: RDF representation of a sample NCP (partial). Each NCP is defined by a set of parameters: rise, rotation, ncp bps and BPIndex. The BPIndex is defined by center position, the axis of rotations and the radius of starting base pair

GSO is intended to capture the 3D conformation of the genome, and has overlaps with other existing ontologies. In order to avoid re-inventing the wheel, we surveyed the existing ontologies for ontology terms that we used in the GSO. We applied the

NCBO annotator on the all the GSP terms and searched for matches for all the ontologies hosted by the NCBO Bioportal. Not surprisingly, we have found that 98 other ontologies contain terms in GSO. We have created a matrix to link the terms in GSO with the terms in the other ontologies. Out of the 98 ontologies, the top 10 ontologies are BIOMODELS, CRISP, MESH, RH-MESH, NIFSTD, AURA, GO, GO-EXT, HINO, and SNML. We will further investigate these ontologies and terminologies and align the GSO terms to them when applicable.

3 Discussion

Built upon our knowledge of constructing a 3D physical model of the human genome, we developed a Genome Structure Ontology (GSO) for eukaryotic genome. GSO captures the basic structure of the genome from the whole genome level to the chromosome level, to the 30nm fibre level, to the nucleosome level and then down to the atomic level of the DNA structure. Combined with quantitative parameters, GSO also defines the details of genome structure at each level.

A mature GSO cannot only define a 3D genomic model, but also annotating existing data used to construct the model. For example, analysis results from Chip-Seq data for the nucleosome position can be annotated with the ontology term such as NCP. Such annotation can help data user to quickly identify data needs to be used for 3D model construction. While annotating a 3D model of a eukaryotic genome, the ontology can help to infer the relationships of different components of a model, thus providing useful insight from spatial relationships. For example, in our original Genome3D paper (11), we gave an example of multiple Single Nucleotide Polymorphisms (SNPs) occurred within an NCP, we can easily infer all the occurrence of such SNPs for the whole genome using GSO. Another example of such application is an enrichment analysis of ontology terms in space: given a transcription hotspot or CHIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing) spot, what are the ontology terms are enriched within a radius of 0.1μm.

Even though the GSO is still a prototype, further refinement of this prototype could make it a very useful tool to annotate and to capture the 3D structure of eukaryotic genome. More semantic annotations can be built on top of the GSO classes in order to support automatic inference for different use cases.

4 Acknowledgement

This work is supported by NSF award #1339470 to WJZ, #1161586 to JT, and NIH/NLM award R56 LM010680 to WJZ and R01LM011829 to JD and CT.

References:

1. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009;462(7269):58-64. Epub 2009/11/06.

2. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289-93. Epub 2009/10/10.

3. P. Hahnfeldt JEH, D.J. Brenner, R.K. Sachs, and Lynn R. Hlatky. Polymer Models for interphase chromosomes. PNAS. 1993;90(16):7854-8.
4. Sachs RK, van den Engh G, Trask B, Yokota H, Hearst JE. A random-walk/giant-loop model for interphase chromosomes. Proc Natl Acad Sci U S A. 1995;92(7):2710-4.
5. Ponomarev AL, Brenner D, Hlatky LR, Sachs RK. A polymer, random walk model for the size-distribution of large DNA fragments after high linear energy transfer radiation. Radiat Environ Biophys. 2000;39(2):111-20.
6. Woodcock CL, Grigoryev SA, Horowitz RA, Whitaker N. A chromatin folding model that incorporates linker variability generates fibers resembling the native structures. Proc Natl Acad Sci U S A. 1993;90(19):9021-5.
7. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295(5558):1306-11.
8. Balaeff A, Mahadevan L, Schulten K. Modeling DNA loops using the theory of elasticity. Phys Rev E Stat Nonlin Soft Matter Phys. 2006;73(3 Pt 1):031919.
9. Beard DA, Schlick T. Computational modeling predicts the structure and dynamics of chromatin fiber. Structure. 2001;9(2):105-14.
10. Sharma S, Ding F, Dokholyan NV. Multiscale modeling of nucleosome dynamics. Biophysical journal. 2007;92(5):1457-70.
11. Asbury TM, Mitman M, Tang J, Zheng WJ. Genome3D: a viewer-model framework for integrating and visualizing multi-scale epigenomic information within a three-dimensional genome. BMC Bioinformatics. 2010;11:444. Epub 2010/09/04.
12. McShan DC, Rao S, Shah I. PathMiner: predicting metabolic pathways by heuristic search. Bioinformatics. 2003;19(13):1692-8.
13. Horrocks I, Patel-Schneider PF, McGuinness DL, Welty CA. OWL: a Description Logic Based Ontology Language for the Semantic Web. The Description Logic Handbook: Theory, Implementation, and Applications 2ed: Cambridge University Press; 2007.
14. van Helden J, Naim A, Mancuso R, Eldridge M, Wernisch L, Gilbert D, et al. Representing and analysing molecular and cellular function using the computer. Biol Chem. 2000;381(9-10):921-35.