Project Acronym: MuG

Project title: Multi-Scale Complex Genomics (MuG)

Call: H2020-EINFRA-2015-1

Topic: EINFRA-9-2015

Project Number: 676556

Project Coordinator: Institute for Research in Biomedicine (IRB Barcelona)

Project start date: 1/11/2015

Duration: 36 months

Deliverable 3.1: A critical evaluation of the problems on data structure the browser has to solve.

Lead beneficiary: Institute for Research in Biomedicine (IRB Barcelona)

Dissemination level: PUBLIC

Due date: 29/01/2016

Actual submission date: 18/02/2016

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676556.

Document history

Version	Main author(s)	Partner	Date	Comments
0.1	Mike Goodstadt François Serra	CNAG-CRG	05/02/2016	First version circulated to Technical Board board for revision
0.2	Marco Pasi	UNOT	11/02/2016	Added information on size scale range for the different levels in section 3.4 (p.7)
0.3	Andrew Yates	EBI	12/02/2016	Minor correction spelling in section 3.7 (p.8)
1.0	Anna Montras	IRB Barcelona	15/02/2016	Merged version, approved by Technical Board
1.0			18/02/2016	Approved by Supervisory Board

Table of contents

Executive summary

The present deliverable evaluates current genomic data structures for use in MuG visualization, detailing requirements; aptness of existing formats, and proposed methodology and formats to achieve. A visualization tool (browser) is being developed and will serve as a proof-of-concept of how this data may be gathered, processed and rendered. This document includes a critical evaluation of the problems on data structure the browser has to solve.

1 INTRODUCTION

Visualization of multi-scale genomics is the concise, appropriate and useful representation of available data and associated research findings to form a comprehensive modeling tool. This visualization would allow the user to gain insights in their research by bringing a greater clarity and utility of the data. Therefore to ensure that the data associated with the MuG project is adequate for this task, this paper gives a critical evaluation of the type of data we would be handling and on their specificities.

2 DESCRIPTION

2.1 The Problems of MuG Data

Cell biology is founded on the discovery of macromolecular structures (our genomes) serving as instruction sets that, until recently, have been characterized and interpreted only as linear sequences. However, this represents only a single dimension of the genome and it is now known that the function and regulation of the cell is, not only regulated by the linear sequence of the genome or by epigenetic marks it carries, but also by its folding.

Developing computational infrastructure for understanding how the genome is arranged in space is at the core of Multi-Scale Complex Genomics (MuG) project (see D2.1). Chromosome Conformation Capture (3C)-like experiments identify interactions within the genome with the data typically plotted as a 2D matrix. These interactions have been shown to indicate proximity; therefore the experienced researcher can infer possible conformations from a visual study of such an interaction matrix. Even so, direct visualization of these 3D objects can assist user interaction with complex data and comprehension of the structures.

Therefore any visualization needs to be able to present the linear, the spatial and the dynamic natures of the genome.

2.2 The Power of Visualization

Graphs in 2D are the typical formats for visualizing for genetic data and used correctly are powerful in their abstraction and reduction for fast and precise communication. Representing data in 3D can cause distortion, obfuscation due to perspective, lighting and the complexities of human visual processing [1]. However for spatial data, human vision can quickly and intuitively comprehend aspects such as arrangement, scale, etc. Therefore carefully constructed 3D visualizations can produce highly accessible representations and facilitate valid insights.

2.3 A Visual Grammar for the Genome

Given the nascent state of this area of investigation, there is yet to emerge a standard form of representation of the spatial nature of the genome that is being discovered. An example of how such a **visual grammar** may develop and the resulting appearance can been seen in the ‘cartoon’ representation of protein structure [2] which abstracts inherent complexity. An example of the added understanding that a browser may bring can be seen in the Aquaria protein browser [3].

3 Analysis

3.1 Discrete Levels of Representation

The term multiscale correctly describes the nature of cellular structure and therefore the needs of any complete visualization of cell biology. There are observable and discrete scales at which the structure

and function of cells can be described and analyzed. This is in part due to the principles of grouping inherent in human perception but also derives from actual components and roles.

The following schema of levels of representation typical of that used to describe **scales of biological structure** where Levels 1 to 4 best reflect the focus of the MuG project and the data under consideration (note that there is no equivalent comprehensive schema within the EBI ontologies):

Within MuG scope	Not in MuG scope
1. Atomic 2. Molecular 2.1. Small Organic Molecules / Biomolecules 2.2. Macromolecules 2.3. Supramolecular Complexes 3. Subcellular Structures 4. Cellular	5. Tissue 6. Organ 7. Organ System 8. Organism 9. Population 10. Community / Biocenoses 11. Biome / Major Habitat Types 12. Biosphere

Categorization of the data using this level schema will help define and assign appropriate visualization at each scale and clarity of transitions between data types. The typical data under consideration in the MuG project does not consider higher levels (5 to 12), however they might be important to consider especially in the case of level 8 for the Organism. Where no data is available, the level can be indicated as deficient, again giving clarity of extent and limitations of what is being represented and also leaving room for possible future applications.

These levels equate to the levels of details used in interactive visualizations and, in particular, in efficient 3D computer graphics. The levels therefore form the basis from which coherent and effective visualization can be constructed.

3.2 Shared Properties of Data

Associations need to be created between disparate data on different levels so as to assemble a multi-scale model. Tying the data together is also needed to the smooth transition between levels in any visualization. To implement this a number of common, cross-scale properties (unifiers) must be identified and included in all data within the MuG project. The following lists suggests unifiers for discussion:

- Domain: Bacteria, Achaea, Eukaryotes
- Organism: *E. Coli*, *Arabidopsis*, Yeast, *Caenorhabditis*, *Drosophila*, Zebrafish, Mouse, Human
- Chromosome: X, Y, 2L, 19 etc.
- Source: publication DOI, experiment protocol, computation configuration
- Valuation: publication status, confidence score from production
- Sample Origin: country (US, EU, CN etc.), geolocation (e.g. 41.34357, 2.03659)

These unifiers could be added as annotation in various ways (filenames, tags, headers, xml, or in an independent database) or for tighter integration of these with the original data, as a codified, single text string (e.g. `01010101-EDSM0X-101038NSMB1936-E09EU4102`). Further discussion is needed to determine the most stable and secure form of annotation. Moreover this annotation could be integrated into the form by which any data is identified within MuG.

3.3 Common Identification of Data

The MuG project brings together multiple types of data from a variety of sources, many of which not easily centralized due to size, ownership, privacy concerns, etc. It is essential for confidence and utility of the project that ease of access and explicit identification and providence is assigned to any data used in the project. This labeling would ideally be done on output from production i.e. by the software or machine but, given backlog or historic data, a process of verification and assignation will be required when importing data into the project.

The Digital Object Identifier (DOI) standard already "provides an infrastructure for persistent unique identification of objects of any type". It also provides for identities which have multiple sources and this may be a method of designating a common identity for data at various scales that are aligned either in experiment, organism, cell cycle etc. This is proposed as the best solution and examples should be produced and tested.

3.4 Data Characterized by Scale

To be able to produce visualizations with standard visual grammar at each scale, the type of the data at each scale needs to be characterized. To help clarify, the following table gives relevant examples of the biology at each scale:

Level	Level Name	Example of Cellular Biology	Size scale
1.0	Atomic	Nucleotide chemistry	Å – nm
2.1	Molecular - Biomolecules	Unpacked nucleotides, linear reads, base pair correlated	1 – 10 nm
2.2	Molecular - Macromolecules	Packed nucleotides, average reads, chromatin types	10 – 100 nm
2.3	Molecular – Supramolecular	TADs, Chromosomes, Centromere anchor points	0.1 – 1 µm
3.0	Subcellular Structures	Organelles, Cytoplasm	1 – 10 µm
4.0	Cellular	Cells	10 – 100 µm

Note that there is an approximate 10^3 jump between scales, which gives a useful guide for defining and creating the visualizations at differing scales. Using these criteria, the following table of the level covered by each data listed in the Data Description Survey of WP4:

Data Type	Levels Covered
Genomic sequences	2.1
Sequence Annotations	2.2
Nucleosome positioning (histones)	2.1 - 2.2
Nucleic Acids 3D Structures	1.0 - 2.2
Nucleic Acids MD Atomistic Trajectories	1.0 - 2.2
Nucleic Acids MD CG Trajectories	1.0 - 2.2
ChIP-Seq, ATAC-Seq, DNA methylation and RNA-Seq data	2.1 - 2.2
HiC sequencing reads	2.2
HiC contact matrix	2.2
HiC contact coverage in close/far-cis, tran	2.2
HiC scaling factor	2.3
HiC TADs	2.3
HiC differential contacts	2.2
HiC contact peak	2.2
HiC contact coverage	2.3

HiC directionality index	2.3
TADs data	2.3
Compartments data	2.3
Ensemble of chromatin 3D structures	2.3

3.5 Data at All Levels

To form a complete multi-scale model, all levels need to have data from which visualization can be formed. There is research being undertaken in this field at all scales of cell biology, however from the above assessment in section 3.4 it is clear that currently there are missing areas of data, namely 3.0 and 4.0. This data, for example FISH images, needs to be sourced and incorporated. Furthermore, to provide for reference and testing, a collection of typical data for all levels should be assembled.

3.6 Data in the Visualization

Ideally, the visualization tools should be able to combine different types of data into a single interactive workspace on screen – a dashboard or storyboard. Additionally tools must be able to process, manipulate and generate the representations without recourse to intermediate, additional format to ensure traceability and simplicity of function. Navigation within a visualization must be able to be performed fluidly across domains and scales e.g. between chromosomes and across scales.

3.7 Connection to external datasets

Given the importance of functionally annotating the models produced by the MuG consortium, it is necessary to develop the visualizers able to connect to external datasets. Those include, but are not limited to the Ensembl (http://www.ensembl.org), the UCSC Genome Browser (http://genome.ucsc.edu), the KEGG (http://www.genome.jp/kegg/) and the Protein Data Bank (http://www.pdb.org), among others. Those databases provide APIs to be directly plugged-in to our visualizers.

4 CONCLUSIONS

This document identified the different data types and their scales that will need to be handled and represented by the MuG visualizer. Our consortium is well placed to further develop such visualizers complying with international standards and accessing/delivering data in proper formats. The MuG groups will help in implementing a new formal representation of genomes and genomic domains that could be shared between the relevant researchers in the 3D genomics community.

5 REFERENCES

1. Heinrich, J., et al., Faraday Discuss, 2014. 169:179-93.

2. Richardson, J.S., Nat Struct Biol, 2000. 7:624-5.

3. O'Donoghue, S.I., et al., Nat Methods, 2015. 12:98-9.