


**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 6.3: Software tools linking multiresolution structural and simulation data

**Lead beneficiary**: The University of Nottingham (UNOT)

**Dissemination level**: PUBLIC

Due date: 31/10/2018

Actual submission date: 31/10/2018



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

# Document history

| Version | Contributor(s) | Partner | Date | Comments |
|---|---|---|---|---|
| 0.1 | Athina Meletiou | UNOT | 08/10/2018 | First draft |
| 0.2 | Charles Laughton | UNOT | 09/10/2018 | First review |
| 0.4 | Athina Meletiou | UNOT | 29/10/2018 | Second review |
| 1.0 | | | 30/10/2018 | Final Version. Approved by Supervisory Board |

# Table of Contents

# 1 Executive summary

One of the main objectives of the MuG VRE is to integrate data from different scales, resolutions, and/or sources. Focusing on structural and simulation data, the outputs of task 6.3 provide VRE users with the suitable tools to produce, process, and analyse data at various scales, as well as tools that utilise data from multiple scales to provide further insights in the field of 4D genomics. This document defines and outlines those tools deployed on the MuG VRE that provide the infrastructure needed to navigate, integrate, and analyse different levels of DNA data resolution or scale, such as 3D structures, experimental data, simple DNA sequence data, and atomistic or coarse-grained trajectories.

# 2 Introduction

Molecular simulations produce vast amounts of data with information about structure and dynamics at the atomistic level, particularly as computational resources become more powerful and accessible. At the same time these data need substantial processing in order to be linked with data from studies at different resolution levels or scales. Further, integrating structural information derived from experiments with simulation data is another challenge facing the genomics community.

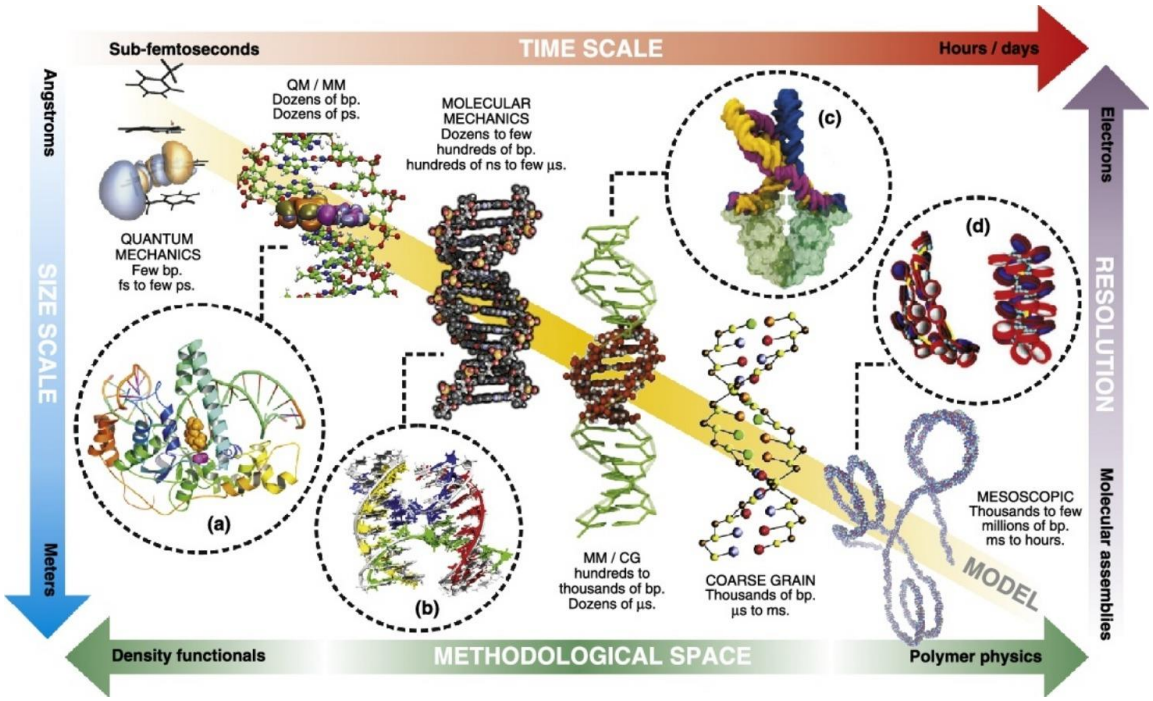



Figure 1. Schematic representation[1] of the intrinsic multiscale nature of DNA; from a relevant review article (https://doi.org/10.1016/j.sbi.2015.11.011) by Orozco *et al*. focusing on recent theoretical methods for multiscale simulation of DNA. The theoretical methods discussed in the review are represented in the schematic above according to five dimensions: (i), the time scale that each model samples; (ii), the system size; (iii), the methodological space; (iv), the model resolution available at each level.

DNA is an inherently multi-resolution molecule whose study requires moving in a wide range of sizes and time scales (Figure 1). The study of DNA covers a broad range of different topics, from the sub-Angstrom details of the electronic distributions of nucleobases, to the mechanical properties of millimeter-long chromatin fibres.[1] In biomolecular simulation, various theoretical methods allow the study of DNA at different resolution levels. Those can be summarised in four groups, namely: electronic, atomistic, coarse-grained, and mesoscopic, as can be seen in Figure 1. QM methods are computationally expensive and their applicability on nucleic acids is limited either to small model systems or combined with atomistic molecular mechanics methods. All-atom MD can handle very large biomolecular systems, however force-field inaccuracies and need for refinement should be taken into consideration. Coarse-graining (CG) methods enable the study of large systems that cannot be negotiated at the atomistic level, and mesoscopic methods aim to tackle the study of chromatin structure.

As computational resources become even more accessible and computer power increases, and while experimental data increase in resolution, it is anticipated that multiscale approaches will steer the future of chromatin studies. In this document, we present tools that are integrated within the MuG VRE that focus on linking data from different levels of resolution together with experimental data.

# 3 Tools

## 3.1 MC-DNA

By Jürgen Walther (IRB Barcelona)

This tool creates 3D all-atom B-DNA conformations of a sequence of interest. The user can obtain either the ground state structure or a molecular dynamics-like trajectory. Even though the outputs are atomistic, the smallest unit of the models is a base-pair (bp) as models are based on a Metropolis Monte Carlo algorithm with bp resolution. Utilising a Monte Carlo algorithm means that the tool runs up to $10^5$ faster than conventional molecular dynamics and at a tiny fraction of the cost, thus enabling the study of much larger systems. The tool therefore takes structural information at base-pair resolution, and outputs data at atomistic resolution.

**Inputs:** txt file with DNA sequence.

## 3.2 Chromatin Dynamics

By Jürgen Walther (IRB Barcelona)

Chromatin Dynamics is an extension of MC-DNA (described above) to model chromatin fiber conformations. The user can design their own chromatin fiber as a "beads-on-a-string" representation. The chromatin fiber can be produced by providing a linker DNA sequence and the nucleosome positions, and it can be either a single structure or a trajectory of possible chromatin conformations. The sampling in this case is done *via* Monte Carlo moves on the bp-step parameters of the linker DNA while the nucleosomal DNA remains untouched.

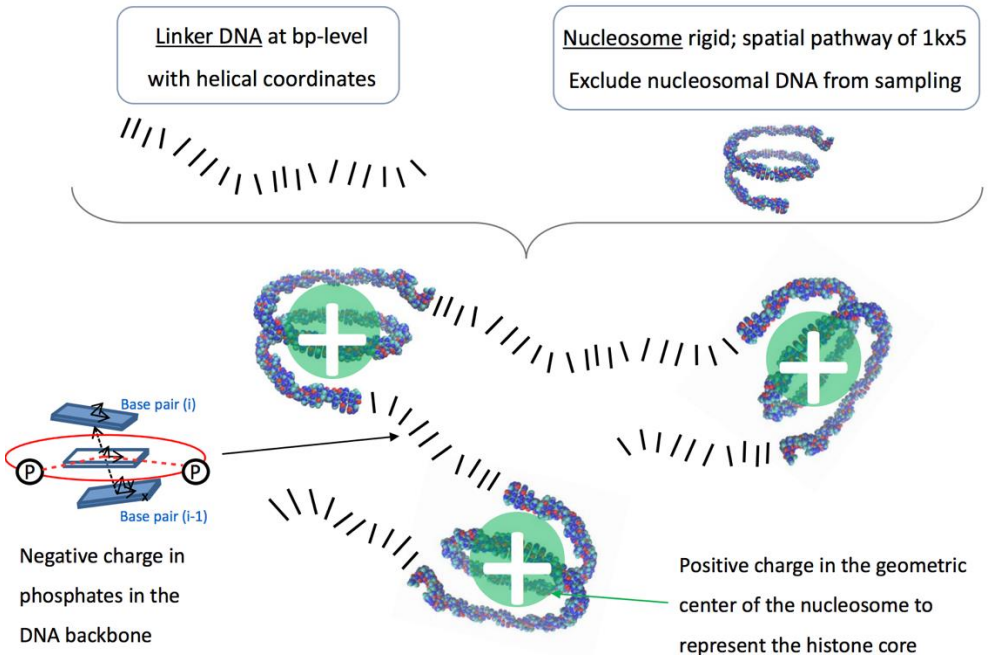



Figure 2. Chromatin Dynamics overview.

Alternatively, the user can obtain chromatin fiber configurations from whole genome nucleosome positioning experiments such as MNase-seq analysis. In the VRE, this can be achieved by utilising the results from NucleR of the Nucleosome Dynamics workflow. The Nucleosome Dynamics server integrated in the VRE offers different tools to analyse nucleosome positioning from MNase-seq data. More specifically, NucleR defines the location of nucleosomes by performing Fourier transform filtering and peak calling. Thus Chromatin Dynamics links structural data at kilobase resolution with data at the base-pair resolution.

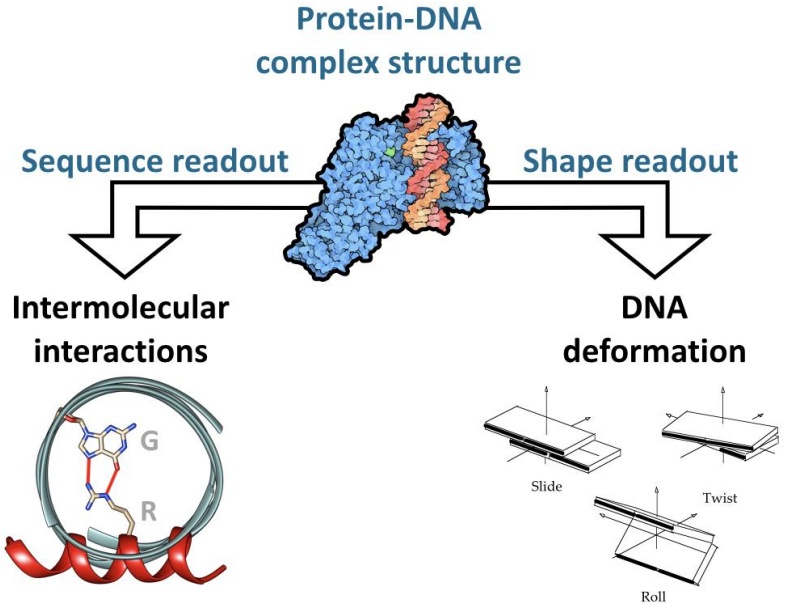
**Inputs:** a DNA sequence and a nucleosome positioning txt file; or a nucleosome positioning gff3 file from NucleR of the Nucleosome Dynamics workflow.

## 3.3 3DConsensus

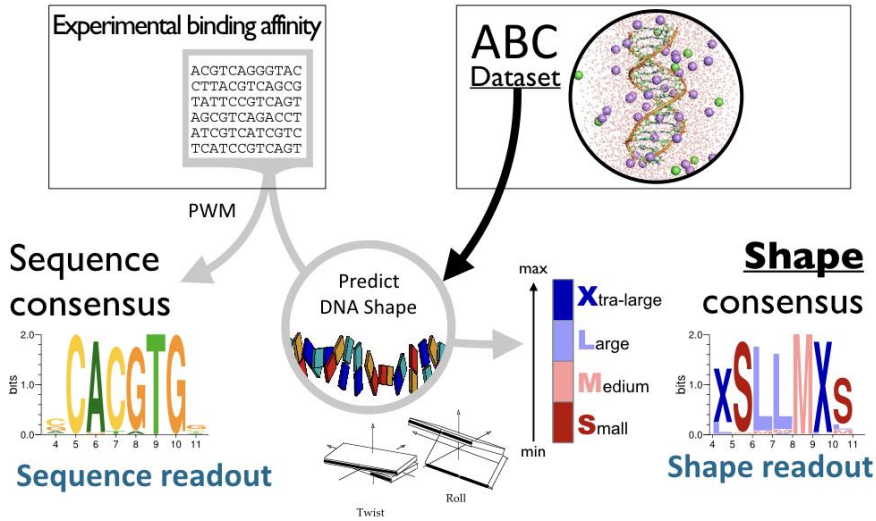
By Marco Pasi (UNOT)

3DConsensus is a tool for the analysis of protein-DNA complexes. It analyses their 3D structure, either experimental or modeled using PyDockDNA, to identify interactions and assess their impact on specific binding. This is done by integrating experimental data on the protein's DNA sequence specificity, obtained for example using ChIP-seq or Protein Binding Matrices.

3DConsensus extracts conformational parameters of the DNA in the complex from the 3D structure, and compares it to the predicted behaviour of a corresponding DNA molecule in the absence of its partner in order to deconvolute the effect of DNA sequence and of protein binding on the structure of DNA. It further identifies the specific interactions that stabilise the complex.




A sequence-only consensus is derived from the experimental data, as well as a “shape” consensus based on the physical properties of DNA, by leveraging the extensive atomic-detail information on the sequence-dependent behaviour of naked DNA available in the ABC Dataset.[2] The mechanical properties of DNA at the base resolution are the basis for the comparison of the results of these analyses directly in the VRE, using the provided graphical interface, to establish meaningful links between structural information at the atomistic scale, and experimental binding specificity data measured at the genomic scale.




**Inputs:** 3DConsensus requires the atomistic structure of the protein-DNA complex (in PDB format) and the experimental binding or relative binding affinity information encoded in a plain-text file.

## 3.4 NAFlex

By Adam Hospital (IRB Barcelona)

NAFlex[3] is a set of tools to analyse molecular dynamics trajectories, produced *in situ* or uploaded by the user, of nucleic acids, isolated or protein-bound, from atomistic MD or coarse-grained simulations. A complete set of flexibility analysis is integrated, including helical parameters, principal components, local and global stiffness, energy decomposition, hydrogen bonds, distance contacts, NMR observables (such as NOEs and J-Couplings), and stacking energies.

**Inputs:**

- *From structure*

| | |
|---|---|
| PDB file | Nucleic acid 3D structure |

- *From trajectory*

| | |
|---|---|
| PDB | Nucleic acid 3D structure |
| PARMTOP | Nucleic acid topology |
| MDCRD, DCD, NETCDF | Nucleic acid trajectory |

# 4 References

1. Dans, P. D.; Walther, J.; Gómez, H.; Orozco, M. Multiscale simulation of DNA. *Curr. Opin. Struct. Biol.* **2016**, *37*, 29-45.
2. Pasi, M.; Maddocks, J. H.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T., 3rd; Dans, P. D.; Jayaram, B.; Lankas, F.; Laughton, C.; Mitchell, J.; Osman, R.; Orozco, M.; Pérez, A.; Petkevičiūtė, D.; Spackova, N.; Sponer, J.; Zakrzewska, K.; Lavery, R. μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **2014**, *42*, 12272-12283.
3. Hospital, A.; Faustino, I.; Collepardo-Guevara, R.; González, C.; Gelpí, J. L.; Orozco, M. NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.* **2013**, *41*, W47-W55.