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

Milestone 22: 4D data from ABC project processed to generate database of DNA sequence-flexibility relationship.

Lead beneficiary: University of Nottingham
Dissemination level: PUBLIC

Due date: 01/05/2016
Actual submission date: 20/05/2016

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

Document History

Version	Contributor(s)	Partner	Date	Comments
0.1	Marco Pasi	UNOT	27/04/16	First draft
0.2	Marco Pasi	UNOT	16/05/16	Final draft
0.3				

The specific sequence of bases that constitute a DNA molecule influence significantly its properties, in particular its structure and its flexibility. These effects are expected to play an important role in how proteins and drugs interact with DNA, as well as in the design and optimisation of DNA-based nanomaterials. Indeed, a number of proteins have been confirmed to rely on the structure and fluctuations of DNA to recognise their target binding sequence. Specific protein binding to DNA is vital for a number of fundamental cellular processes, such as DNA replication and repair, chromatin compaction and transcription regulation.

Molecular Dynamics (MD) simulations have been used to integrate experimental data in order to make systematic studies of how sequence affects DNA. In particular, the Ascona B-DNA Consortium (or ABC) has pioneered this field (Beveridge et al., 2004; Dixit et al., 2005) and managed to recently obtain microsecond-scale MD trajectories of a purposely designed set of oligomers, which provides complete information on DNA sequence effects up to tetranucleotides (Lavery et al., 2010; Pasi et al., 2014).

The ABC strategy relies on packing the 136 distinct[1] tetranucleotide sequences in 39 B-DNA oligomers of 18 bp, each constructed according to the pattern 5'-gc-CD-ABCD-ABCD-ABCD-gc-3', where upper case letters indicate sequences that vary between oligomers and lower case letters indicate fixed sequences (dashes have been added for clarity). Microsecond-scale MD simulations were performed on each oligomer in explicit water and physiological salt concentration, using state-of-the-art force fields and protocols. The analysis of this data allowed to establish that tetranucleotides are the minimal-length unit of sequence required to accurately describe sequence effects on B-DNA.

This large set of MD data, encompassing more than 9 TB of stored trajectories, and just under 36 million DNA structures, has been analysed to provide to the users of the MuG VRE a complete description of the structure and flexibility of DNA, as a function of its local tetranucleotide sequence. The results of these analyses have been uploaded to the MuG section of the BIGNASim repository (Hospital et al., 2015), which complies with the requirements set out in the Data Management Plan (https://3.basecamp.com/3126297/buckets/97795/uploads/122576165) for this type of

[1] In this context, a tetranucleotide is a 4bp double-stranded fragment of B-DNA featuring canonical Watson-Crick base pairing. It follows that the sequence of bases on one strand is immediately implied by the sequence of bases on the other strand, and therefore each tetranucleotide can be uniquely defined by the sequence of bases on either strand, by convention read in the 5'-3' direction.

data, and are accessible through the MuG interactive flexibility browser (http://www.multiscalegenomics.eu/flexBrowser.html), which is part of the MuG portal. More complete integration within the visualisation and analysis tools of the MuG VRE are expected as an outcome of Milestone 23.

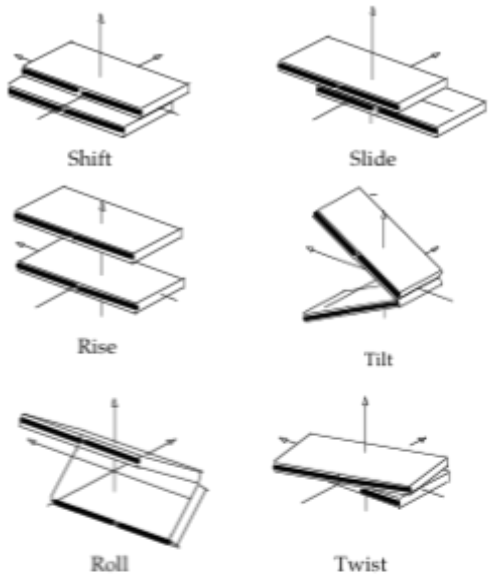

Figure 1. Inter-base pair helical parameters used to describe the relative position and orientation of a base pair with respect to a neighbouring, reference base pair, using three translations (Shift, Slide and Rise), in Å, and three rotations (Tilt, Roll and Twist), in degrees. Specifically, the reference base pair is the one including the base located on the 5' side of the reference strand.

To analyse DNA flexibility, the conformational variations of DNA are described as a function of the well-known and widely used Curves+ helical parameters of DNA (Lavery et al., 2009). These represent a set of independent coordinates, as a function of which averages and covariances were calculated along the microsecond MD trajectories (Hospital et al., 2013). Results shown in the flexibility browser pertain to the inter-base pair helical parameter of the central dinucleotide of each of the 136 distinct tetranucleotides (Figure 1). The choice of representing DNA flexibility in this way was defined taking into account the specific requirements of Pilot Projects 7.2 and 7.3 (WP7), who will be the main beneficiaries of this data within the MuG project, as well as its use within Task 6.2. In particular, a coarse-grain model of DNA, that is being developed as part of Pilot Project 7.2, is using this same definition of flexibility in order to estimate the elastic deformation energy of DNA.

References

Beveridge,D.L., *et al.* (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. Biophys. J., 87, 3799–3813.

Dixit,S.B., *et al.* (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. Biophys. J., 89, 3721–3740.

Lavery,R., Zakrzewska,K., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dixit,S., Jayaram,B., Lankas,F., Laughton,C. *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.

Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F. *et al.* (2014) □ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. Nucleic Acids Res., 42, 12272–12283.

Hospital,A., Faustino,I., Collepardo-Guevara,R., Gonzalez,C., Gelpi,J.L. and Orozco,M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. Nucleic Acids Res., 41, W47–W55.

Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Res., 37, 5917–5929.

Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Goñi,R., Orozco,M. and Gelpí,J.L. (2016) BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. Nucl. Acids Res., 44, D272-D278