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 7.3: Report on the use of MuG VRE on the integration of DNA simulation data

Lead beneficiary: Centre National de la Recherche Scientifique (CNRS)

Dissemination level: PUBLIC

Due date: 31/10/2018

Actual submission date: 06/11/2018

Copyright© 2015-2018 The partners of the MuG Consortium

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	30/10/2018	First draft
0.2	Charles Laughton	UNOT	30/10/18	Review
0.3	Marco Pasi	UNOT	05/11/2018	Minor correction in section 3.2
1.0			06/11/2018	Approved by Supervisory Board

Table of Contents

1 Executive summary

Extensive MD simulations on a set of transcription factor-DNA complexes have been performed and analysed using a range of VRE tools. The current document presents an outline of our feedback on the usability and completeness of the current VRE tool offering for the analysis and integration of MD simulation data on transcription factor-DNA complexes. A detailed report of the outputs from each VRE tool utilised is presented, as well as a brief discussion of external tools that were required in areas where there is currently no VRE equivalent. Finally, the document offers some recommendations for improvement and enhancement of VRE functionality.

2 Introduction

Eukaryotic transcription factors (TFs) play a central role in regulatory networks within the genome. They recognise DNA in a specific manner, and the mechanisms that steer this specificity have been identified for many TFs based on 3D structures of TF–DNA complexes.[1] However, exactly how TFs can select their binding sites within a cellular environment *in vivo* has not yet been fully understood. Closely related TFs bind to distinct binding sites to execute different *in vivo* functions, but at the same time the mechanisms that make paralogous TFs to select very similar, but not identical, binding sites are not completely understood.[1]

Reading and recognition of DNA by TFs is done *via* both a direct (base readout) and an indirect (shape readout) way (Figure 1). Most TFs use interplay between the base- and shape-readout modes to recognise their DNA binding sites, although the contribution of each mechanism to TF-DNA binding specificity may vary across different TF families.[1] Understanding the effect of shape-readout is complex, since it requires detailed knowledge of not only the structure of naked and protein-bound DNA, but also of the physical properties of both. It is well known that certain TFs cause DNA bending,[2] and that TF binding can be promoted (or inhibited) by DNA that is under torsional or bending strain (e.g. due to being in a relatively small loop).[3] However, our understanding of the cross-talk between TF-DNA interaction and DNA deformation is based on a very limited amount of data at the atomistic level of detail.

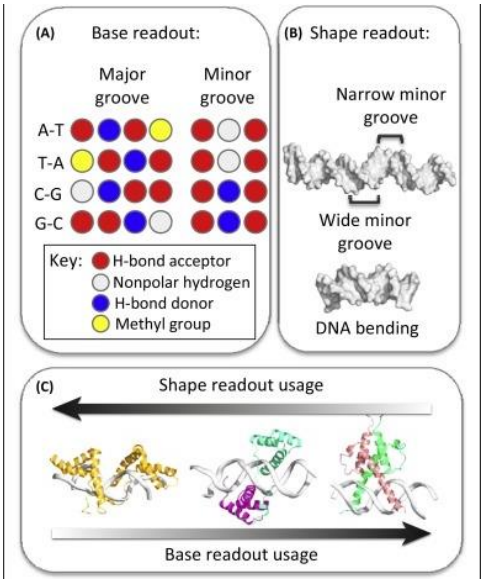


Figure 1. Reading and recognition of DNA by TFs is done via both a direct (A - base readout) and an indirect (B - shape readout) way. Most TFs use interplay (C) between the base- and shape-readout modes to recognise their DNA binding site.[1] Figure from Slattery et. al.[1]

In order to generate insights into the cross-talk between TF-DNA interaction and DNA deformation, we have conducted extensive fully atomistic MD simulations of selected TF-DNA complexes. The systems of choice for this work have been FOXO3[4] (RCSB PDB ID: 2uzk) and FOXA3[5] (RCSB PDB ID: 1vtn). Both are members of the forkhead-box (FOX) TF family.[6] Among other functions, FOXO3 has been shown to participate in cellular senescence,[7] while both FOXO3 and FOXA3 are thought to be interacting with and remodelling chromatin as hybrid and pioneering factors respectively.[6] Further,

FOXA TFs are thought to prefer pre-bent DNA and to also cause significant deformations to the DNA upon binding.[6]

The systems modelled in this work are outlined in Figure 2 below. Each TF has been modelled with its cognate DNA and with the DNA of the other TF, while control experiments have been done by simulating the naked DNA structures. Fully atomistic MD simulations of six replicates of each of these systems have been performed, using GROMACS.[8-10]

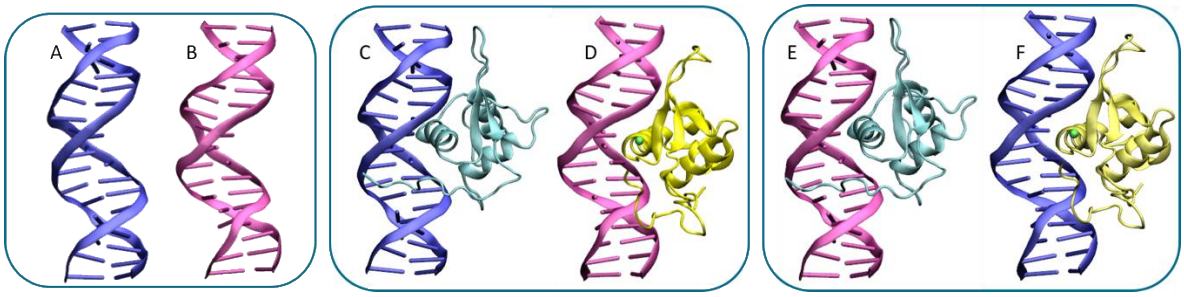


Figure 2. Schematic representation of TF-DNA systems modelled. (A) naked DNA of FOXO3. (B) naked DNA of FOXA3. (C) FOXO3 with its cognate DNA. (D) FOXA3 with its cognate DNA. (E) FOXO3 with FOXA3 DNA. (F) FOXA3 with FOXO3 DNA.

The two TFs have been crystallised with different DNA sequences (Figure 3). FOXO3 has been crystallised[4] with DNA containing the FOXO consensus sequence GTAAACAA, while FOXA3 has been crystallised[5] with DNA containing the strong transthyretin promoter binding site TAAGTCA.

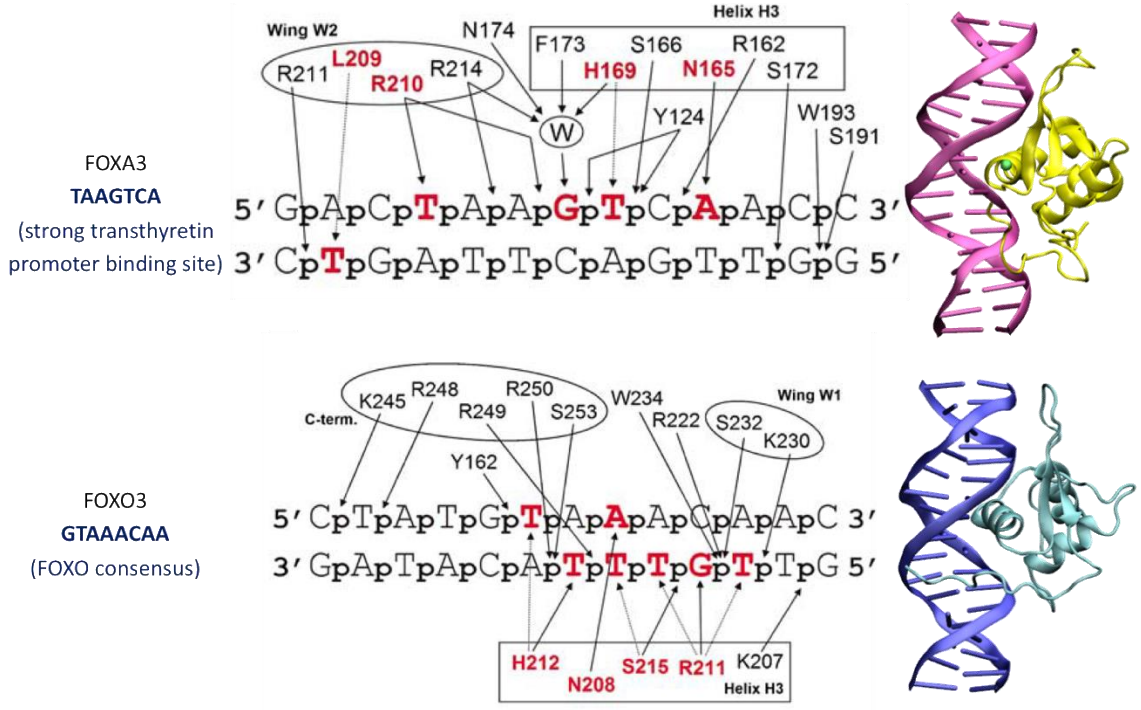


Figure 3. FOXA3 and FOXO3 cognate DNA and TF-DNA interactions. The residues and bases that participate in specific contacts between TFs and DNA are shown in red, polar interactions are shown with arrows, and Van der Waals contacts are shown as dashed arrows.[11] Figure modified from Obsil and Obsilova.[11]

3 Feedback on the usage of VRE tools

3.1 MDweb

by Andrio Pau (BSC), Hospital Adam (IRB), and Gelpí Josep Lluis (BSC)

MDWeb is a molecular dynamics workflow to energetically minimise a 3D structure. The user uploads a single pdb structure and the tool follows the worfklow outlined in the schematic representation below (Figure 4) in order to output a refined structure that can be used for visualisations or further simulations. Integration is currently done using GROMACS[8-10] package MD tools.

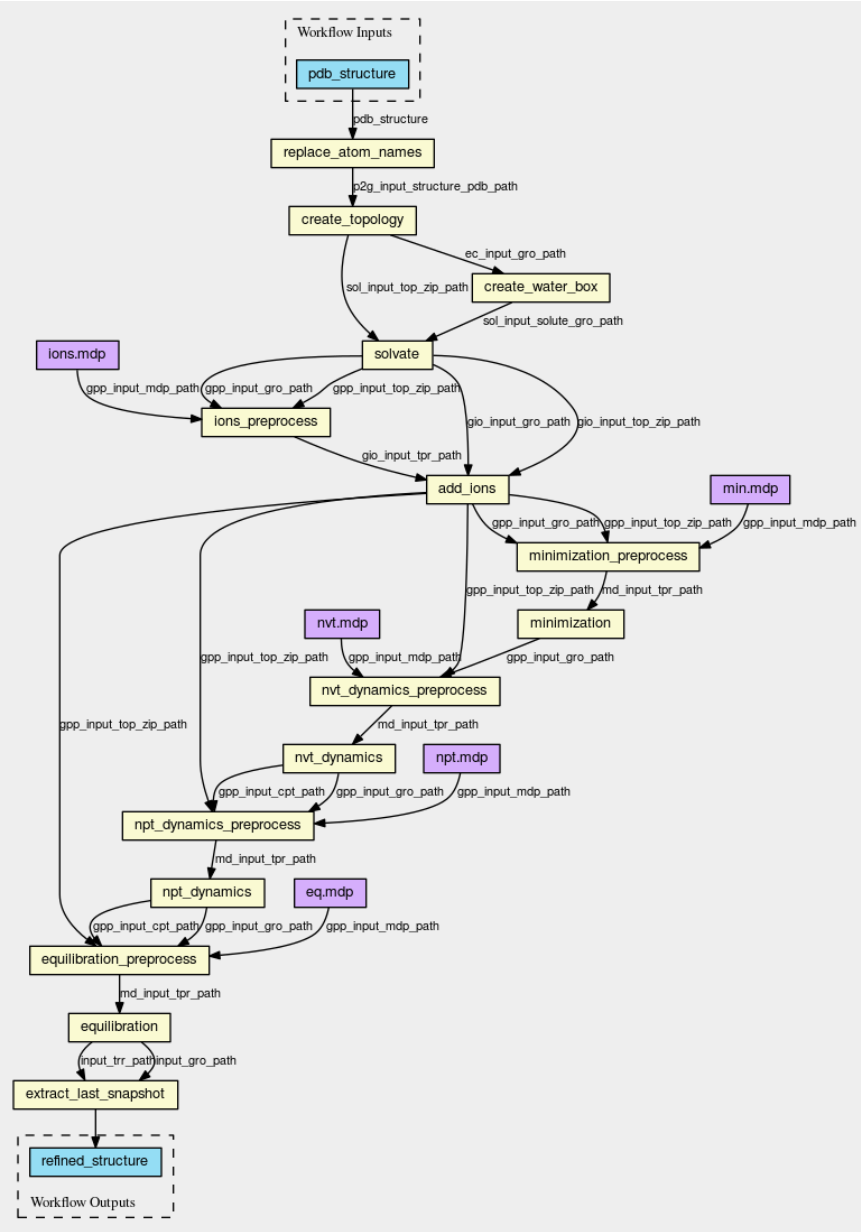


Figure 4. MDWeb workflow. From the MDWeb help section on the MuG VRE: "Optimised structures often correspond to a substance as it is found in nature. Finding configurations for which the energy is a minimum, that is, finding a point in configuration space where all of the forces on the atoms are balanced, is a usual step to perform after modelling a structure. Stable conformations of a molecule can be identified by simply minimising its energy."

MDweb has been used for the energy refinement of FOXO3.pdb. MDWeb is a user-friendly energy refinement tool that only requires a single pdb structure as an input file. The tool in its current form only outputs a single pdb file of the refined 3D structure, while all the equilibration results and, crucially, topology files are only produced in the background and are not currently written to the user workspace. It could therefore be argued that MDWeb may not be currently accomplishing its full potential as a simulation setup tool, but it nevertheless offers a very easy-to-use interface for users who do not have access to a working GROMACS[8-10] installation, as well as users who are not comfortable using tools without a GUI, and who want to generate an energetically minimised pdb structure for visualisation, analysis, or further simulations. A new version of MDWeb that writes all the interim files in the user worskpace is being developed. Such an enhancement would decisively increase MDWeb's applicability and would upgrade the tool into a complete simulation setup suite.

```
_______MUG REFINEMENT_______

2018-10-29 17:39:55,613 [MainThread  ] [INFO ]  in ----------- Get PDB structure
2018-10-29 17:39:55,650 [MainThread  ] [INFO ]  sed ---------- Replacing atom names
2018-10-29 17:39:55,784 [MainThread  ] [INFO ]  pdb2gmx ------ Create gromacs topology
2018-10-29 17:39:56,488 [MainThread  ] [INFO ]  editconf ----- Define box dimensions
2018-10-29 17:39:56,578 [MainThread  ] [INFO ]  solvate ------ Fill the box with water molecules
2018-10-29 17:39:58,176 [MainThread  ] [INFO ]  grompp_ions -- Preprocessing: Adding monoatomic ions
2018-10-29 17:39:58,862 [MainThread  ] [INFO ]  genion ------- Running: Adding monoatomic ions
2018-10-29 17:39:58,865 [MainThread  ] [INFO ]                     To neutralize the system charge
2018-10-29 17:39:59,331 [MainThread  ] [INFO ]  grompp_min --- Preprocessing: Energy minimization
2018-10-29 17:40:00,517 [MainThread  ] [INFO ]  mdrun_min ---- Running: Energy minimization
2018-10-29 17:40:21,383 [MainThread  ] [INFO ]  grompp_nvt --- Preprocessing: nvt constant number of molecules, volume and temp
2018-10-29 17:40:22,472 [MainThread  ] [INFO ]  mdrun_nvt ---- Running: nvt constant number of molecules, volume and temp
2018-10-29 17:41:54,802 [MainThread  ] [INFO ]  grompp_npt --- Preprocessing: npt constant number of molecules, pressure and temp
2018-10-29 17:41:56,272 [MainThread  ] [INFO ]  mdrun_npt ---- Running: npt constant number of molecules, pressure and temp
2018-10-29 17:42:07,973 [MainThread  ] [INFO ]  grompp_eq ---- Preprocessing: 100ps Molecular dynamics Equilibration
2018-10-29 17:42:08,913 [MainThread  ] [INFO ]  mdrun_eq ----- Running: 100ps Molecular dynamics Equilibration
```

Figure 5. MDWeb output. Top: refined structure of FOXO3.pdb, visualised with NGLViewer,[12-13] also integrated within the MuG VRE. Bottom: log file outlining GROMACS steps in the MDWeb workflow.

3.2 3DConsensus

By Marco Pasi (UNOT)

3DConsensus is a tool for the analysis of protein-DNA interactions. 3DConsensus calculates a sequence-only consensus from the experimental data, as well as a consensus based on the physical properties of DNA, by leveraging extensive atomic-detail information on the sequence-dependent behaviour of naked DNA.[14] It uses Curves+[15] to analyse the 3D structure of a protein-DNA complex to

extract conformational parameters, identify interactions, and study their impact on specific binding by integrating experimental data on the protein's DNA specificity. Analysis of the helical parameters, as these are outlined in the schematic representation in Figure 6 below, allows the user to evaluate the extent and direction of DNA deformations in the protein-DNA complex.

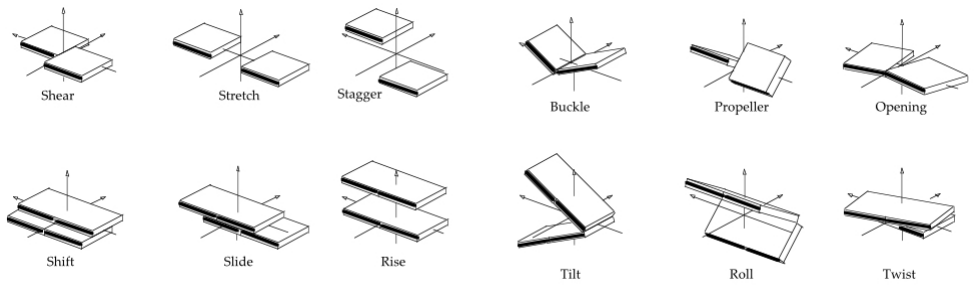


Figure 6. DNA helical parameters.

3DConsensus outputs plots for various helical parameters[15] (Figure 6) and shape consensus, as well as offers a variety of structure visualisation options. Selected 3DConsensus outputs can be seen in Figure 7 below.

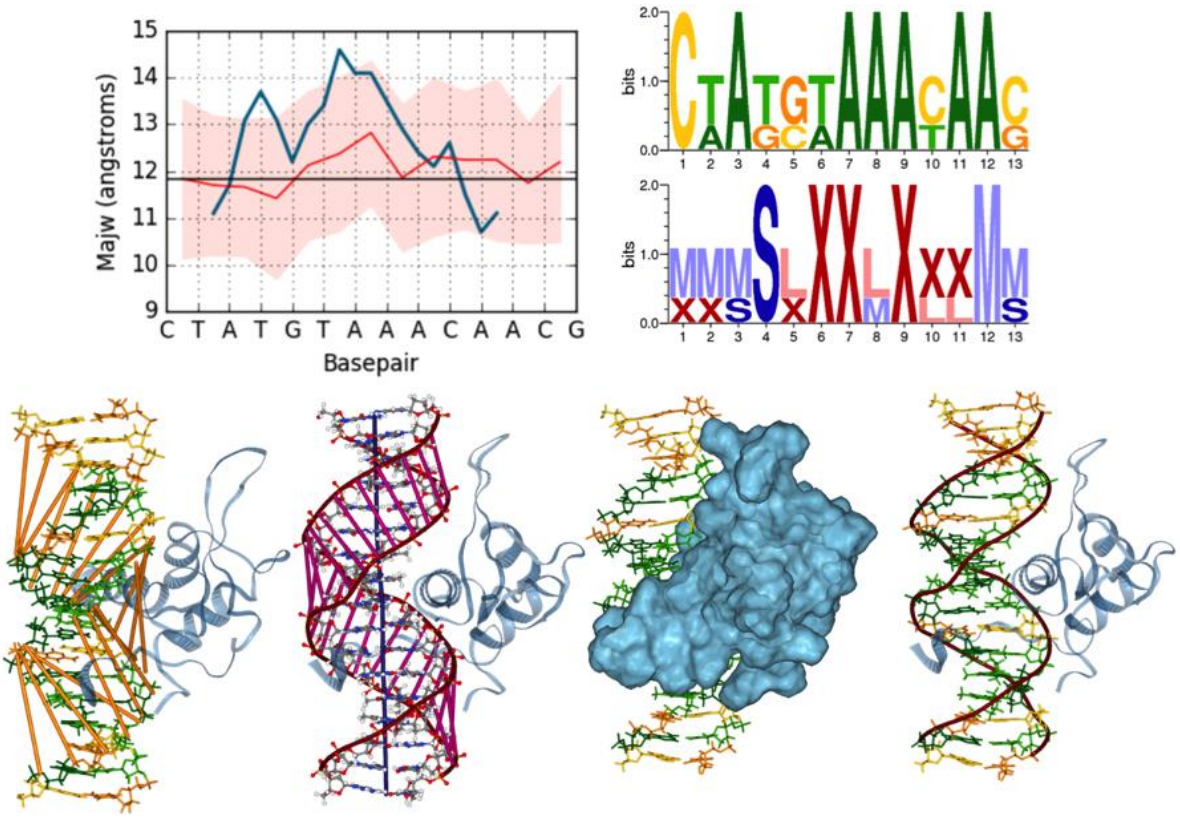


Figure 7. 3DConsensus output. Top left: selected helical parameters results. Top right: selected sequence and shape consensus results. Bottom: selected different options for visualisations of the target structure.

3DConsensus generates detailed metrics that can be utilised for comparisons between snapshots from different systems. Although it does not currently offer the option of trajectory analyses, it nevertheless constitutes an analysis suite integrating atomistic level DNA data with experimental data in a way that provides the user with useful insights for the interpretation of the protein's binding specificity.

3.3 NAFlex

by Adam Hospital (IRB Barcelona)

NAFlex[16-17] is a tool for the analysis of nucleic acids structures or trajectories, either atomistic or coarse-grained. The user can either upload their own trajectory or create one *in situ* using MC DNA. Using Curves+,[15] NAFlex can produce a complete analysis of nucleic acids helical parameters. Further, NAFlex can deliver detailed metrics on principal components,[18] hydrogen bonds, distance contacts, and NMR observables. Currently it is only double-stranded, standard nucleic acids that can be analysed with the NAFlex version integrated within the MuG VRE. Mis-paired nucleotides, single-stranded nucleic acids, triplexes, and quadruplexes are not currently supported. Finally, modified nucleotides are only allowed in some of the flexibility analysis operations.

Selected plots from NAFlex analyses can be seen in Figure 8 below. NAFlex produces detailed metrics that can be utilised for comparisons between snapshots from different systems, as well as trajectory analyses. Currently NAFlex is the only tool in the MuG VRE doing trajectory analyses, which makes it an excellent option for users interested in gaining insights into the flexibility of nucleic acids over time during a molecular dynamics simulation.

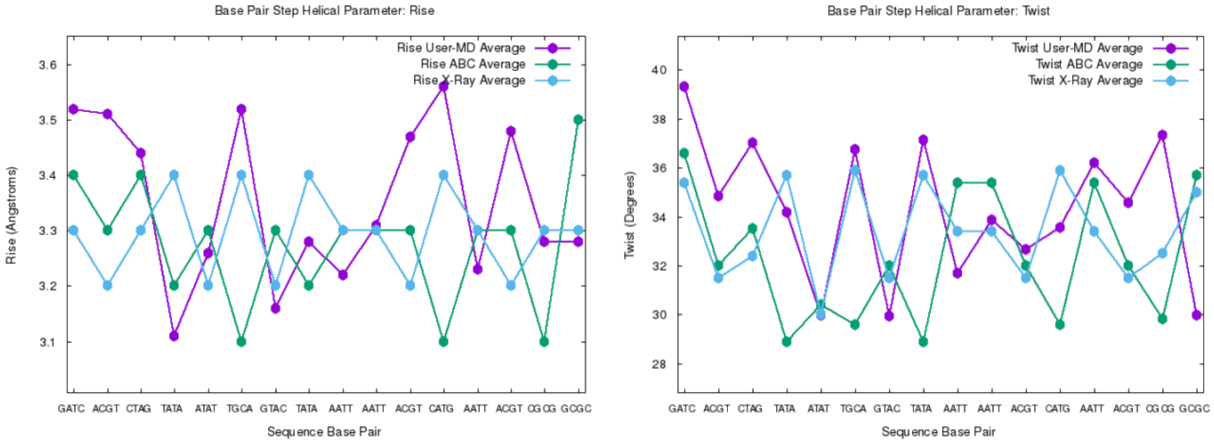


Figure 8. Selected plots from NAFlex output.

3.4 pyDockDNA

by Brian Jiménez (BSC)

pyDockDNA[19] is a tool for the structural prediction of protein-DNA interactions. Starting from the 3D coordinates of an interacting protein and DNA molecule, the tool outputs the best rigid-body docking orientations as these are generated by FTDock[20] and evaluated by the scoring function of pyDockDNA, which includes electrostatics energy and limited van der Waals contribution.

An example of pyDockDNA[19] output can be seen in Figure 9 below. The output is an energy table outlining the calculated energy terms (electrostatics, desolvation, and van der Waals) for each of the predicted models. The tool also offers the option to download various files generated by the method in a compressed TAR-GZ folder. The compressed folder contains, among other files, the top scoring predicted models in PDB format, as well as a plain text file containing the energy table for the top 10,000 poses predicted by the method. pyDockDNA was an excellent option for this experiment in

order to evaluate and compare the predicted poses against the conformations produced by the random-seed velocity generator in GROMACS.[8-10] pyDockDNA[19] generated the predicted poses quickly and the fact that it offers the option for a direct download of the produced structures in PDB file format certainly increases the applicability and ease-of-use of the tool.

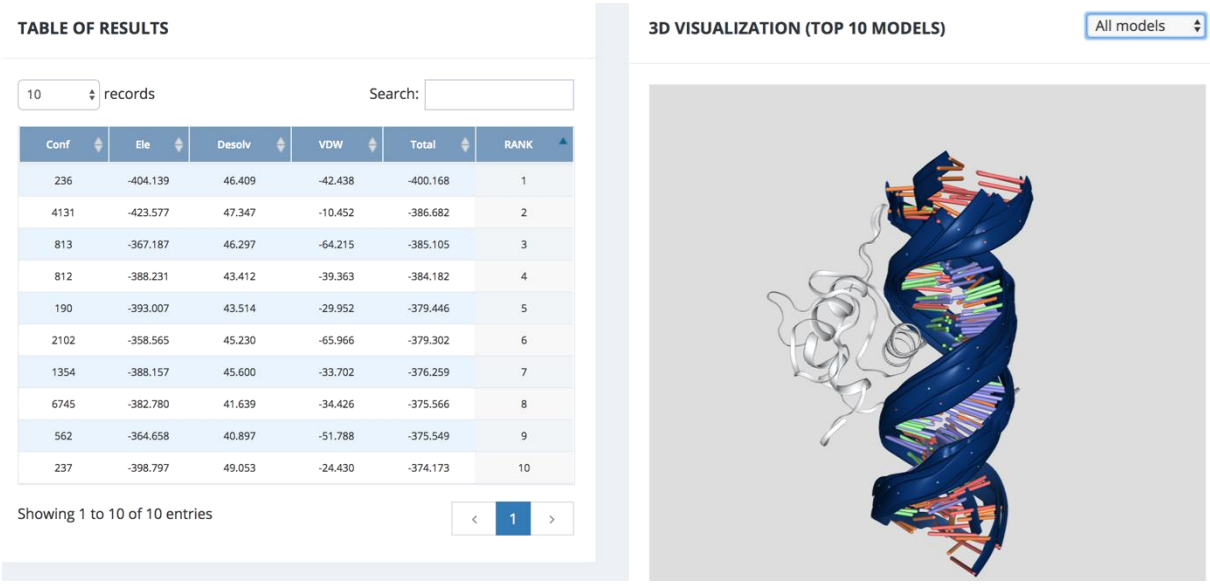


Figure 9. pyDockDNA output. Left: energy table with electrostatics, desolvation and Van der Waals terms. Right: the top ten models of the docking.

3.5 MCDNA

by Jürgen Walther (IRB)

This tool creates 3D all-atom B-DNA conformations of a sequence of interest, either as a single structure or as a molecular dynamics-like trajectory. The models are based on a Metropolis Monte Carlo algorithm with bp resolution. The user only needs to provide a simple text file containing the DNA sequence of interest. MC DNA was utilised in the current task in order to evaluate and compare the single DNA structure in a relaxed state, defined as the state of minimum potential energy according to the bp-step parameters, as it was generated by the tool, against the DNA conformations resulting from GROMACS[8-10] simulations. Further MC DNA outputs (Figure 10) plots of bending properties that can be useful in preliminary analyses of DNA bending.

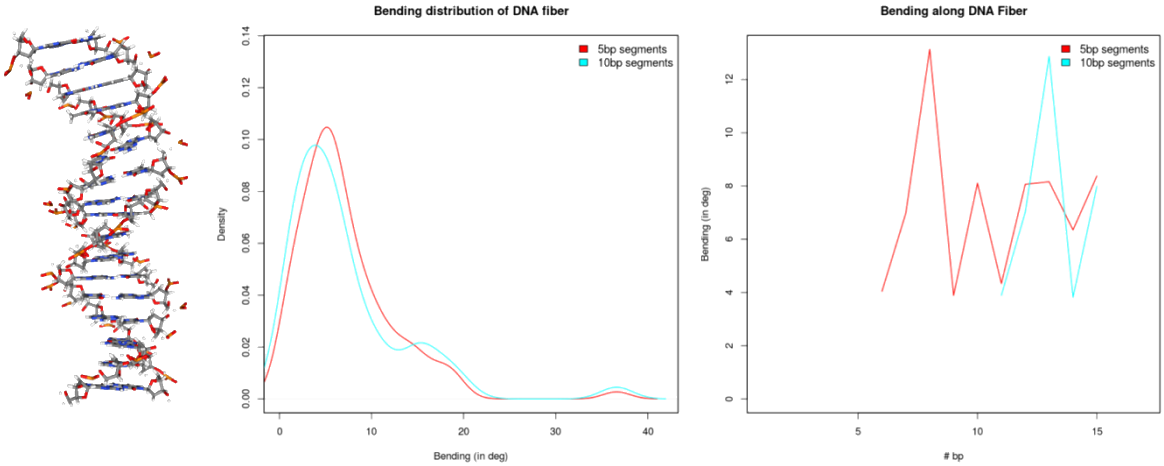
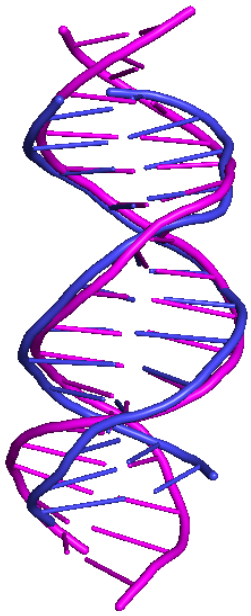


Figure 10. MC DNA output. DNA 3D structure and plots outlining bending properties of the DNA input structure.

4 External tools used

During the initial simulation setup stages and while aligning the two protein-DNA complexes using PyMOL,[21] it was clear that they needed to be elongated by adding extra bases on their terminal ends in order to become of equal length. This was done in order to remove the bias introduced by the different lengths, so as to be able to conduct unbiased comparative modelling studies. The DNA sequence of FOXA3 was elongated on the 3' end, whereas the DNA sequence of FOXO3 was elongated on both 5' and 3' ends. A schematic representation of the initial alignment, the added bases (in red and lower case), as well as the initial DNA structure alignment in PyMOL can be seen in Figure 11 below. Further, base mutations (in square boxes, seen in Figure 11), were performed in order to be able to swap the two DNA sequences between the two protein-DNA complexes. The DNA rebuilding tool that was used both for DNA elongation and base mutations was x3DNA.[22-24]



G A C T A A G T C A A C C a c g c
g a C T A T G T A A A C A A C g c

Figure 11. Left: initial alignment of DNA structures in PyMOL.[21] The length of FOXA3 DNA (magenta) stretches beyond that of FOXO3 DNA (purple). Right: schematic representation of the initial alignment, and outline of bases added (red font, lower case). In squares: base mutations performed.

In addition to the external tools (PyMOL[21] and x3DNA[22-24]) already mentioned, other external tools utilised for this work were Gromacs[8-10] and VMD.[25]

5 Conclusions

The current document presents an outline of our feedback on VRE usage specifically for the study of TF-DNA simulations complexes. Overall, and in terms of user friendliness, the MuG VRE has been shown to be very easy to use, it is highly intuitive for first-time users, and the layout of the user workspace is pleasant and efficient. In terms of tool selection, it is our opinion that the MuG VRE has a sufficient number of tools for our needs, namely: for the analysis of the interactions of protein-DNA complexes (3DConsensus, pyDockDNA[19]), for the analysis of DNA trajectories and structures (NAFlex,[16] MCDNA), for simulation setup and energy refinement (MDWeb), while also offering a native 3D viewer (NGLViewer[12-13]).

Potential suggestions for improvement of the current toolkit deployed on the VRE would be expanding the functionality of MDWeb to transform it into a complete simulation setup tool, as has already been discussed earlier in the text. Additionally, we have identified a need for more MD trajectory analysis tools, as currently there is only a single tool (NAFlex[16]) fulfilling this role. Finally, and taking into consideration our needs for external tools (particularly for DNA rebuilding tools such as x3DNA[22-24]) as those were outlined earlier in this document, we have identified a potential for deployment of a DNA rebuilding tool. Further, more sophisticated visualisers, that directly allow the user to render publication-quality images within the VRE environment, is another potential for future tool deployment. The addition of such tools would further enhance the usability of the VRE.

6 References

1. Slattery, M.; Zhou, T.; Yang, L.; Dantas Machado, A. C.; Gordân, R.; Rohs, R. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences* **2014**, *39*, 381-399.
2. Kerppola, T. K.; Curran, T. Fos-Jun heterodimers and jun homodimers bend DNA in opposite orientations: Implications for transcription factor cooperativity. *Cell* **1991**, *66*, 317-326.
3. Yuan, H. S.; Finkel, S. E.; Feng, J. A.; Kaczor-Grzeskowiak, M.; Johnson, R. C.; Dickerson, R. E. The molecular structure of wild-type and a mutant Fis protein: relationship between mutational changes and recombinational enhancer function or DNA binding. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 9558-9562.
4. Tsai, K.-L.; Sun, Y.-J.; Huang, C.-Y.; Yang, J.-Y.; Hung, M.-C.; Hsiao, C.-D. Crystal structure of the human FOXO3a-DBD/DNA complex suggests the effects of post-translational modification. *Nucleic Acids Res.* **2007**, *35*, 6984-6994.
5. Clark, K. L.; Halay, E. D.; Lai, E.; Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **1993**, *364*, 412.
6. Lalmansingh, A. S.; Karmakar, S.; Jin, Y.; Nagaich, A. K. Multiple modes of chromatin remodeling by Forkhead box proteins. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **2012**, *1819*, 707-715.
7. Xie, Q.; Peng, S.; Tao, L.; Ruan, H.; Yang, Y.; Li, T.-M.; Adams, U.; Meng, S.; Bi, X.; Dong, M.-Q.; Yuan, Z. E2F Transcription Factor 1 Regulates Cellular and Organismal Senescence by Inhibiting Forkhead Box O Transcription Factors. *J. Biol. Chem.* **2014**, *289*, 34205-34213.
8. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435-447.
9. Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J Mol Model* **2001**, *7*, 306-317.
10. van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701-1718.
11. Obsil, T.; Obsilova, V. Structure/function relationships underlying regulation of FOXO transcription factors. *Oncogene* **2008**, *27*, 2263.
12. Rose, A. S.; Hildebrand, P. W. NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.* **2015**, *43*, W576-W579.
13. Rose, A. S.; Bradley, A. R.; Valasatava, Y.; Duarte, J. M.; Prlić, A.; Rose, P. W. NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics* **2018**, *34*, 3755-3758.
14. 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.
15. Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **2009**, *37*, 5917-5929.
16. 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.
17. Hospital, A.; Andrio, P.; Cugnasco, C.; Codo, L.; Becerra, Y.; Dans, P. D.; Battistini, F.; Torres, J.; Goñi, R.; Orozco, M.; Gelpí, J. L. BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.* **2016**, *44*, D272-D278.
18. Meyer, T.; Ferrer-Costa, C.; Pérez, A.; Rueda, M.; Bidon-Chanal, A.; Luque, F. J.; Laughton, C. A.; Orozco, M. Essential Dynamics: A Tool for Efficient Trajectory Compression and Management. *J. Chem. Theory Comput.* **2006**, *2*, 251-258.
19. Rodríguez-Lumbreras, L. Á.; Jiménez-García, B.; Fernández-Recio, J. pyDockDNA: a new approach for protein-DNA docking. *In Book of abstracts, Barcelona Supercomputing Center* **2017**, 49.

20. Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. E. Modelling protein docking using shape complementarity, electrostatics and biochemical information11Edited by J. Thornton. *J. Mol. Biol.* **1997**, *272*, 106-120.
21. The PyMOL Molecular Graphics System, Version 2.2.0, Schrödinger, LLC.
22. Colasanti, A. V.; Lu, X.-J.; Olson, W. K. Analyzing and building nucleic acid structures with 3DNA. *Journal of visualized experiments : JoVE* **2013**, e4401-e4401.
23. Lu, X.-J.; Olson, W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols* **2008**, *3*, 1213.
24. Zheng, G.; Lu, X.-J.; Olson, W. K. Web 3DNA—a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.* **2009**, *37*, W240-W246.
25. Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **1996**, *14*, 33-8, 27-8.