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 6.2: Software Tools for Protein-DNA Interactions

Lead beneficiary: University of Nottingham
Dissemination level: PUBLIC

Due date: 31/10/2016
Actual submission date: 31/10/2016

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	14/10/16	First draft
0.2	Brian Jimenez	BSC	21/10/16	Comments on tools, typo fixing
0.3	Juan Fernández-Recio	BSC	21/10/16	Fixed typos
0.4	Marco Pasi	UNOT	24/10/16	Final details
1.0	-	-	31/10/16	Final version approved by technical and supervisory boards.

Table of Contents

1 EXECUTIVE SUMMARY

Protein-DNA interactions play a fundamental role in shaping the regulatory processes in the cell. The integrative approach that the MuG VRE aims to implement has the potential of overcoming many of our limitations in understanding and predicting these interactions, and ultimately to understand chromatin structure and gene expression regulation. In this document we define, with sustainability in mind, which tools to predict and analyse protein-DNA interactions will be initially available in the MuG VRE, with particular attention to *(i)* what data will be retrieved and stored in the data repositories, and exchanged among tools, as well as to *(ii)* how the tools will be integrated in the computational infrastructure of the VRE.

2 INTRODUCTION

It has become clear in recent years that many human diseases are caused by malfunction in the complex intracellular processes that regulate gene expression; these include cancer, cardiovascular disease, diabetes, autoimmunity, neurological disorders, and obesity [Lee 2013]. Our ability to devise effective treatments for some of these conditions is greatly limited by our poor understanding of gene regulation and transcriptional control. DNA-binding proteins play a fundamental role in these processes, and in particular transcription factors (TFs) regulate the expression of specific genes by binding with high affinity to their "consensus" sequence (the TF binding site or TFBS, e.g. promoters or enhancers).

Unraveling the molecular details of TF sequence specificity is a major challenge in deciphering the spatiotemporal gene expression patterns that sustain life at the cellular level. However, very early structural information available on protein-DNA complexes made it clear that it is impossible to define a simple code for protein-DNA binding, based exclusively on contacts between specific amino acid side chains and specific DNA bases [Luscombe 2000, Jen-Jacobson 2000, Garvie 2001, Matthews 1998]. Although predictive models that rely exclusively on DNA sequence have been successful in a large number of cases [Stormo 1982, Stormo 2013], these are unable to account for a variety of structural and dynamic factors that affect the binding affinity [O'Flanagan 2005], but are difficult to predict from DNA sequence alone. These effects are often collectively referred to as "indirect readout" [Lavery 2005, Sarai 2005], in contrast to "direct readout" which instead depends directly upon sequence, and their understanding requires detailed knowledge of the structure and flexibility of both interaction partners. To complicate things further, both theoretical and experimental studies have shown that protein-DNA interfaces are highly dynamic [Zandarashvili 2013, Chen 2015, Etheve 2015], and therefore a single structure is often insufficient to analyse recognition, and in particular to identify the key residues making the most important contributions to recognition. In this context atomistic molecular simulations have reached a level of accuracy that allows them to be a powerful tool to complement the available experimental structural information to study in detail the mechanical properties of the interaction partners alone [Beveridge 2004, Dixit 2005, Lavery 2010, Pasi 2014, Dans 2014], as well as how they are affected by their binding.

Several other features contribute to the specific binding of proteins to DNA, generating a complexity which severely limits our ability to predict TFBSs and ultimately to understand the regulatory role of each TF [Slattery 2015]. These include cooperative binding of multiple TFs, and the spatiotemporal variations in accessibility of chromatin, nucleosome occupancy, and DNA methylation. Integrating all this information is an ambitious goal of the MuG project, and the MuG VRE is the perfect environment to establish and exploit these connections.

In this document a full description is provided of the software tools relevant for the study of protein-DNA interactions within MuG (see Figure 1 for a summary). On the basis of the guidelines defined in the software architecture specifications (See Deliverable 6.1 [D6.1]), a tool within the VRE is defined by four parameters: *(i)* functionality, *(ii)* required input data, *(iii)* resulting output data and *(iv)* type of integration within the VRE and execution environment requirements. In the following sections, each tool will be defined in these terms, with particular focus to how it interacts with data management (points *ii* and *iii*) and with the computational infrastructure (point *iv*).

3 PROTEIN-DNA TOOLS

3.1 General considerations

Figure 1 shows the protein-DNA interaction (PDI) information network in MuG, which summarises the flow of information related to PDIs within the MuG VRE. In particular, it defines the main types of data that are pertinent to studying PDIs, connected through the relevant tools available in the VRE: details on these entities are provided in the following sections.

Connecting heterogeneous data from different sources and bridging the gap between the various size and time-scales relevant to the understanding of multiscale genomics is at the heart of the ambition of the MuG VRE. In this spirit, the PDI information network connects sequence information on proteins and DNA to structural representations of these macromolecules and their complexes at the atomistic scale, and then further, through molecular simulations to coarse-grain models of chromatin at various scales, and through binding affinity analysis to genome-wide protein-DNA binding specificity analysis (ChIP-seq).

The field of 4D genomics is experiencing a phase of exponential growth, and this results in a high rate of appearance of new data and new tools. Given this situation it is likely that a given tool might be supplanted by a newer, more performing tool in the near future. Furthermore, during an initial requirement survey carried out within the MuG project [IC6.0], it was pointed out by several prospective users that the VRE should include multiple tools to perform the same scientific tasks, in order to leave the freedom of choice to the user. For these reasons, tools are grouped in **Tool Categories** that share a common functionality, and the PDI network is designed to define the relationship among these tool categories, to establish a structure that is more sustainable and likely to be persistent for longer. Each tool category is therefore characterised by a particular functionality and by a specific input and a specific output **Data Type** (see Section 3.2), as outlined in Figure 1 and described in detail in Section 3.3. For each category, one or more specific tools are implemented in the VRE using the most appropriate strategy according to the guidelines defined in the software architecture specification [D6.1]: details on the integration of each tool in the computational infrastructure of the VRE are given in Section 3.4.

The choice of tools to integrate in the VRE was made primarily based on their functionality, making sure they are the best match for the envisaged use cases within the field of multiscale genomics. Tools that are in active development and that benefit from a large and lively user community were preferred. Other choice criteria include the tool's simplicity of integration within the VRE and its ease of licensing for community usage.

3.2 Data types

Data will be stored and annotated according to the specifications set in Deliverables 4.1 (Data types, Processing and Data Model Specification [D4.1]) and 4.2 (Data Management Plan [D4.2]). In the following list, the data types mentioned in the protein-DNA interaction (PDI) information network (Figure 1) are listed, with references to relevant sections of the Data Management Plan (DMP).

- **DNA Sequence, Protein Sequence**: see Section 3.3 of the DMP.
- **DNA Structure, Protein Structure, Protein-DNA Complex, Protein-protein Complex**: Three-dimensional (3D) structure of macromolecules or macromolecular complexes, see Section 3.5 of the DMP [D4.2].
- **Sequence Specificity**: information on a DNA-binding protein's preference to bind to specific DNA sequences. The representation may vary according to the model used to

calculate the specificity; Position Weight Matrices [Stormo 1982] are commonly used, although they have several limitations.
- **ChIP-seq**: Raw sequencing results are stored according to Section 3.3 of the DMP [D4.2].
- **Binding sites**: see Section 3.3 of the DMP [D4.2].

3.3 Tool categories

Tools within each category share three features (see D6.1): *(i)* their functionality, *(ii)* required input data, *(iii)* resulting output. In this section, for each of the tool categories outlined in the PDI information network (Figure 1), these three features are described. The fourth and final feature (type of integration within the VRE and execution environment requirements) is instead specific to each tool and will be described in the next section.

3.3.1 Feature prediction

- **DNA Structure modelling**
  - *Function*: Derive the atomic-detailed structure of a DNA fragment from its sequence
  - *Input data*: DNA sequence
  - *Output data*: DNA structure

- **Protein Structure modelling**
  - *Function*: Derive the atomic-detailed structure of a Protein from its sequence
  - *Input data*: Protein sequence
  - *Output data*: Protein structure

- **DNA Flexibility prediction**
  - *Function*: Predict the flexibility of DNA from its sequence
  - *Input data*: Protein sequence
  - *Output data*: DNA Flexibility information (see Section 4.3.3 of the DMP [D4.2])

3.3.2 Structural analysis

- **DNA Structure analysis**
  - *Function*: Analyse the structure of DNA employing a set of relevant internal coordinates that provide a meaningful description of DNA conformation by simplifying the atomic-detailed information [Lavery 2009].
  - *Input data*: DNA structure
  - *Output data*: plain-text tabular output

- **DNA binding site prediction**
  - *Function*: Predict putative regions on the surface of a protein that might act as interaction interfaces with DNA.
  - *Input data*: Protein structure
  - *Output data*: The output is often output as a likelihood score, encoded as the B-factor column of a PDB structure. Additional plain-text output may be required to further interpret the results.

(continues on page 9)

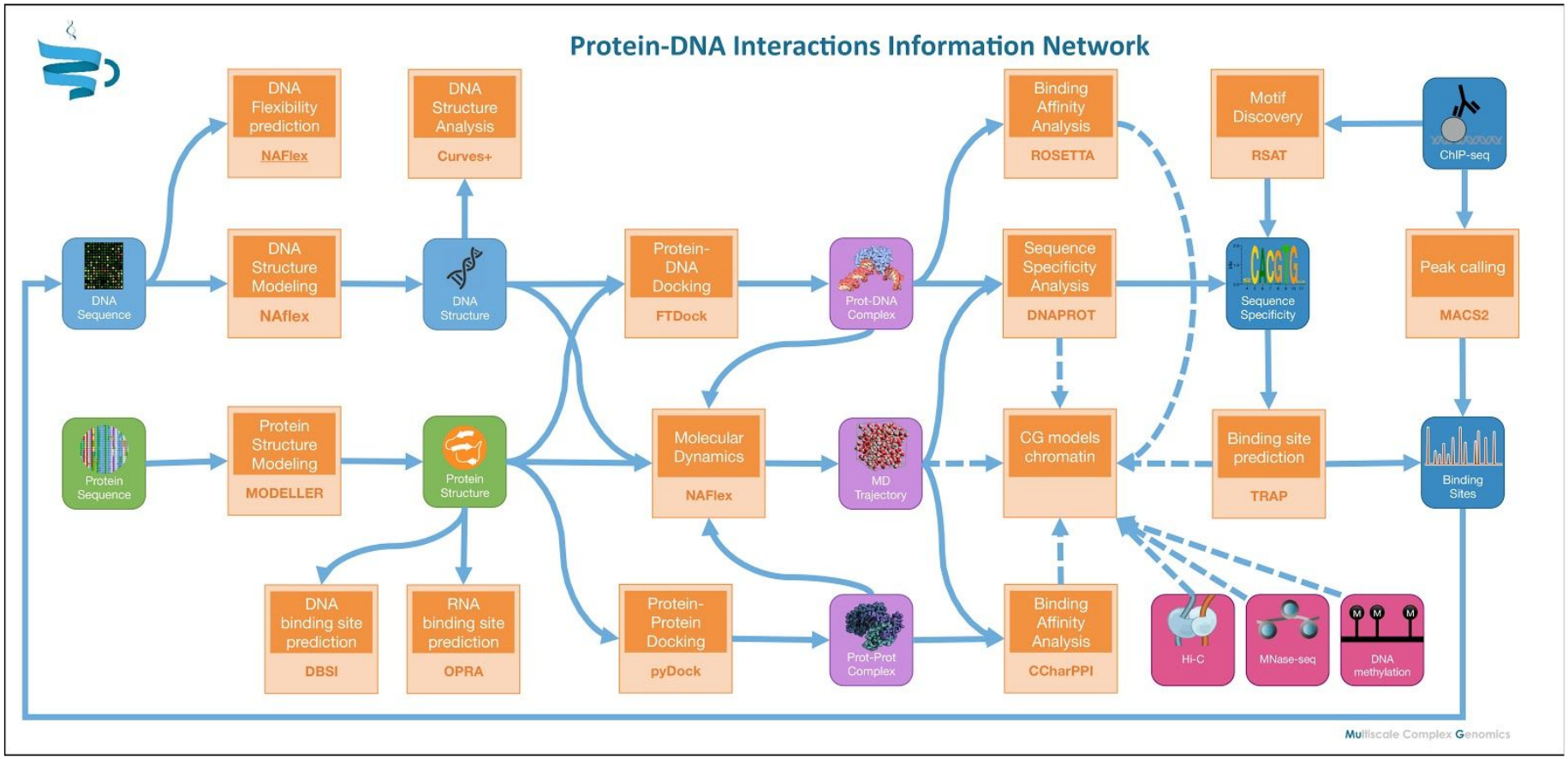

Figure 1. The Protein-DNA Interactions Information Network in MuG. This scheme resumes the flow of information relating to protein-DNA interactions within MuG, by connecting the relevant data types (represented as colored curved rectangles) through the pertinent tools (represented as orange boxes). Tools are described as belonging to a tool category (specified in the shaded box) with the specific tool included in the initial selection specified in the lower part of the box. Dashed arrows represent complex workflows that are included in the scheme but not strictly related to protein-DNA interactions. See the main text for more information.

(...continued from page 7)

- **RNA binding site prediction**
  - *Function*: As for DNA binding site prediction, but for interfaces with RNA.
  - *Input data*: Protein structure
  - *Output data*: Protein structure and plain-text (see DNA binding site prediction)

- **Molecular Dynamics**
  - *Function*: Provide detailed information on the time evolution of the macromolecule or macromolecular complex in physiological conditions, exposing the mechanical, dynamical and thermodynamic properties of the system.
  - *Input data*: 3D structure of the molecule/complex
  - *Output data*: Molecular dynamics trajectory (see Section 3.5 of the DMP [D4.2]); further analyses can be performed on the trajectories using BIGNASim within the VRE [Hospital 2014].

3.3.3 Docking and Affinity analysis

- **Protein-DNA Docking**
  - *Function*: Identify the best conformation (or a set of most likely conformations) for the complex between a target protein and a DNA fragment.
  - *Input data*: Protein structure, DNA structure
  - *Output data*: Protein-DNA complex 3D structure (or a set of structures, grouped as multiple models in a single file)

- **Protein-Protein Docking**
  - *Function*: As for protein-DNA docking, but for complexes between proteins.
  - *Input data*: Two protein structures
  - *Output data*: Protein-protein complex 3D structure (or a set of structures, see Protein-DNA docking)

- **Binding Affinity Analysis (Protein-DNA)**
  - *Function*: Analyse one or more 3D structures of the complex to determine the binding affinity of the two partners.
  - *Input data*: Protein-DNA complex 3D structure or MD trajectory
  - *Output data*: Various plain-text files.

- **Binding Affinity Analysis (Protein-Protein)**
  - *Function*: As for Protein-DNA complexes, but for protein-protein complexes.
  - *Input data*: Protein-protein complex 3D structure or MD trajectory
  - *Output data*: Various plain-text files.

- **Sequence Specificity Analysis (Protein-DNA)**
  - *Function*: Analyse one or more 3D structures of the complex to make predictions about the protein's preferential binding to specific DNA sequences.
  - *Input data*: Protein-DNA complex 3D structure or MD trajectory.
  - *Output data*: Sequence specificity

3.3.4 Motif Discovery and Binding Site prediction

- **Motif Discovery**
  - *Function*: Analyse ChIP-seq data to determine the binding specificity of the target protein-DNA binding protein.
  - *Input data*: ChIP-seq data (in the form of sequencing reads).
  - *Output data*: Sequence specificity.

- **Binding site prediction (genome)**
  - *Function*: Scan a sequence (e.g. a genome) to identify potential binding sites of a DNA-binding protein, given information on its sequence specificity.
  - *Input data*: Sequence specificity, a Sequence.
  - *Output data*: Binding sites.

- **Peak calling**
  - *Function*: Analyse ChIP-seq data to determine precisely to what sequences the target DNA-binding protein was bound.
  - *Input data*: ChIP-seq data (in the form of sequencing reads).
  - *Output data*: Binding sites.

3.4 Tool selection

At least one tool will be integrated in the VRE for each of the categories outlined in the previous section. In this section, the protein-DNA interaction tools are listed, together with details that are specific to each tool. These include the fourth and final defining feature for tools, namely *(iv)* type of integration within the VRE and execution environment requirements (see D6.1). Tools are categorised according to their **integration type** in the VRE: **external** when tools do not run on the computational infrastructure of the VRE, that is when they are used as external services, for example through a REST API. Conversely, **internal** when tools are run by calling an executable or by using a Python API: these tools require the definition of their interaction with the computational infrastructure of the VRE, as they may require to be installed and run in a specific execution environment. The execution environment is defined as the operating system (OS) and the minimum and maximum number of cores/processors that can be used; more complex tools may also have specific software dependencies. Each *internal* tool will carry enough information for the COMPSs runtime to handle this heterogeneity (for more details see Deliverable 6.1 [D6.1]). Although an in depth description of each of these tools functionality and method is outside the scope of this document, literature and online references are provided for further details. Finally, when available, information about the licensing policies of the tools are specified.

3.4.1 Feature prediction

- DNA Structure modelling: **NAFlex**
  - *Integration type*: EXTERNAL; web service (REST API)
  - *Execution environment*: NA
  - *Reference*: http://mmb.irbbarcelona.org/NAFlex/, [Hospital 2013]
  - *License*: Server is accessible free of charge

- Protein Structure modelling: **Modeller**
  - *Integration type*: INTERNAL; executable
  - *Execution environment*:
    - *OS*: GNU/Linux

    - *#CPUs*: min 1, max 1
  - *Reference*: https://salilab.org/modeller/, [Webb 2014]
  - *License*: Free for academics, author required authorization

- DNA Flexibility prediction: **DNAFlexBrowser (NAFlex)**
  - *Integration type*: EXTERNAL; web app
  - *Execution environment*: NA
  - *Reference*: http://www.multiscalegenomics.eu/MuGVRE/flexibility-browser/, [Hospital 2013]
  - *License*: Free

3.4.2 Structural analysis

- DNA Structure analysis: **Curves+**
  - *Integration type*: INTERNAL; executable
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max 1
  - *Reference*: http://gbio-pbil.ibcp.fr/Curves_plus, [Lavery 2009]
  - *License*: Free

- DNA binding site prediction: **DBSI**
  - *Integration type*: INTERNAL; executable
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max 1
    - *Dependencies*: NACCESS, DSSP, PSI-BLAST and SVM-light
  - *Reference*: https://mitchell-lab.biochem.wisc.edu/DBSI/, [Sukumar 2016]
  - *License*: This program and any other programs supplied with it are free to use for non-commercial purposes. Those wishing to distribute modified versions of the code can request permission to do so from the authors.

- RNA binding site prediction: **OPRA**
  - *Integration type*: INTERNAL; executable
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max 1
  - *Reference*: https://life.bsc.es/pid/opra, [Perez-Cano 2010]
  - *License*: see pyDock

- Molecular Dynamics: **NAFlex**
  - *Integration type*: INTERNAL; various executables
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max NA (application specific)
    - *Dependencies*: Amber/GROMACS
  - *Reference*: http://mmb.irbbarcelona.org/NAFlex/, [Hospital 2013]
  - *License*: Server is accessible free of charge

3.4.3 Docking and Affinity analysis

- Protein-DNA Docking: **FTDock**
  - *Integration type*: INTERNAL; executable
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max 128
  - *Reference*: http://bioinformatics.oxfordjournals.org/content/29/13/1698.long
  - *License*: GNU GPLv1

- Protein-Protein Docking: **pyDock**
  - *Integration type*: INTERNAL; executable
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max NA (1 core per pose)
  - *Reference*: https://life.bsc.es/pid/pydock/, [Cheng 2007]
  - *License*: Free for academics

- Binding Affinity Analysis (Protein-Protein): **CCharPPI**
  - *Integration type*: EXTERNAL; webapp
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max NA
  - Reference: https://life.bsc.es/pid/ccharppi, [Moal 2015]
  - License: Server is accessible free of charge for academics

- Binding Affinity Analysis (Protein-DNA): **ROSETTA**
  - Integration type: INTERNAL; executable
  - Execution environment:
    - OS: GNU/Linux
    - #CPUS: min 1, max 1
  - Reference: [Morozov 2005]
  - License: Free for academics

- Sequence Specificity Analysis : **DNAPROT**
  - *Integration type*: INTERNAL; executable
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max 1
  - *Reference*: http://161.111.227.80/compbio/soft/dnaprot.php, [Angarica 2008]
  - *License*: Free for academics

3.4.4 Motif Discovery and Binding Site prediction

- Motif Discovery: **RSAT (peak-motifs)**
  - *Integration type*: web service (SOAP API)
  - *Execution environment*: NA
  - *Reference*: http://rsat01.biologie.ens.fr/, [Medina-Rivera 2015]
  - *License*: Free for all

- Binding site prediction (genome): **TRAP**
  - *Integration type*: INTERNAL; executable
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max 1
  - Reference: http://trap.molgen.mpg.de, [Thomas-Chollier 2011]
  - License: Unspecified

- Peak calling: **MACS2**
  - *Integration type*: INTERNAL; Python library
  - *Execution environment*:
    - *OS*: GNU/Linux
    - *#CPUs*: min 1, max 1
  - Reference: http://liulab.dfci.harvard.edu/MACS/, [Zhang 2008]
  - License: Free for all

4 CONCLUSION

Protein-DNA interactions play a fundamental role in shaping the regulatory processes in the cell. Although the field is in active development, our understanding of these interactions, and in turn our ability to predict them, is still limited. The integrative approach that the MuG VRE aims to implement has the potential of overcoming some of these limitations, ultimately paving the way towards understand chromatin structure and gene expression regulation.

This document outlines the protein-DNA section of the MuG VRE, based on available tools that match the currently envisaged use cases within the field of multiscale genomics, establishing well-defined connections with the VRE's data management facilities and computational infrastructure. The aim of this document is also to sustainably delineate the structures and policies that will characterise the future addition of tools and pipelines related to protein-DNA interactions, as more workflows and more complex applications become pertinent further on in the project.

5 REFERENCES

- [D4.1] MuG Deliverable 4.1.
- [D4.2] MuG Deliverable 4.2.
- [D4.3] MuG Deliverable 4.3.
- [IC6.0] Report for the internal consultation IC6.0, available on the MuG website.
- [D6.1] MuG Deliverable 6.1.
- [Angarica 2008] Angarica, V.E., Perez, A.G., Vasconcelos, A.T., Collado-Vides, J. and Contreras-Moreira, B. (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. BMC Bioinformatics, 9:436.
- [Beveridge 2004] Beveridge, D.L., Barreiro, G., Byun, K.S., Case, D.A., Cheatham, T.E.3., Dixit, S.B., Giudice, E., Lankas, F., Lavery, R., 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.
- [Chen 2015] Chen, C., Esadze, A., Zandarashvili, L., Nguyen, D., Pettitt, B.M. and Iwahara, J. (2015) Dynamic Equilibria of Short-Range Electrostatic Interactions at Molecular Interfaces of Protein--DNA Complexes. The journal of physical chemistry letters, 6, 2733-2737.
- [Cheng 2007] Cheng, T.M.-K., Blundell, T.L., Fernandez-Recio, J. (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins, 68, 503–515.
- [Dans 2015] Dans, P.D., Faustino, I., Battistini, F., Zakrzewska, K., Lavery, R. and Orozco, M. (2015) Unraveling the sequence-dependent polymorphic behavior of d (CpG) steps in B-DNA. Nucleic Acids Res, 42, 11304-11320.
- [Dixit 2005] Dixit, S.B., Beveridge, D.L., Case, D.A., Cheatham, T.E.3., Giudice, E., Lankas, F., Lavery, R., Maddocks, J.H., Osman, R., 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.
- [Etheve 2015] Etheve, L., Martin, J. and Lavery, R. (2015) Dynamics and recognition within a protein- DNA complex: a molecular dynamics study of the SKN-1/DNA interaction. Nucleic Acids Res, 44, 1440-1448.
- [Garvie 2001] Garvie, C.W. and Wolberger, C. (2001) Recognition of specific DNA sequences. Mol Cell, 8, 937-946.
- [Hospital 2013] 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, W47–55.
- [Hospital 2015] Hospital, A., Andrio, P., Cugnasco, C., Codo, L., Becerra, Y., Dans, P.D., Battistini, F., Torres, J., Goni, R., Orozco, M. et al. (2015) BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. Nucleic Acids Res, 44 (D1), D272-D278.
- [Jen-Jacobson 2000] Jen-Jacobson, L., Engler, L.E. and Jacobson, L.A. (2000) Structural and thermodynamic strategies for site-specific DNA binding proteins. Structure Fold Des, 8, 1015-1023.
- [Lavery 2005] Lavery, R. (2005) Recognizing DNA. Q Rev Biophys, 38, 339-344.
- [Lavery 2009] 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.

- [Lavery 2010] Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T., Dixit, S., Jayaram, B., Lankas, F., 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.
- [Lee 2013] Lee, T.I. and Young, R.A. (2013) Transcriptional regulation and its misregulation in disease. Cell, 152, 1237-1251.
- [Luscombe 2000] Luscombe, N.M., Austin, S.E., Berman, H.M. and Thornton, J.M. (2000) An overview of the structures of protein-DNA complexes. Genome Biol, 1, reviews001.
- [Matthews 1988] Matthews, B. W. (1988). Protein-DNA interaction. No code for recognition. Nature, 335, 294-295.
- [Medina-Rivera 2015] Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., et al. and van Helden, J. (2015) RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Res, 43, W50-6.
- [Moal 2015] Moal, I.H., Jimenez-Garcia, B., Fernandez-Recio, J. (2015) CCharPPI web server: computational characterization of protein-protein interactions from structure. Bioinformatics, 31(1), 123-5.
- [Morozov 2005] Morozov, A.V., Havranek, J.J., Baker, D. and Siggia, E.D. (2005) Protein-DNA binding specificity predictions with structural models. Nucleic Acids Res. 33, 5781-98.
- [O'Flanagan 2005] O'Flanagan, A., Paillard, G., Lavery, R. and Sengupta, A.M. (2005) Non-additivity in protein-DNA binding. Bioinformatics, 21, 2254-2263.
- [Pasi 2014] 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) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. Nucleic Acids Res, 42, 12272- 12283.
- [Perez-Cano 2010] Perez-Cano, L., Fernandez-Recio, J. (2010) Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. Proteins, 78(1), 25-35.
- [Sarai 2005] Sarai, A. and Kono, H. (2005) Protein-DNA recognition patterns and predictions. Annu Rev Biophys Biomol Struct, 34, 379-398.
- [Slattery 2015] Slattery, M., Zhou. T., Yang. L., Dantas Machado. A.C., Gordan. R. and Rohs, R. (2015) Absence of a simple code: how transcription factors read the genome. Trends in Biochemical Sciences, 39(9), 381-399.
- [Stormo 1982] Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res, 10, 2997-3011.
- [Stormo 2013] Stormo, G.D. (2013) Modeling the specificity of protein-DNA interactions. Quantitative Biology, 1(2): 115–130.
- [Sukumar 2016] Sukumar, S., Zhu, X., Ericksen, S.S., Mitchell, J.C. (2016) DBSI server: DNA binding site identifier. Bioinformatics, 32(18), 2853-5.
- [Thomas-Chollier 2011] Thomas-Chollier, M., Hufton, A., Heinig, M., O'Keeffe, S., Masri, NE., Roider, HG., Manke, T., Vingron, M. (2011) Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. Nat Protoc, 6(12), 1860-9.
- [Webb 2014] Webb, B., Benjamin, W., Andrej, S. (2014) Comparative Protein Structure Modeling Using MODELLER. Current Protocols in Bioinformatics, 5.6.1–5.6.32

- [Zandarashvili 2013] Zandarashvili, L., Esadze, A. and Iwahara, J. (2013) NMR studies on the dynamics of hydrogen bonds and ion pairs involving lysine side chains of proteins. Adv Protein Chem Struct Biol, 93, 37-80.
- [Zhang 2008] Zhang et al. (2008) Model-based Analysis of ChIP-Seq (MACS). Genome Biol, 9 (9), R137-.