

Docking assays of Musashi-1 allosteric RNA-binding sites



Universitat
Oberta
de Catalunya



UNIVERSITAT_{DE}
BARCELONA

Lucas Goiriz Beltrán

MU Bioinf. i Bioest.
Àrea de treball final

Nombre Tutor/a de TF

Emmanuel Fajardo

**Profesor/a responsable
de la asignatura**

Emmanuel Fajardo

Fecha Entrega

12 de Enero 2023



Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada [3.0 España de Creative Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

FICHA DEL TRABAJO FINAL

Título del trabajo:	<i>Docking assays of Musashi-1 allosteric RNA-binding sites</i>
Nombre del autor:	<i>Lucas Goiriz Beltrán</i>
Nombre del consultor/a:	<i>Emmanuel Fajardo</i>
Nombre del PRA:	<i>Emmanuel Fajardo</i>
Fecha de entrega:	<i>01/2023</i>
Titulación o programa:	<i>Máster Universitario en Bioinformática y Bioestadística</i>
Área del Trabajo Final	<i>Drug Design and Structural Biology</i>
Idioma del trabajo:	<i>Inglés</i>
Palabras clave:	<i>Molecular Docking, RNA-binding protein, Allosteric Regulation</i>

Resumen del Trabajo

Máximo 250 palabras, con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo

Abstract

A maximum of 250 words, detailing the purpose, context of application, methodology, results and conclusions of the work

Index

1	Introduction	1
1.1	Context and justification	1
1.2	Objectives	3
1.3	Impact on sustainability, social-ethical and diversity	3
1.3.1	Sustainability	3
1.3.2	Social and ethical responsibility	5
1.3.3	Diversity, gender and human rights	5
1.4	Approach and methods	5
1.5	Work plan	6
1.6	Summary of obtained products	8
1.7	Summary of other sections of this report	9
2	State of the art	10
2.1	Synthetic Biology	10
2.2	RNA-binding proteins	11
2.3	Molecular Docking	11
3	Materials and methods	12
4	Results	13
5	Conclusions and further work	14
6	Glossary	15
7	References	16
8	Appendix	I

List of Figures

1	3D RRM1 structure of the MSI-1 protein constructed by means of protein crystalization and RMN.	1
2	3D mouse MSI-1 structure as predicted by AlphaFold.	2
3	Work plan summarized in a Gnatt chart.	8

1 Introduction

1.1 Context and justification

The Musashi-1 (MSI-1) protein belongs to a family of neural RNA-binding proteins (RBPs) that play a fundamental role in neural development in both vertebrates and invertebrates [1, 2, 3, 4]. In particular, the mouse MSI-1 consists of 362 aminoacids with a molecular mass of 39 kDa [2], and contains 2 sequences of around 80 to 90 aminoacids that interact with RNA (RRMs; short for RNA-recognition motifs) following the consensus sequence $RU_nAGU^{1,2}$ [4, 6].

Despite multiple efforts, the complete structure of any of the MSI-1 protein homologs remains a mystery. The most notable successes have only been able to completely elucidate MSI-1's RRM1 and RRM2 domains [7, 8, 9, 10], see for example **Figure 1** for the 3D structure of RRM1.

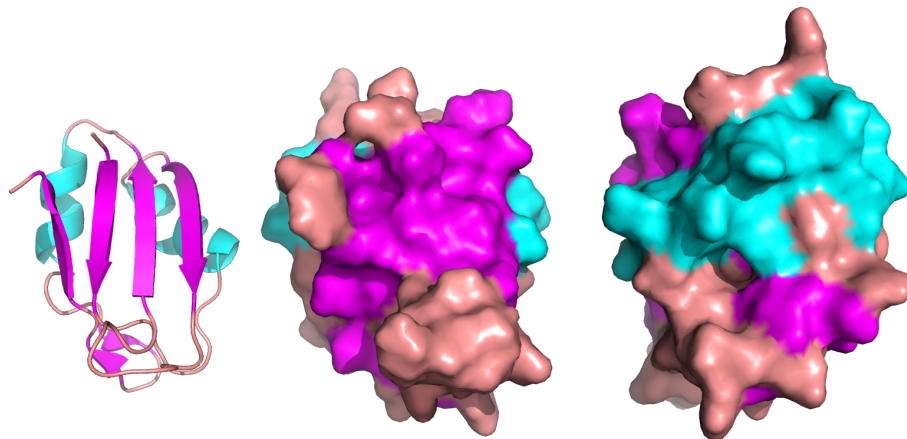


Figure 1: 3D RRM1 structure of the MSI-1 protein constructed by means of protein crystalization and RMN. From left to right: RRM1 in *cartoon* setting, RRM1 in *surface* setting and RRM1 in *surface* setting turned 180°. Colors represent the secondary structures found within the protein: cyan corresponds to α -helices, magenta corresponds to β -sheets and salmon corresponds to *random coil*. The model was retrieved from [Uniprot entry Q61474](#). Visualized through [Pymol](#).

¹Following the IUPAC standard for degenerate base symbols, R represents both an Adenine (A) and a Guanine (G). Refer to [5] for the complete IUPAC standard.

² U_n stands for repeating U within the consensus sequence n times. In most cases, $n \in \{1, 2, 3\}$ [4].

At the moment, the available complete MSI-1 structures only originate from computational predictions, such as the one shown in **Figure 2** which is an aminoacid-sequence-based prediction yielded by [AlphaFold](#) [11].

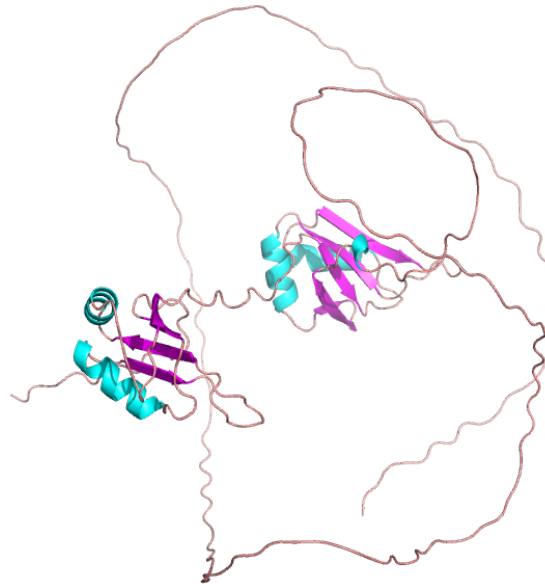


Figure 2: 3D mouse MSI-1 structure as predicted by [AlphaFold](#), in *cartoon* setting. Colors represent the secondary structures found within the protein: cyan corresponds to α -helices, magenta corresponds to β -sheets and salmon corresponds to *random coil*. The model was retrieved from [Uniprot entry Q61474](#). Visualized through [Pymol](#).

With the surge of synthetic biology during the past two decades, RBPs such as MSI-1 have gained a special interest for their applicability as post-transcriptional regulators of gene expression in synthetic gene circuits [12, 13, 14]. In particular, Dolcemascolo and colleagues pioneered by designing a synthetic genetic circuit employing MSI-1 as post-transcriptional regulator [15]. In addition, they also exploited MSI-1's particularity of binding to fatty acids in addition to RNA, out of which the *cis*-9-Octadecenoic acid (also known as oleic acid) presents a higher binding affinity. The latter interaction weakens the MSI-1 protein's RNA-binding affinity, resulting in a dissociation between MSI-1 and the RNA sequence [15, 16]. In fact, the pocket in which the fatty acids interact with the RMM1 domain of MSI1 can be easily seen in the rightmost protein in **Figure 1**.

Therefore, in order to confidently add MSI-1 as an allosteric regulator to the synthetic biologist's toolbox, there is a need to understand how the underlying interaction mechanisms work. Clingman and colleagues were the first

who performed docking simulations with the MSI-1 core RNA-binding-motif and several fatty acids [16]. Nevertheless, these results only hold for the aforementioned RNA-motif: it is expected that any mutation on the RNA-motif would drastically impact the manner in which MSI-1 interacts with the RNA.

1.2 Objectives

This Master's Thesis aims to extend the knowledge behind the allosteric interactions occurring in the MSI-1 proteins by performing docking simulations between MSI-1's RRM1 domain, several RNA motifs and fatty acids, with the objective of obtaining a model for each of the combinatorial cases.

This general objective breaks down into the following specific objectives:

1. *In silico* Generate the 3D structures of 5 mutant MSI-1 RNA-binding motifs used in [15]. The 3D structures of the motifs will depend on the sequence-based-prediction of their secondary structures.
2. Perform docking simulations between MSI-1 and each of the mutants.
3. Perform docking simulations between MSI-1 and several fatty acids of interest: oleic acid, linoleic acid, palmitoleic acid, arachidonic acid and stearic acid.
4. Keep the top 2 fatty acids whose interaction with MSI-1 is the clearest.
5. Repeat the docking simulations of the mutant RNA-motifs in presence of each of the selected fatty acids.

1.3 Impact on sustainability, social-ethical and diversity

1.3.1 Sustainability

The result of this Master's Thesis has had a negligible but negative impact on sustainability. This impact originates on the energy consumption during the following steps:

1. Design and debugging of the docking pipeline followed in this work.
2. Computation of the different models.

This impact can be considered negligible because the net amount of energy consumed during the computations is very small. Nevertheless, it is important to highlight this aspect as there is room for improvement: the source code of the docking software employed in this work is mainly composed of two programming languages: C (72%) and Python (28%).

Pereira and colleagues developed a benchmark to order programming languages with respect of their energy consumption when executing well-known computer science algorithms and data structures [17]. Their finding was that the least energy consuming programming language was C, while Python was one of the most energy consuming.

It can be discussed that, in order to reach minimal energy consumption, the programming language to be used in software that is going to be distributed and executed multiple times should be exclusively C. The disadvantages of C are that it takes more time to write a full program than with higher level languages (such as Python), it is more difficult to maintain a codebase written in C and that C is not a particularly accessible programming language for non-computer scientists.

There is a trade-off that has to be considered. Let L be a programming language, E_{W_L} be the energy consumed by a computer during the crafting of a software written in L , E_{E_L} be the energy consumed by a computer during one execution of the software written in L (in this particular case, one docking simulation with some fixed parameters) and n be the number of times the software is executed, then it is possible to approximate $E_{T_L}(n)$ (the total energy consumed by the software written in L at the moment n) as follows:

$$E_{T_L}(n) \approx E_{W_L} + n \times E_{E_L} \quad (1)$$

For the pair Python and C, we have that $E_{W_{\text{Python}}} \lll E_{W_C}$ and $E_{E_{\text{Python}}} \ggg E_{E_C}$. So the larger n becomes (the number of times the software is executed), by **Equation 1**, Python becomes less and less energy efficient when compared to C.

Nevertheless, as mentioned previously, the codebase of the software used is in its majority written in C, the energy consumption of a personal computer is low and the number of executions n in this work is small (less than 10^2). Therefore this negative impact on energy consumption can be considered negligible.

1.3.2 Social and ethical responsibility

The technical nature of this work separates it from any positive nor negative impacts on socio-ethical aspects. There are no conflicts of interests with the results, and this work's material (models and source codes) are freely available to anyone through a public `GitHub` repository.

In addition, this work doesn't have any socio-ethical motivations, as the main objective is to gather more technical knowledge for the field of molecular biology.

1.3.3 Diversity, gender and human rights

Again, the technical nature of this work separates it from any positive nor negative impacts on diversity, gender and human rights. The references and materials employed within this work were used without any gender nor race bias.

Furhermore, this work doesn't have any diversity, gender nor human rights motivations. However, this work attempts to be accessible to as many people as possible, in the spirit of the universal right for knowledge and education. Moreover, this work exclusively employed license-free and open sourced software so that all the dependencies are accessible to anyone.

1.4 Approach and methods

This work aims to extend the work of Dolcemascolo and colleagues [15] where they designed a synthetic genetic circuit that employed MSI-1 as post-transcriptional regulator. Thanks to the plasticity of RNA, several mutants of MSI-1's RNA-binding core motif were designed, and it was shown that gene regulation (which is tightly dependent on the MSI-1-RNA interaction) was altered for all mutants: some of them displayed an increase in regulation fold change while others displayed a decrease. In addition, the effects of MSI-1 induced regulation were partially nullified in presence of oleic acid.

The initial step of this work will consist in the retrieval of several 3D structures from the appropriate databases:

- the 3D structure of mouse MSI-1's RMM1 is retrieved from [UniProt](#).
- the 3D structures of several fatty acids (oleic acid among them) are retrieved from [chemspider](#).

Next, some of the RNA sequences employed in [15] are retrieved and their 3D model structures are computed by means of [NUPACK](#) and [RNAComposer](#) web server.

Then, all the docking simulations are performed:

1. Between each RNA sequence and MSI-1 RMM1.
2. Between each fatty acid and MSI-1 RMM1.
3. Between the models obtained in the simulations of step 1 and the fatty acids.

The docking software will be [LightDock](#) because it is license-free and open source. The best models will be selected based on the chosen scoring function. In addition, all the models will be visualized with [Pymol](#).

1.5 Work plan

This work is divided in the following tasks (and respective subtasks):

1. Perform docking simulations of MSI-1 and a collection of mutants of the MSI-1 RNA-binding-motif.
 - (a) Generate in silico a collection of up to 5 interesting mutant MSI-1 RNA-binding motifs.
 - (b) Perform docking simulations between MSI-1 and each of the mutants.
 - (c) Discard those mutants whose interaction with MSI-1 is poor.
2. Perform docking simulations of MSI-1 and a collection of fatty acids.
 - (a) Perform docking simulations between MSI-1 and several fatty acids of interest: oleic acid, linoleic acid, palmitoleic acid, arachidonic acid and stearic acid.
 - (b) Keep the top 2 fatty acids whose resulting model is the best.
3. Perform docking simulations of MSI-1 and a subset of mutants of the MSI-1 RNA-binding-motifs in presence of fatty acids.

4. Defense preparation

- (a) Manuscript
- (b) Presentation

The landmarks of this work are distributed among the following continuous assessment tests:

- Continuous Assessment Test 1: Thesis Definition and Work Plan.
- Continuous Assessment Test 2: Work Development (Phase 1).
Landmarks:
 - Generate in silico a collection of up to 5 interesting mutant MSI-1 RNA-binding motifs.
 - Perform docking simulations between MSI-1 and each of the mutants *(partially)*.
- Continuous Assessment Test 3: Work Development (Phase 2).
 - Perform docking simulations between MSI-1 and each of the mutants.
 - Discard those mutants whose interaction with MSI-1 is poor.
 - Perform docking simulations between MSI-1 and several fatty acids of interest: oleic acid, linoleic acid, palmitoleic acid, arachidonic acid and stearic acid *(partially)*.
- Continuous Assessment Test 4: Thesis Closure and Presentation.
 - Perform docking simulations between MSI-1 and several fatty acids of interest: oleic acid, linoleic acid, palmitoleic acid, arachidonic acid and stearic acid.
 - Keep the top 2 fatty acids whose resulting model is the best.
 - Manuscript closure.
 - Presentation closure.

The work plan is summarized in **Figure 3**:

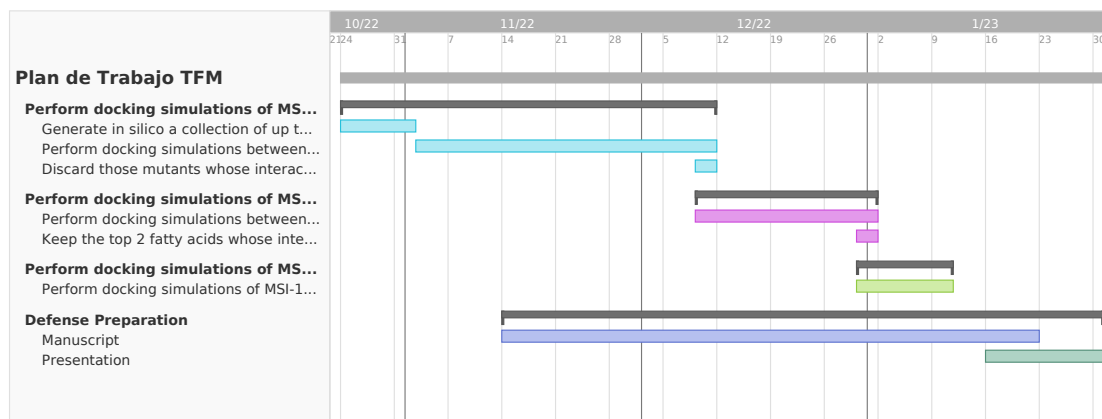


Figure 3: Work plan summarized in a Gantt chart.

1.6 Summary of obtained products

The main products of this work are:

1. The current Master's Thesis manuscript.
2. 3D structures of 5 MSI-1 RNA-motifs used in [15].
3. *los modelos o lo que salga*
4. *los scripts del pipeline*
5. [A step-by-step tutorial on how to perform a simple protein-RNA docking with LightDock.](#)

Resulting products are available in different repositories:

1. The direct products of this work are present in [this GitHub repository](#).
2. The step-by-step tutorial on how to perform a simple protein-RNA docking with LightDock is available at [the LightDock GitHub repository](#), and soon on the [LightDock tutorials page](#).

1.7 Summary of other sections of this report

The remaining sections of this manuscript are:

- **State of the art**
- **Materials and methods**
- **Results**
- **Conclusions and further work**
- **Glossary**
- **References**
- **Appendix**

2 State of the art

2.1 Synthetic Biology

Synthetic Biology is a modern niche of Biology in which the characteristic “know-how” from engineering fields is combined with classical Molecular Biology techniques. Guided by physicist Richard Feynman’s famous quote “*what I cannot create, I do not understand*”, synthetic biologists have developed a deep understanding of how cells control and regulate the processes of gene transcription, translation, metabolite biosynthesis and degradation.

It is “synthetic” in the sense that synthetic biologists engineer novel (absent in nature) genetic constructions with biological “pieces”, which usually originate from various different organisms, in order to achieve a phenotype in a target organism that lacks said phenotype in nature. Furthermore, this design usually comes with a mathematical model based in physical first principles of how the system should behave.

One of the first successful constructions of this type, and most likely the most popular one, is Gardner’s toggle switch [18] where they engineered a synthetic toggle switch within *E. coli* by using two different promoters that repressed each other, achieving this way bistability (either one promoter is repressed, or the other). Simultaneously, Elowitz’ repressilator [19] was published, where they engineered a synthetic “clock” within *E. coli* by chaining 3 repressive promoters that repressed one another in chain, in such a way that the genes controlled by them had periodic expression.

These constructions, and all the ones that followed, would not have been possible without a (by then, small) repository of genetic parts, which in turn would not have been possible without the arduous work of an innumerable amount of traditional molecular biologists that sequenced and characterized said parts. In fact, the iGEM foundation keeps a repository of standardized biological parts [20] along with various data such as the assembly methods used, levels of expression, polymerase used, etc...

However, there is still a need for more biological parts, especially parts that are “orthogonal”, i.e. they can co-exist and function correctly within the same organism without deviating from their theoretical behavior.

2.2 RNA-binding proteins

2.3 Molecular Docking

3 Materials and methods

4 Results

5 Conclusions and further work

6 Glossary

MSI-1

RBP

RRM

NMR

7 References

- [1] M. Nakamura, H. Okano, J. A. Blendy, and C. Montell, "Musashi, a neural rna-binding protein required for drosophila adult external sensory organ development," *Neuron*, vol. 13, no. 1, pp. 67–81, 1994.
- [2] S. Sakakibara, T. Imai, K. Hamaguchi, M. Okabe, J. Aruga, K. Nakajima, D. Yasutomi, T. Nagata, Y. Kurihara, S. Uesugi, T. Miyata, M. Ogawa, K. Mikoshiba, and H. Okano, "Mouse-musashi-1, a neural rna-binding protein highly enriched in the mammalian cns stem cell," *Developmental Biology*, vol. 176, no. 2, pp. 230–242, 1996.
- [3] P. Good, A. Yoda, S.-i. Sakakibara, A. Yamamoto, T. Imai, H. Sawa, T. Ikeuchi, S. Tsuji, H. Satoh, and H. Okano, "The humanmusashi homolog 1(msi1) gene encoding the homologue of musashi/nrp-1, a neural rna-binding protein putatively expressed in cns stem cells and neural progenitor cells," *Genomics*, vol. 52, no. 3, pp. 382–384, 1998.
- [4] T. Imai, A. Tokunaga, T. Yoshida, M. Hashimoto, K. Mikoshiba, G. Weinmaster, M. Nakafuku, and H. Okano, "The neural rna-binding protein musashi1 translationally regulates mammalian numb gene expression by interacting with its mrna," *Molecular and Cellular Biology*, vol. 21, no. 12, pp. 3888–3900, 2001.
- [5] A. Cornish-Bowden, "Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984," *Nucleic Acids Research*, vol. 13, no. 9, pp. 3021–3030, 1985.
- [6] N. R. Zearfoss, L. M. Deveau, C. C. Clingman, E. Schmidt, E. S. Johnson, F. Massi, and S. P. Ryder, "A conserved three-nucleotide core motif defines musashi rna binding specificity," *Journal of Biological Chemistry*, vol. 289, no. 51, pp. 35530–35541, 2014.
- [7] T. Nagata, R. Kanno, Y. Kurihara, S. Uesugi, T. Imai, S.-i. Sakakibara, H. Okano, and M. Katahira, "Structure, backbone dynamics and interactions with rna of the c-terminal rna-binding domain of a mouse neural rna-binding protein, musashi1," *Journal of Molecular Biology*, vol. 287, no. 2, pp. 315–330, 1999.
- [8] Y. Miyanoiri, H. Kobayashi, T. Imai, M. Watanabe, T. Nagata, S. Uesugi, H. Okano, and M. Katahira, "Origin of higher affinity to rna of the n-terminal rna-binding domain than that of the c-terminal one of a mouse neural protein, musashi1, as revealed by comparison of their structures, modes of

- interaction, surface electrostatic potentials, and backbone dynamics,” *Journal of Biological Chemistry*, vol. 278, no. 42, pp. 41309–41315, 2003.
- [9] T. Ohyama, T. Nagata, K. Tsuda, N. Kobayashi, T. Imai, H. Okano, T. Yamazaki, and M. Katahira, “Structure of musashi1 in a complex with target rna: The role of aromatic stacking interactions,” *Nucleic Acids Research*, vol. 40, no. 7, pp. 3218–3231, 2011.
- [10] L. Lan, M. Xing, M. Kashipathy, J. Douglas, P. Gao, K. Battaile, R. Hanzlik, S. Lovell, and L. Xu, “Crystal and solution structures of human oncoprotein musashi-2 n-terminal rna recognition motif 1,” *Proteins: Structure, Function, and Bioinformatics*, vol. 88, no. 4, pp. 573–583, 2019.
- [11] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, and et al., “Highly accurate protein structure prediction with alphafold,” *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
- [12] B. J. Belmont and J. C. Niles, “Engineering a direct and inducible protein-rna interaction to regulate rna biology,” *ACS Chemical Biology*, vol. 5, no. 9, pp. 851–861, 2010.
- [13] J. Cao, M. Arha, C. Sudrik, A. Mukherjee, X. Wu, and R. S. Kane, “A universal strategy for regulating mrna translation in prokaryotic and eukaryotic cells,” *Nucleic Acids Research*, vol. 43, no. 8, pp. 4353–4362, 2015.
- [14] N. Katz, R. Cohen, O. Solomon, B. Kaufmann, O. Atar, Z. Yakhini, S. Goldberg, and R. Amit, “Synthetic 5’ utrs can either up- or downregulate expression upon rna-binding protein binding,” *Cell Systems*, vol. 9, no. 1, 2019.
- [15] R. Dolcemascolo, M. Heras-Hernández, L. Goiriz, R. Montagud-Martínez, A. Requena-Menéndez, R. Márquez-Costa, R. Ruiz, A. Pérez-Ràfols, R. A. Higuera-Rodríguez, G. Pérez-Ropero, W. F. Vranken, T. Martelli, W. Kaiser, J. Buijs, and G. Rodrigo, “Repurposing the mammalian rna-binding protein musashi-1 as an allosteric translation repressor in bacteria,” *bioRxiv*, 2022.
- [16] C. C. Clingman, L. M. Deveau, S. A. Hay, R. M. Genga, S. M. Shandilya, F. Massi, and S. P. Ryder, “Allosteric inhibition of a stem cell rna-binding protein by an intermediary metabolite,” *eLife*, vol. 3, p. e02848, 2014.
- [17] R. Pereira, M. Couto, F. Ribeiro, R. Rua, J. Cunha, J. P. Fernandes, and J. Saraiva, “Energy efficiency across programming languages: How do energy, time, and memory relate?,” *Proceedings of the 10th ACM SIGPLAN International Conference on Software Language Engineering*, 2017.

- [18] T. S. Gardner, C. R. Cantor, and J. J. Collins, “Construction of a genetic toggle switch in escherichia coli,” *Nature*, vol. 403, no. 6767, pp. 339–342, 2000.
- [19] M. B. Elowitz and S. Leibler, “A synthetic oscillatory network of transcriptional regulators,” *Nature*, vol. 403, no. 6767, pp. 335–338, 2000.
- [20] The iGEM Foundation, “igem registry of standard biological parts.” http://parts.igem.org/Main_Page, 2022.

8 Appendix