Single-molecule fluorescence in sequence space

Severins, I.W.H.


1

Chapter 1

Introduction:
The importance of sequence

"You can never know everything, and part of what you know is always wrong. Perhaps even the most important part. A portion of wisdom lies in knowing that. A portion of courage lies in going on anyways."
- al'Lan Mandragoran (The Wheel of Time)

1.1 INTRODUCTION

Biomolecules are the result of millions of years of evolution, carefully tuned to form a network of interacting elements, working together to form the living systems that populate the planet today. The ability to evolve such networks is largely due to the ability to store and process information, which became possible by the development of sequence-defined molecules: DNA, RNA and proteins. Compared to the evolutionary process that formed these molecules we are now extremely rapidly understanding how these molecules function. Since the discoveries of DNA, RNA, and proteins, we have learned how they are synthesized, and have identified thousands of genes encoding proteins to perform various tasks inside the cell. The sequence of these molecules dictates for a large part their structure, and in turn their function, for example by enabling or disabling interactions of various strengths.

To study the sequence-dependence of biomolecular structure and function with high throughput, we have mainly relied on bulk approaches for which broadly two categories can be distinguished: solution-based and surface-based methods. Examples of the former are SELEX (Komarova and Kuznetsov 2019) and RNA Bind-n-Seq (Lambert et al. 2014) that are used to study protein binding to nucleic acids by separating the bound and unbound sequences. The bound nucleic acids are then purified and sequenced to identify the protein-binding motifs. These solution-based methods can be used to study huge libraries, but provide mostly qualitative information on high-affinity interactions. On the other hand, surface-based assays study the effects of sequence using clusters of DNA molecules immobilized to a surface. These clusters can be located on microarrays (Bulyk 2007) or on Illumina sequencing chips where they are formed during the sequencing process (Nutiu et al. 2011). Observing cluster binding of fluorescently-labelled proteins allows studying their sequence preferences (Bulyk 2007; Nutiu et al. 2011). While surface-based approaches allow obtaining more quantitative information, they are mostly limited to simple two-state systems and require synchronization to determine reaction rates.

Conversely, single-molecule fluorescence methods give a detailed view of the inter- and intramolecular kinetics, shedding light on complex, multi-state reactions. They do not require synchronization and can show heterogeneities within populations. However, only a limited number of sequences can be studied due to constraints in time and cost. For example, varying 6 nucleotides would result in a sequence library of 4^6 and thus 4096 sequences. Performing single-molecule experiments in series with a speed of 20 experiments per day would result in a total experiment duration of 9 months. Furthermore, obtaining all DNA sequences separately, costing about 25 euros per oligo, would result in a total expense of roughly 100,000 euros. These numbers underscore the necessity of a parallel approach.

In this thesis, therefore, we introduce a new method to perform high-throughput Single-molecule Parallel Analysis for Rapid eXploration of Sequence space, in short SPARXS, where we perform single-molecule fluorescence experiments and next generation Illumina sequencing on the same sequencing flow cell. In Chapter 2 we give an overview of the previously established approaches that perform surface-based biophysical and biochemical assays on sequencing clusters, varying either DNA, RNA or protein sequence. In addition, we discuss the various possibilities to perform such assays at the single-molecule level. Chapter 3 describes the detailed protocol to perform SPARXS, discussing, among others, differences with

regular single-molecule fluorescence experiments and experimental design considerations to take into account. In Chapter 4 we zoom in on the specific step in this protocol that registers the single-molecule and sequencing datasets to determine corresponding molecules and sequences. We discuss the process of registration and of estimating a suitable distance threshold for determining correspondence. Chapter 5 shows the application of SPARXS to the sequence-dependent kinetics of the four-way DNA Holliday junction structure, highlighting that we can obtain mechanistic insights from the observed kinetic behaviour. Finally, in Chapter 6 we look ahead to the future of SPARXS, discussing potential applications, improvements and expansions.

1.2 REFERENCES

Bulyk, Martha L. 2007. "Protein Binding Microarrays for the Characterization of DNA–Protein Interactions." In *Analytics of Protein–DNA Interactions*, 104:65–85. Springer Berlin Heidelberg. https://doi.org/10.1007/10_025.

Komarova, Natalia, and Alexander Kuznetsov. 2019. "Inside the Black Box: What Makes SELEX Better?" *Molecules* 24 (19): 3598. https://doi.org/10.3390/molecules24193598.

Lambert, Nicole, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A. Sharp, and Christopher B. Burge. 2014. "RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins." *Molecular Cell* 54 (5): 887–900. https://doi.org/10.1016/j.molcel.2014.04.016.

Nutiu, Razvan, Robin C Friedman, Shujun Luo, Irina Khrebtukova, David Silva, Robin Li, Lu Zhang, Gary P Schroth, and Christopher B Burge. 2011. "Direct Measurement of DNA Affinity Landscapes on a High-Throughput Sequencing Instrument." *Nature Biotechnology* 29 (7): 659–64. https://doi.org/10.1038/nbt.1882.