Single-molecule fluorescence in sequence space

Severins, I.W.H.


S

Summary / Samenvatting

Summary

Engineering biological systems, for example designing drugs for curing diseases or enzymes for improving chemical synthesis, requires understanding of the physical processes that govern biology. These processes have been studied for many decades on a wide range of scales from whole ecosystems, down to organisms, organs, tissues and cells. However, behavior at all of these scales ultimately originates from the properties of, and interactions between, individual biomolecules. Many of these biomolecules – DNA, RNA and proteins – consist of a specific sequence of building blocks that determines their molecular structure and function. Therefore, to understand their behavior, we have to know the effects of sequence.

The relation between sequence and function has been traditionally studied using ensemble assays, where many molecules are studied together, as an ensemble. Initially, only solution-based assays were performed using a library of sequences in solution. Sequences with specific properties are selected from the library and identified, for example DNA sequences that bind to a specific protein. Such experiments, however, provide mostly qualitative information. With the introduction of next-generation sequencing it became possible to perform surface-based experiments that provide quantitative information. The sequencer uses a library of sequences to create spatially separated clusters of DNA molecules on the surface of a flow cell. Sequencing of the clusters reveals their identity, which allows them to be employed for biochemical and biophysical assays. These cluster-based assays have been applied to study protein-DNA interactions, such as those of the CRISPR-Cas protein complex. The DNA of each cluster can also be transcribed to RNA and subsequently translated to proteins. This allowed studying the influence of RNA sequence on RNA or protein binding, and even uncovered the effect of protein sequence on the enzymatic conversion of a small-molecule substrate. Such quantitative data about the influence of sequence yielded new mechanistic understanding and allowed the construction and verification of thermodynamic models.

While high-throughput studies on sequencing clusters reveal interesting sequence-dependent properties and behavior, they provide limited kinetic information. Studying the ensemble of molecules in a cluster obscures heterogeneities within a population and variations in time, and only allows the observation of simple two-state kinetics. Overcoming these limitations requires studying each molecule separately, which can be achieved using single-molecule fluorescence studies that excel in observing multi-state dynamic processes. However, these studies provide only low sequence throughput, as obtaining and handling each sequence individually is laborious and costly. To perform these experiments on a wide range of sequences a parallel approach is thus essential. Hence, we introduce SPARXS: Single-molecule Parallel Analysis for Rapid eXploration of Sequence space. This platform enables simultaneous profiling of thousands of different sequences at the single-molecule level by coupling single-molecule fluorescence microscopy with next-generation high-throughput sequencing.

SPARXS is initiated with the immobilization of a fluorescently-labelled DNA sequence library onto a commercially-available Illumina sequencing flow cell. Then, a single-molecule fluorescence experiment is performed by imaging the flow cell on a custom fluorescence microscope, performing automated scanning to cover a large area. Next, the flow cell is transferred to the Illumina sequencer, where individual molecules undergo surface-based amplification to form clusters, which are then sequenced. Subsequently, the coordinate systems of the single-molecule and sequencing datasets are aligned, allowing coupling of kinetics to sequences. Finally, this results in a kinetic or energy landscape in sequence space.

While the concept of combining single-molecule experiments and sequencing appears simple, its implementation provided several challenges. For example, sequencing imposes strict requirements on the homogeneity and composition of the DNA library. Additionally, single-molecule experiments on the sequencing flow cell required the removal of single-molecule-like autofluorescence through photobleaching. Furthermore, performing sequencing after single-molecule experiments required altering several of the default sequencing steps. Aligning the single-molecule and sequencing datasets was complicated due to the large datasets and the large number of missing points. As a solution, we developed a three-step approach starting with global alignment between the instruments using a geometric hashing algorithm, followed by an experiment-specific alignment using a cross-correlation algorithm and, finally, a step to fine tune each field of view using a kernel correlation algorithm. Additionally, we developed a way to estimate an appropriate distance threshold to link the kinetics and sequencing data.

To show the utility and effectiveness of SPARXS, we applied it to study the sequence dependence of the four-way DNA Holliday junction that forms during homologous recombination. The Holliday junction structure is known to switch between two stacked states, where the rate of switching depends on the sequence at the junction core. By fluorescently labelling two of the Holliday junction arms, the switching kinetics could be observed using Förster resonance energy transfer (FRET). In this way the dynamic behavior of millions of Holliday junction molecules covering thousands of distinct core sequences could be measured, a result unattainable with previous techniques. We obtained a variety of quantitative parameters: the fraction of molecules showing dynamic behavior, the FRET efficiencies for the observed states, and the equilibrium constant and transition rates for multi-state molecules. The obtained transition rates could be correlated to the purine-pyrimidine distribution and the GC content at the junction core, and in turn to the theoretical stacking energies. However, no such correlation was observed for the equilibrium constant, indicating the involvement of other structural influences, likely dependent on the specific base pair identities.

While we show the application of SPARXS to study the structure of a DNA system, other applications may be envisioned as well. By using transcription and translation, the sequence-dependence of RNA and proteins may be studied at the single-molecule level. Furthermore, in addition to intra-molecular structures, inter-molecular binding and chemical reactions may be studied by introducing DNA, RNA, proteins or small molecules in solution. Overall, the rich datasets, the high throughput, and the broad applicability highlight the utility of the new method. SPARXS thus opens an entirely new and unexplored avenue of study, providing novel insights into the working mechanisms of biological systems, possibly contributing to next steps in biological engineering and design.