A STUDY OF BIOMOLECULAR INTERACTIONS IN KAPPA-CASEIN AND SLEEPING BEAUTY TRANSPOSASE

by

Venkatesh V Ranjan

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Chemistry and Nanoscale Science

Charlotte

2024

Approved by:

_____
Dr. Irina V. Nesmelova

_____
Dr. Kirill A. Afonin

_____
Dr. Donald J. Jacobs

_____
Dr. Jerry M. Troutman

ABSTRACT

VENKATESH V. RANJAN. A Study of Biomolecular Interactions in Kappa-Casein and Sleeping Beauty Transposase (Under the direction of DR. IRINA V. NESMELOVA)

This thesis investigates the model intrinsically disordered protein (IDP) κ-casein and the multidomain protein Sleeping Beauty (SB) transposase. Using advanced techniques such as pulsed-field gradient NMR and time-resolved FRET, we reveal that κ-casein exhibits continuous self-association, significantly impacting its translational diffusion. At low volume fractions, κ-casein self-associates, leading to macroscopic phase separation, while at higher concentrations, it forms labile gel-like networks. For SB transposase, we employ microscale thermophoresis to determine its DNA binding affinity to transposon direct repeats, providing crucial insights into transpososome assembly. Furthermore, we experimentally tested the predicted model of the transpososome complex in solution using FRET-based distance restraints. Our analysis identified discrepancies between the computational model of the paired-end complex and the experimentally derived inter-residue distances. These findings have broad implications for both basic science and applied biotechnology, offering potential advancements in gene therapy, and genetic engineering, and highlighting the complex interplay between protein structure, function, and environment in elucidating IDP interactions. This research enhances our understanding of IDP behavior in crowded environments and contributes to optimizing transposon-based gene delivery systems.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Irina Nesmelova, for her guidance and support throughout my doctoral journey. Her expertise, patience, insightful feedback, and constructive criticisms have been invaluable in shaping both this thesis and my growth as a researcher. I am also profoundly grateful to my committee members, Dr. Donald Jacobs, Dr. Jerry Troutman, and Dr. Kirill Afonin, for their readiness to always help me with my research questions, and for their assistance in setting realistic expectations that helped refine my research projects. Special thanks go to Dr. Yuri Nesmelov for guiding me through a significant part of the experiments and generously sharing his expertise. His incisive questions about and enthusiasm for the subject matter were truly inspiring. I am indebted to my colleagues in the Nesmelova Lab and in the Department of Chemistry for creating a stimulating and friendly research environment.

I would like to extend my heartfelt thanks to Dr. Michael Walter, Dr. Tom Schmedake, and Dr. Bernadette Donovan-Merkert at the Department of Chemistry. The faculty and staff at the Department of Chemistry have consistently demonstrated goodwill in every interaction, for which I am deeply appreciative.

I am eternally grateful to my family for their unconditional love, understanding, and sacrifices throughout this journey. I also thank my grandparents, whose memories and blessings are always with me. Finally, special thanks to my fellow Survivors of Fall'18, now known as Dr. Tyler Adams and Dr. Abhishek Shibu, for their camaraderie, intellectual discussions, and moral support. To them and to all my friends who have made this journey more bearable and enriching, I extend my sincerest thanks. You know who you are. You guys made everything worthwhile.

DEDICATION

To those on the other side of the world:

Mumma, Papa, Amiya

I couldn’t have done this without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AF3	AlphaFold 3
AIRs	Ambiguous interaction restraints
BMRB	Biological Magnetic Resonance Data Bank
BSA	Bovine serum albumin
CAPRIN1	Cell cycle associated protein 1
CARA	Computer Aided Resonance Assignment
CD	Circular dichroism
CNS	Crystallography & NMR Systems
Ddx4	DEAD-box helicase 4
DHMRI	David H. Murdock Research Institute
DMSO	Dimethyl sulfoxide
DR	Direct repeat
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electrophoretic mobility shift analysis
FRET	Förster resonance energy transfer
GuHCl	Guanidine hydrochloride
HADDOCK	High Ambiguity Driven protein-protein DOCKing
HeLa	Henrietta Lacks (cell line)

HSQC	Heteronuclear single quantum coherence
IDP	Intrinsically disordered protein
IgG	Immunoglobulin G
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IR-DRs	Inverted repeat direct repeats
IRF	Instrument response function
KD	Dissociation constant
LB	Lysogeny broth
Li	Left inner (transposase binding site)
LLPS	Liquid-liquid phase separation
Lo	Left outer (transposase binding site)
MAKE-NA	Make Nucleic Acid program
METRIC	Molecular Education, Technology, and Research Innovation Center
MODSTE	Modified five-pulse stimulated echo sequence
MST	Microscale thermophoresis
NACCESS	Solvent accessibility calculation program
NHS	N-hydroxysuccinimide
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
NOE	Nuclear Overhauser Effect

NS	Non-specific (DNA sequence)
OD600	Optical density at 600 nm
PAI	PAI subdomain of SB transposase
PB	PiggyBac
PD-DOCK	Protein-DNA docking program
PDB	Protein Data Bank
PEC	Paired-end complex
PFG	Pulsed-field gradient
PMSF	Phenylmethylsulfonyl fluoride
PROCHECK	Program to check the stereochemical quality of protein structures
PSDO	Pseudo atoms
PYMOL	PyMOL Molecular Graphics System
RED	RED subdomain of SB transposase
RMSD	Root-mean-square deviation
SB	Sleeping Beauty
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SETMAR	SET domain and mariner transposase fusion protein
STE	Stimulated-echo pulse sequence
TALOS	Torsion angle likelihood obtained from shift and sequence similarity
TBST	Tris-buffered saline with Tween 20
TCC	Transposase/transposon end/target DNA complex

TCEP	Tris(2-carboxyethyl)phosphine
TIRs	Terminal inverted repeats
TMR	Tetramethylrhodamine
TOCSY	Total correlation spectroscopy
Tol2	Transposable element derived from medaka fish
UV-Vis	Ultraviolet-visible spectroscopy
XPLOR-NIH	X-PLOR software package from NIH

CHAPTER 1: INTRODUCTION

This thesis investigates biomolecular interactions and their effect on the translational diffusion of an intrinsically disordered protein (IDP) using the model IDP κ-casein (part one) and the formation of the nucleoprotein complex by the multidomain protein Sleeping Beauty (SB) transposase (part two). In recent years it became increasingly recognized that the occurrence of unstructured regions of significant size (more than 50 residues) is also common in functional proteins[5, 6]. These disordered regions are characterized by great structural flexibility and plasticity invoking the analogy to flexible synthetic polymers. However, due to the heterogeneous composition of charged, polar, and nonpolar amino acids, proteins are never random coils and always have some residual structure[7, 8]. The degree of compactness of the polypeptide chain depends on the amino acid residue composition of a given protein and on environmental conditions, including the concentration of the protein itself and/or the crowders. Hence, there is a great interest to understand how the intrinsically disordered proteins (IDPs) behave in the wide range of concentrations, from dilute to highly concentrated solutions. In particular, understanding the translational diffusion of IDPs, which is the major mode of macromolecular transport in biological or chemical systems (e.g., the self-diffusion, hereafter denoted simply as diffusion), becomes important. However, thus far only a few diffusion coefficient measurements have been performed for IDPs or proteins unfolded by different denaturants[9-14]. Although the differences in the diffusion coefficients between folded and unfolded proteins have been reported, it is still not clear whether hydrodynamically the IDP can be pictured as similar to globular proteins, flexible synthetic polymers, or as species with unique features.

κ-Casein is a well-studied Intrinsically Disordered Protein (IDP) in the context of its role in milk and dairy products[15, 16]. Kappa casein is considered to be a model IDP[17]. Lacking a stable three-dimensional structure, the amino acid sequences of IDPs are characterized by low hydrophobicity and high net charge[18]. IDPs have a predominance of disorder-causing residues like Alanine (A), Arginine (R), Glycine (G), Glutamic acid (E), Proline (P), Serine (S), Lysine (K), and Glutamine (Q)[19]. The lack of a stable structure in IDPs poses significant challenges for their study. As a result, the behavior of κ-casein in crowded cellular environments, where it may encounter high concentrations of other macromolecules, is less understood[17, 20]. It is known for its ability to self-associate and form micelles, which are crucial for stabilizing milk proteins[15, 17, 21]. The self-association of κ-casein is a critical factor in its function. In a crowded environment, such as within a cell, κ-casein molecules may interact with each other and with other proteins, affecting their diffusion and function[17, 22]. The self-association process can lead to the formation of larger complexes and even liquid-liquid phase separation (LLPS)[23], impacting the protein's mobility and its ability to reach target sites within the cell[24]. Translational diffusion is the primary mechanism by which proteins move through cellular environments[25]. For κ-casein, understanding how its diffusion is affected by self-association is essential for elucidating its biological roles.

The Sleeping Beauty (SB) transposase is a multidomain protein of significant biological and clinical importance[26-30], including applications in humans[31]. This enzyme is involved in the process of DNA transposition, where DNA segments move from one location to another within a genome[32]. The SB transposase adapts its structure for optimal interaction with DNA, facilitating the formation of the transpososome, a nucleoprotein complex essential for transposition[33-36]. The domains of SB transposase are connected by flexible linkers that enable the transposase to

accommodate the four different DNA-binding sites in the SB transposon for the reaction of transposition[36-42]. However, studying the SB transposase poses significant challenges. The crystallographic structure of the full-length transposase could not be obtained due to the inability to crystallize the protein, and the only experimental structural information is available for individual domains of the SB transposase without DNA[4, 43, 44]. In this thesis, I use a combination of complementary alternative approaches to investigate the formation of a nucleoprotein complex (the transpososome) by the SB transposase.

Overall, the findings reported in this thesis have broad implications for both basic science and applied biotechnology.

CHAPTER 2: TRANSLATIONAL DIFFUSION AND SELF-ASSOCIATION OF AN INTRINSICALLY DISORDERED PROTEIN K-CASEIN USING NMR WITH ULTRA-HIGH PULSED-FIELD GRADIENT AND TIME-RESOLVED FRET[1]

2.1 INTRODUCTION

The conformational plasticity of a protein in response to environmental conditions is determined by the physicochemical properties of amino acids and their arrangement in the protein sequence, which dictates which intramolecular interactions or side chain interactions with the solvent are more favorable[45]. Water is a poor solvent for the protein backbone, and in a natural environment, many proteins adopt a compact three-dimensional structure stabilized by a variety of interactions, such as ionic and hydrophobic interactions as well as hydrogen and disulfide bonds[46, 47]. However, protein sequences with low hydrophobicity and high net charge preferentially adopt disordered, extended conformations[48-50]. These intrinsically disordered proteins (IDPs) play a functional role in signaling pathways and regulatory processes[51-53]. Understanding how IDPs move in crowded environments is particularly important because IDPs are commonly found in

[1] (With minor modifications this work was published in *The Journal of Physical Chemistry B* (2024): Melnikova, Daria L., Venkatesh V. Ranjan, Yuri E. Nesmelov, Vladimir D. Skirda, and Irina V. Nesmelova. "Translational Diffusion and Self-Association of an Intrinsically Disordered Protein κ-Casein Using NMR with Ultra-High Pulsed-Field Gradient and Time-Resolved FRET.")

cellular compartments and regions with high local concentrations of proteins, DNA, and RNA[54-56].

In the absence of a rigorous theoretical framework, an effective approach to understanding the translational diffusion of IDPs is to compare their translational diffusion coefficients at various concentrations with two well-studied limiting cases: flexible synthetic polymers and globular proteins. In accordance with polymers and globular proteins having fundamentally different structures, the behavior of their diffusion coefficients mirrors these differences, which are particularly evident as solutions transition from dilute to crowded[57]. The master curve for the concentration dependence of translational diffusion coefficients of synthetic polymers, constructed based on de Gennes' dynamic scaling theory[58, 59], is valid for solvent quality ranging from θ (where polymer-polymer interactions equal polymer-solvent and solvent-solvent interactions) to good (where polymer-solvent interactions prevail over polymer-polymer interactions). This master concentration dependence curve shows a gradual increase from dilute solutions, where interactions between molecules are negligible and the polymer molecule moves as an impenetrable to solvent molecules coil, to concentrated solutions, where polymer molecules entangle, and their motion is much more complex. In contrast, the master concentration dependence curve for globular proteins[60] demonstrates a more sharp transition from dilute to concentrated solutions that follows the theoretical concentration dependence of the diffusion coefficient of rigid Brownian spheres[61]. Furthermore, the diffusion regime in concentrated solutions of globular proteins qualitatively differs from that of polymers because the maximum solubility of the globular proteins is approximately close to the concentration of close packing for hard spheres, where entanglement is not expected.

Since IDPs do not form well-defined structures, they explore a large number of conformations. An empirical expression $R_H \sim N^v$, relating the hydrodynamic radius R_H of an IDP to the number of amino acid residues N comprising it, was proposed based on the analysis of experimental data[62-64]. Reported Flory exponent v values for IDPs range from 0.49289 to 0.50987 and correspond to the value of v for homopolymers in θ (indifferent) solvents[65]. Note that in dilute solutions, even random coils move as hydrodynamically compact species. We have observed such behavior in solutions of the IDP α-casein[66], where the diffusion coefficient of α-casein follows the same trend as the concentration dependence of the diffusion coefficients of globular proteins and rigid Brownian spheres. However, increasing IDP concentration can significantly increase their conformational heterogeneity. Additionally, IDPs have much shorter chains and heterogeneous charge distributions along the amino acid sequence, resulting in varying degrees of compaction[62, 67, 68]. Therefore, it is unclear whether the diffusion behavior of an IDP in crowded space corresponds to that in a concentrated polymer solution, in which long and fully flexible polymer chains entangle, forming transient networks.

One of the normalization parameters, used in constructing the master curves for synthetic polymers or globular proteins[60, 69], is critical concentration $\hat{\varphi}$. We have shown that in the case of globular proteins $\hat{\varphi}$ reflects the tendency of molecules to self-associate[60]. For example, $\hat{\varphi}$ is equal to 0.16 (expressed as a volume fraction) in solutions of myoglobin, where no association is observed, whereas it is equal to 0.08 in the solution of lysozyme with pH 7.4–7.8, where lysozyme molecules form aggregates[60, 70-72]. The shape of the master curve is also sensitive to the key features of the aggregation process. In concentrated solutions, the diffusion coefficient of an IDP α-casein demonstrates a much stronger concentration dependence, φ^{-12}, than that of globular proteins or linear flexible polymers[66]. This strong dependence results from the continuous self-assembly of

α-casein molecules into labile supramolecular gel networks, restricting the translational mobility of the molecule as a whole due to the formation of multiple protein-protein interactions that lead to gel formation and are not accounted for in the construction of the master curve[66]. In this regard, comparing experimental data with master curves allows identifying the presence of interactions leading to the formation of supramolecular structures in the studied solutions. Although not shown before for proteins, based on the data for several polymer systems[73, 74], we also expect that the concentration dependence of an IDP may be sensitive to liquid-liquid phase separation (LLPS), which is a ubiquitous phenomenon in IDP proteins[75]. However, due to limited information on the translational diffusion of IDPs in concentrated solutions, both under self-crowding conditions and in the presence of crowding molecules of different nature, more experimental data are needed to assess the impact of various types of self-association on the translational diffusion of IDPs and to determine the applicability of scaling laws and master curve analysis in such conditions.

The goal of this work was to investigate the translational diffusion and supramolecular assembly of κ-casein, an IDP from the casein family[76-78]. Caseins comprise approximately 80% of milk protein, and their primary function is to serve as a source of amino acids, calcium, and phosphorus[79]. There are four types of caseins in mammals: α_{s1}, α_{s2}, β, and κ-casein[80]. Due to the amphipathic nature of their molecules, containing both polar and hydrophobic domains, all caseins display a strong tendency to self-associate[81, 82]. The degree and type of association vary between different caseins due to differences in amino acid composition and their distribution in the sequence[83, 84] (Appendix A and Figure S1). κ-casein is unique among caseins as it has only one or two phosphorylated residues, and thus forms fewer interactions with calcium compared to other caseins and remains soluble in the presence of calcium[85]. Independently, κ-casein exists

as a dynamic, oligomeric ensemble, the properties of which are highly dependent on concentration as well as solution pH and buffer composition[86-89]. In the mixture, κ-casein interacts with highly phosphorylated α and β-caseins and prevents their aggregation and precipitation in the presence of high concentrations of calcium[83, 90], leading to the formation of a thermodynamically stable complex with calcium phosphate known as casein micelle[80, 83]. In the casein micelle, κ-casein is believed to play a stabilizing role by forming a polyelectrolyte "brush," a highly hydrated layer on the surface that provides electrostatic stability to the micelle in good solvents and determines the micelle size by preventing further aggregation of caseins[91]. Currently, the model describing how a casein micelle forms is still under debate[85]. Therefore, understanding the association and nature of intermolecular interactions of casein molecules, both with themselves and with each other, remains important.

Pulsed-field gradient Nuclear Magnetic Resonance (PFG NMR) diffusion measurements are particularly suitable for studying protein association[57, 92], as the translational diffusion coefficient is inversely proportional to the size of diffusing species and, thus, highly sensitive to size changes. However, the informativeness of results on molecular association depends on the ability to measure very slowly moving molecules. Previously, using ultra-high PFG NMR, enabling us to measure the diffusion coefficients as low as 10^{-15} m^2/s, we detected and characterized three-dimensional gel-like structures in α-casein solutions[66, 93]. In this work, we utilized ultra-high PFG NMR to investigate κ-casein solutions and found that the molecules of κ-casein also form geometrically similar gel-like networks in concentrated solutions but exhibits fundamentally different behavior from α-casein at concentrations below the gel formation threshold. Specifically, we observed self-association of κ-casein molecules even at very low protein concentrations and macroscopic phase separation when solutions were stored at 4°C.

2.2 MATERIALS AND METHODS

2.2.1 Materials

Bovine κ-casein (C0406) was purchased from Sigma-Aldrich and used without further purification. We verified the purity and homogeneity of κ-casein by gel electrophoresis (SDS-PAGE). Under reducing conditions, κ-casein migrated as a single band at 19 kDa (Appendix A Figure S2), which corresponds to its molecular weight. We also confirmed that κ-casein is unstructured by circular dichroism spectroscopy (Appendix A Figure S3). All measurements were carried out using fresh samples within several hours after preparation, except otherwise mentioned.

2.2.2 NMR Diffusion Measurements

For NMR diffusion measurements, the lyophilized powder of κ-casein was dissolved in D_2O to minimize the signal from water protons in NMR spectra. Protein concentrations ranged from 0.1 to 20% (w/v %, hereafter) or 0.001 to 0.147 volume fractions. The volume fraction, φ, of κ-casein was calculated using the following relation:

$$\varphi = \frac{1}{1 + \frac{\rho_2 \cdot \omega_1}{\rho_1 \cdot (1 - \omega_1)}}, \qquad (1)$$

where ρ_1 and ρ_2 are the densities of water and κ-casein, respectively, and ω_1 is the weight fraction of water. The density of κ-casein was calculated using its partial specific volume value of 0.689 cm^3/g, determined previously[94, 95].

All NMR measurements were performed at 298 K on a 400 MHz Bruker Avance-III TM spectrometer equipped with a gradient system that allowed an ultra-high gradient, *g*, with the maximum value of 28 T/m (2800 G/cm). Self-diffusion coefficients (hereinafter referred to as diffusion coefficients) were measured using the stimulated-echo pulse sequence (STE)[96] and the

modified five-pulse stimulated echo sequence (MODSTE)[2] (Appendix A Figure S4). The integrated area of the protein peak between 0.16 and 3.61 ppm was used to characterize the κ-casein signal. The experiments were carried out using 48 different values of g and gradient pulse durations δ of 1, 2, or 5 ms. The time interval between the first and second radiofrequency pulses was kept constant in all experiments at τ_1 = 10 or 16 ms to exclude the influence of spin-spin relaxation time $T2$ on the shape of diffusion attenuation. The diffusion time t_d varied from 50 ms to 800 ms by changing τ_2 in the MODSTE pulse sequence (Appendix A Figure S4). The standard experimental error of measured diffusion coefficients was below 5%.

Multi-exponential diffusion attenuations were described by the spectrum of diffusion coefficients, D_i, according to the equation:

$$\frac{A(g^2)}{A(0)} = \sum_i p_i' \exp\left(-\gamma^2 g^2 \delta^2 t_d D_i\right), \qquad (2)$$

where $A(0)$ is the spin-echo amplitude at $g = 0$, γ is the gyromagnetic ratio for protons, and p_i' is the relaxation-weighted fraction of the component with the diffusion coefficient D_i given by the equation:

$$p_i' = \frac{p_i \exp\left(-\frac{2\tau_1}{T2_i} - \frac{\tau_2}{T1_i}\right)}{\sum_{i=1}^{N} p_i \exp\left(-\frac{2\tau_1}{T2_i} - \frac{\tau_2}{T1_i}\right)}, \qquad (3)$$

where p_i is the fraction of the component with the diffusion coefficient D_i, $T2_i$ and $T1_i$ are the proton spin-spin and spin-lattice relaxation times of κ-casein molecules within an i-th aggregate, and τ_2 is the interval between the second and third 90° radiofrequency (RF) pulses in a stimulated-echo pulse sequence.

To describe the spectrum of observed diffusion coefficients, we used the average diffusion coefficient $\langle D \rangle$ defined in accordance with the equation:

$$\langle D \rangle = \sum_i p_i' D_i \,. \qquad (4)$$

<*D*> is determined with high accuracy from the initial slope of the diffusion attenuation ($g \to 0$). In the presence of molecular exchange between species with different diffusion coefficients, to evaluate the fraction of molecules and their lifetime within a given species, it is necessary to maintain the contribution of relaxation factors (Eq. 3) constant for different diffusion times. This can be achieved using the MODSTE pulse sequence presented in Appendix A Figure S4. In this pulse sequence, by maintaining the $\tau_2+\tau_4$ sum constant, the diffusion attenuation can be recorded at different diffusion times while keeping the contribution of *T*1 the same. We demonstrated that the MODSTE pulse sequence enables the unambiguous characterization of the exchange process between species with different diffusion coefficients D_i, even if each of the exchanging species is also characterized by distributions of both T1 and T2 relaxation times, by applying this approach to evaluate the lifetime of a "guest-host" complex formed by the antitumor agent 5-FU and carrier β–CD [2].

To obtain the diffusion coefficient distribution from non-exponential diffusion attenuations, we used the estimate of lifetime and a home-written software based on the Tikhonov regularization algorithm[97, 98]. The main advantage of this software is the minimal number of fitting parameters. Initially, from the analysis of the diffusion attenuation, a physically justified range of expected values for diffusion coefficients is set. The primary regularization parameter, which determines the accuracy of the fit of the diffusion attenuation with the calculated distribution of diffusion coefficients, is the number of iterations N_i. The effect of N_i on the spectrum for both non-exponential (0.1% aqueous solution of k-casein) and exponential (water) diffusion attenuations is demonstrated in Appendix A Figure S5. In our fits, N_i was set to 100.

2.2.3 Time-Resolved FRET Measurements

For FRET measurements, the lyophilized powder of κ-casein was dissolved in H_2O and incubated with the donor (EDANS-C2-maleimide, Anaspec) or acceptor (DABCYL-C2-maleimide, Anaspec) at room temperature for two hours. The labeling was done at a 1:1 protein:label ratio, and unreacted labels were removed using Amicon® ultra centrifugal filters. We assume that only one cysteine per protein was labeled. For titration experiments, the concentration of the donor-labeled κ-casein was kept constant at 0.01% (5 μM), and the concentration of the acceptor-labeled κ-casein varied from 0.01 to 0.38% (5 to 200 μM).

Time-resolved FRET was measured using a home-built transient fluorimeter equipped with a QuadraCentric sample compartment with a cuvette holder and a Peltier element for temperature control (Horiba Scientific), a passively Q-switched microchip YAG laser (SNV-20F-100, 355 nm, 20 kHz, Teem Photonics), a photomultiplier (H6779-20, Hamamatsu), and a fast digitizer (Acqiris DC252, Agilent). A 420 nm cutoff filter and a polarizer set at the magic angle were used in the detection arm. All experiments were done at the temperature of 293 K. The labeled protein solution was loaded into the observation cuvette, and the time-resolved donor fluorescence waveform was acquired by averaging fluorescence transients from one thousand laser pulses. We used a non-fluorescent acceptor to analyze the donor fluorescence only. The obtained waveforms of donor fluorescence were best fitted by three exponential components, convoluted with the instrument response function measured separately from the light scatter before each experiment. The component with the shortest fluorescence decay time (τ_D = 0.7 ns) remained constant during the titration at all concentrations of acceptor-labeled protein. The other two components showed an identical decrease of τ_D. Because the component with the longest decay time comprised approximately 70% of the measured waveform, it was used to extract the values of the donor

fluorescence decay time for further analysis. FRET efficiency was determined from each independent experiment using the equation:

$$E = 1 - \frac{\tau_{DA}}{\tau_D}, \quad (5)$$

where τ_D is the fluorescence decay time of the donor-labeled protein alone, and τ_{DA} is the fluorescence decay time of the donor-labeled protein in the presence of bound acceptor-labeled protein. The Förster distance, R_0, in our experiments was calculated to be 2.6 nm (Appendix A).

To describe κ-casein self-association, we used the model of indefinite isodesmic association where the addition of each successive monomer to an associate involves an equal change in free energy. In this model, the following set of equilibria is considered:

$$M_{i-1} + M \xleftrightarrow{K} M_i, \quad i = 1,2,... \quad (6)$$

where M denotes a monomer, and M_i denotes an i-mer. The total molar concentration of protein, C, is the sum of the molar concentrations of all i-mers in solution, which can be written using the molar concentration of monomers under the assumption of isodesmic association:

$$C = \sum_{i=1}^{\infty} i c_i = \sum_{i=1}^{\infty} i K^{i-1} c_1^i . \quad (7)$$

For indefinite association, the summation of series gives a simple expression for the total concentration:

$$C = \frac{c_1}{(1 - K c_1)^2} . \quad (8)$$

From the quadratic equation, the concentration of monomers as a function of total protein concentration is then found:

$$c_1 = \frac{2KC + 1 - \sqrt{4KC + 1}}{2K^2 C} . \quad (9)$$

Using Eq. (8), molar fractions for each species in solution can be found according to the following relations:

$$\alpha_1 = \frac{c_1}{C}, \quad \alpha_2 = \frac{2Kc_1^2}{C}, \quad ..., \quad \alpha_i = \frac{iK^{i-1}c_1^i}{C}, ... \tag{10}$$

The FRET efficiency of the *i*-th κ-casein species in the presence of multiple acceptors is given by the equation[99]:

$$E_j = \alpha_j \frac{jR_0^6}{jR_0^6 + r_0^6}, \tag{11}$$

where R_0 is 2.6 nm. Index *j* indicates the number of acceptors near a donor, and $j = 1$ corresponds to one acceptor near a donor. Accordingly, the cumulative FRET efficiency detected experimentally is written as a sum of all individual contributions:

$$E = \sum_{j=1}^{\infty} \alpha_j \frac{jR_0^6}{jR_0^6 + r_0^6} \tag{12}$$

2.3 RESULTS

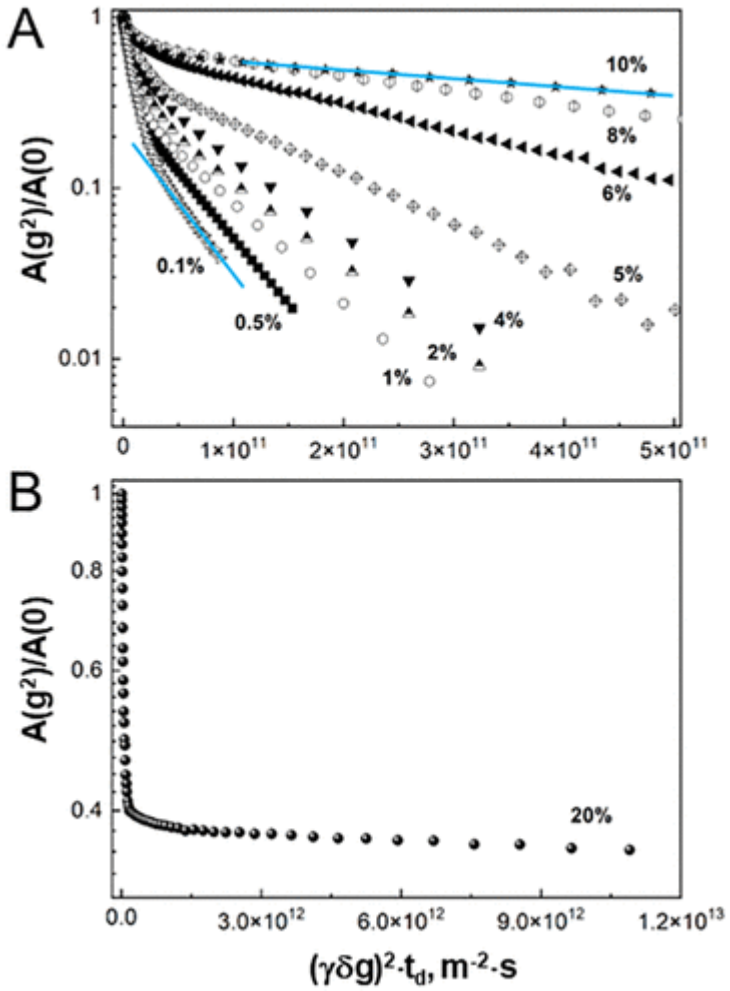

Figure 2.1 Diffusion attenuations of spin–echo signal in solutions of κ-casein. Diffusion attenuations, recorded at td = 50 ms, are shown for protein concentrations in the range from 0.1 to 10% (A) and 20% (B). Solutions were prepared in 100% D2O at pH 7.0. The measurements were done at 298 K. The deviation from a monoexponential attenuation is observed for all protein concentrations. D_{min} decreases as the concentration of κ-casein indicated by solid blue lines drawn to 0.1 and 10% curves.

Figure 2.1 presents the semi-logarithmic plots of spin-echo intensity recorded as a function of the pulsed field gradient amplitude, g, at κ-casein concentrations ranging from 0.1 to 20%. The chosen semi-logarithmic coordinates clearly reveal the deviation of the signal attenuation curves, $A(g^2)$, from mono-exponential behavior at all concentrations studied. Since the κ-casein samples are monodisperse in molecular weight (Appendix A Figure S2), the non-exponential nature of the

diffusion attenuation reflects the high propensity of κ-casein to self-associate. As expected for self-associating molecules, as the protein concentration in solution increases, the change of the shape of diffusion attenuations reflects the emergence of increasingly smaller diffusion coefficients. This is clearly observed by the change in the slope of the corresponding part of the diffusion attenuation, as exemplified by blue lines for the 0.1 and 10% solutions.

To verify that the observed self-association in our experiments is not due to the formation of disulfide bonds, previously suggested as one of the mechanisms of casein micelle formation[100], we performed the reversibility test as described before[66]. Since the formation of disulfide bonds is irreversible and is expected to be more pronounced at higher protein concentrations, we compared the diffusion attenuations of 5% κ-casein samples prepared freshly and by dissolving a 20% sample. Both diffusion attenuations were identical, ruling out the formation of associates due to intermolecular disulfide links (Appendix A Figure S6).

To compare the concentration dependence of the translational diffusion coefficient of κ-casein to flexible polymers and globular proteins, we used the average diffusion coefficient $<D>$ (Eq. 4), determined from the initial slope of the diffusion attenuation. The value of $<D>$ is several orders of magnitude larger than D_{min} and, therefore, primarily reflects the contribution of fast moving κ-casein molecules. Furthermore, without using ultra-high pulsed-field gradients, all available information on self-diffusion of κ-casein would rely only on average diffusion coefficients. Figure 2.2 shows the dependence of $<D>$ presented in logarithmic coordinates on protein concentration,

recalculated as volume fraction using Eq. 1 to facilitate the comparison with master curves for globular proteins and flexible polymers.

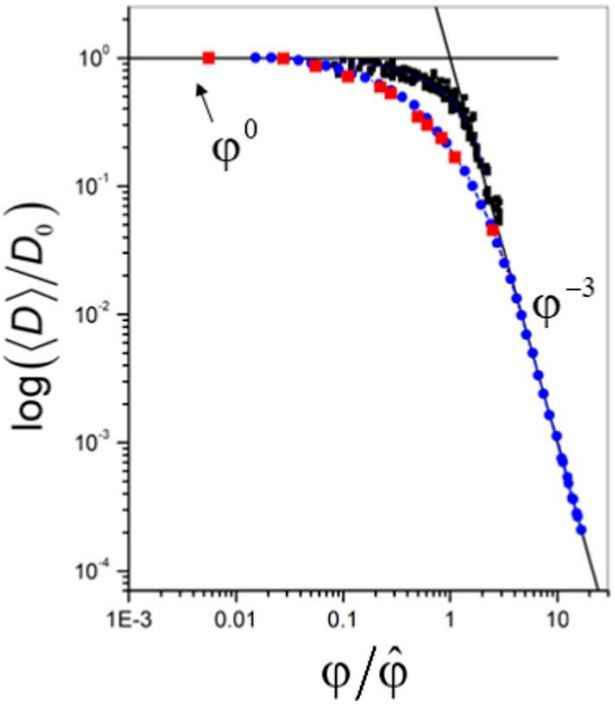

Figure 2.2 The concentration dependence of the κ-casein diffusion coefficient. The normalized concentration dependence of the normalized κ-casein diffusion coefficient ⟨D⟩ is shown by red squares. For comparison, the master curves are shown for the concentration dependence of the diffusion coefficient of linear flexible polymers (blue circles) and globular proteins (black squares). Solid lines indicate the asymptotes with the slopes of φ^0 and φ^{-3}.

We also performed the same normalization procedure established for flexible synthetic polymers and later applied to globular proteins[59]. First, to eliminate temperature dependence, we divided the diffusion coefficient of κ-casein at each concentration by the value of the diffusion coefficient at infinite dilution, determined by extrapolating the experimental data to zero κ-casein concentration. Then, the volume fraction was normalized by the critical concentration, determined from the intersection of the asymptotes with zero slope (dilute solutions) and slope φ^{-3} (concentrated solutions), as indicated by the solid lines. The value of $\hat{\varphi}$ was equal to 0.08. In agreement with observed self-association of κ-casein, this value was smaller than 0.16, determined previously for solutions of non-associating globular proteins[60]. The φ^{-3} asymptote was chosen

because this asymptotic behavior was theoretically predicted for the diffusion coefficient of synthetic polymers and empirically determined for globular proteins. The normalization to critical concentration is equivalent to shifting the entire curve along the log(φ) axis without changing its shape. Similarly to globular proteins and α-casein[60, 66], normalization of the diffusion coefficient by the contribution of internal dynamics, as done for flexible polymers, was not necessary.

Remarkably, the concentration dependence of the κ-casein diffusion coefficient shows the same gradual increase as the master curve for flexible synthetic polymers over the entire range of concentrations studied.

Our data, indicating that the κ-casein self-associate even at the lowest used concentration of 0.1%, agree with gel-filtration data[101] and with reports that reduced and carboxymethylated bovine *κ*-casein self-associates into micelle-like structures at concentrations above 0.05%[86, 102]. In contrast, no association of α-casein was observed at protein concentrations up to 2%, and the diffusion attenuations for these α-casein solutions were mono-exponential[66].

The value of D_{min} in a 0.1% κ-casein solution is approximately $1.5 \pm 0.06 \times 10^{-11}$ m^2/s, which is about an order of magnitude lower than the expected diffusion coefficient for κ-casein monomers, based on comparisons with proteins of similar size[57, 60]. Using the Stokes-Einstein formula, $D=kT/6\pi\eta R$, where k is the Boltzmann constant, T is the temperature, and η is the viscosity of the pure solvent (D_2O, 1.1×10^{-3} Pa·s[103]), we roughly estimate that the hydrodynamic radius R of species diffusing with the diffusion coefficient D_{min} is approximately 13 nm. This is more than three times greater than the expected value of κ-casein R_H, which is 3.56 nm, estimated using the empirical expression[64] $R_H = 2.84N^{0.493}$ with $N = 169$. We also obtained similar linear dimensions for a κ-casein molecule using a secondary structure prediction method PSIPRED[3] (Appendix A

Figure S7). Consequently, considering a simplified geometrical model of random close packing of spheres to account for void volume[104], the number of κ-casein molecules in the associated species can reach up to about 31. This is in agreement with the polymerization value of 30 determined for κ-casein from viscosity and sedimentation data[102] and within the range of sizes reported for κ-casein associates measured by different experimental techniques[80, 87, 105].

The diffusion attenuation of residual H_2O in a 0.1% κ-casein solution is mono-exponential and does not show t_d dependence. This could indicate that either the water content in the κ-casein associate is small (below the detection sensitivity) and/or the exchange with bulk water is fast on the time scale of our experiments (<< 50 ms). Additionally, this suggests that κ-casein associates are protein-dense structures without a large volume of confined water, which would be expected to demonstrate the features of restricted diffusion.

To explore the lability of associated κ-casein species, we studied the dependence of diffusion attenuation $A(g^2)$ on diffusion time t_d for a 0.5% κ-casein solution and assessed the fraction and lifetime of κ-casein molecules within the associates. To exclude the influence of spin-lattice relaxation time $T1$ on the shape of diffusion attenuation (Eq. 3), we used the modified five-pulse stimulated echo pulse sequence MODSTE[2]. Figure 2.3A shows diffusion attenuations recorded at different t_d values. The values of the average and the lowest detected diffusion coefficients $<D>$ and D_{min} do not depend on the diffusion time. In contrast, the fraction of the slowest diffusing molecules, p_{min}, decreases with increasing diffusion time, indicating the presence of molecular exchange between different κ-casein species. We estimated the average lifetime τ* of κ-casein molecules in the associated state using the following approach. The probability for a κ-casein molecule to leave the associate species diffusing with D_{min} at least once is given by the integral

$\int_0^{t_d} \Psi_i(\tau_i) d\tau_i$, where τ is the lifetime distribution of κ-casein molecules for an *i*-th snapshot of the system. In this case, the dependence of $p_{\min}$ on t_d can be calculated according to the equation:

$$p_{\min}(t_d) = p_{\min}(0)\left(1 - \int_0^{t_d} \Psi_i(\tau_i) d\tau_i\right) \quad (13)$$

where $p_{\min}(0)$ is the fraction of the κ-casein molecules in the associates at t_d approaching zero. Figure 2.3B shows that the dependence $p_{\min}(t_d)$ can be well described by an exponential function. By fitting the $p_{\min}(t_d)$ data with an exponent, we determined the fraction and lifetime of κ-casein molecules within the associate to be equal to $p_{\min}(0) = 0.230 \pm 0.002$ and $\tau^* = 0.53 \pm 0.06$ s, respectively. Note that the value of $p_{\min}(0)$ includes the relaxation contribution as described by Eq. 3, and thus provides the lower limit estimate.

To corroborate the NMR diffusion data and determine the dissociation constant for κ-casein species forming in dilute solutions, we carried out FRET titration experiments. During the titration, the concentration of the donor-labeled κ-casein (EDANS-labeled) was kept constant at 0.01% (5 μM), whereas the concentration of the acceptor-labeled κ-casein (DABCYL-labeled) was incrementally increased from 0.01 to 0.38% (5 to 200 μM). Figure 4 shows that FRET efficiency

increases as a function of κ-casein concentration, reflecting the formation of donor-acceptor complexes.

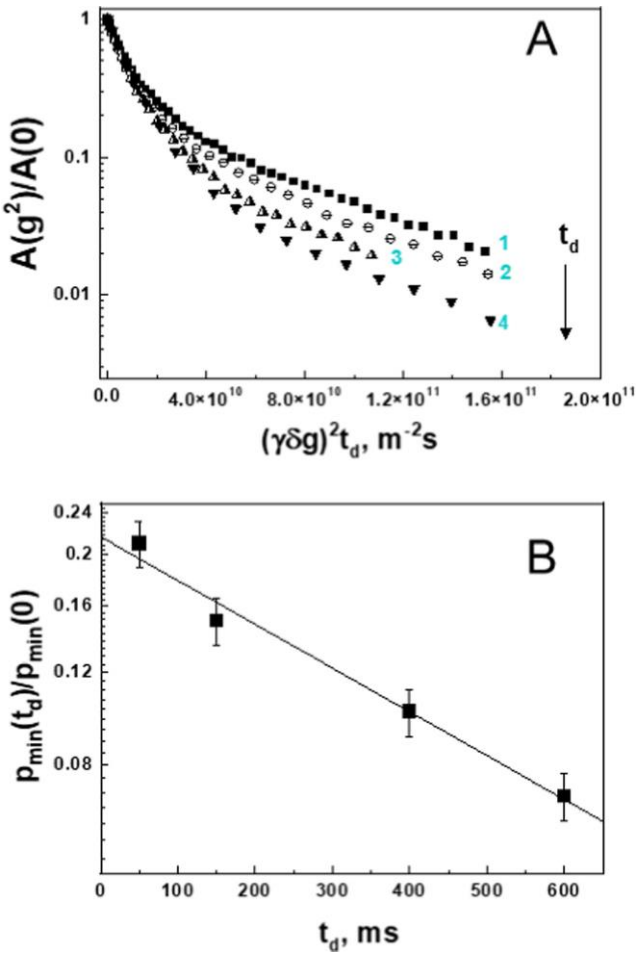

Figure 2.3 The dependence of the diffusion attenuation on diffusion time in 0.5% κ-casein solution. (A) Curves 1–4 represent diffusion attenuations collected at t_d = 50, 150, 400, and 600 ms, respectively. (B) The dependence of the fraction of slowly diffusing κ-casein species, p_{min}, on t_d. The solid line shows the best fit of experimental data to eq 13.

Using FRET efficiency data, we estimated the dissociation constant for adding a monomer to an associate of κ-casein under the assumptions of indefinite isodesmic association and the contribution of acceptors in the first layer around the donor only. The latter assumption is based on the fact that the linear dimensions of the κ-casein molecule are comparable to R_0 = 2.6 nm (see Appendix A), and the FRET efficiency for the donor-acceptor pair at a distance of ~$2R_0$ apart

becomes less than 2%. By fitting Eq. 12 to the experimental data (solid red line in Figure 2.4), the value of the κ-casein equilibrium dissociation constant, K_D, (the inverse of K, Eq. 7) is estimated to be 9.5 ± 1.5 μM. We note that the self-association of κ-casein is highly sensitive to buffer conditions. We provide the estimate for the aqueous solution of κ-casein, whereas the addition of 5 mM sodium phosphate buffer leads to about a 3-fold decrease in the equilibrium association constant (e.g., stronger affinity of binding, data not shown). Given the difference in experimental conditions, our K_D value reasonably agrees with surface plasmon resonance data for casein-casein interactions[101].

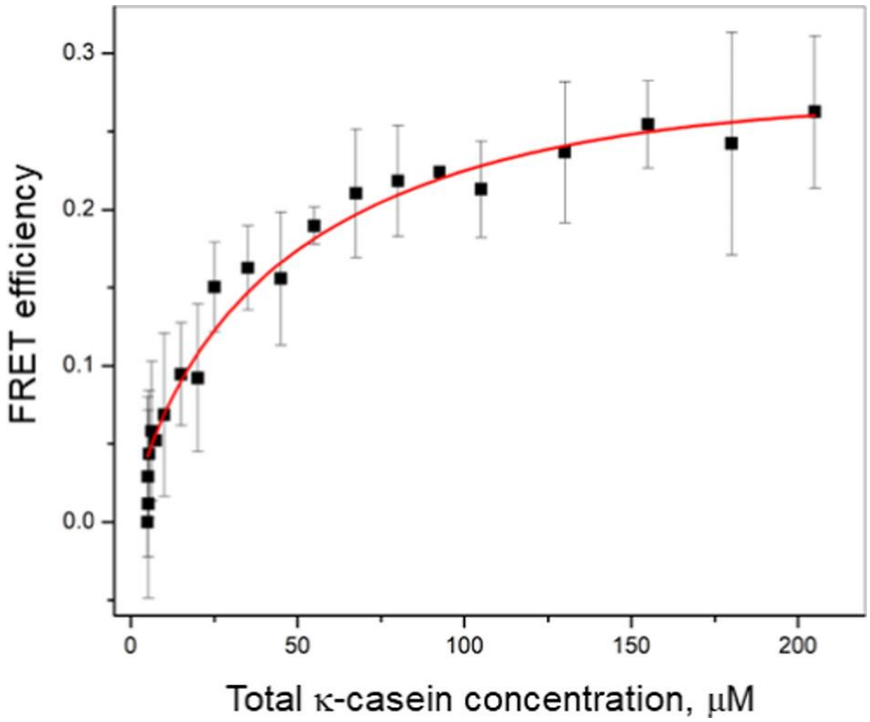

Figure 2.4 The dependence of FRET efficiency on κ-casein concentration. Experimental error is shown as standard deviation from at least three independent measurements. The solid line represents the best fit of the isodesmic association model to experimental data.

Using the estimate of a lifetime and a home-written software based on the Tikhonov regularization algorithm[97], we calculated the spectra of diffusion coefficients from the non-exponential diffusion attenuations, recorded for κ-casein solutions in the range of concentrations from 0.1% to

10%. (Figure 2.5). These spectra clearly show a bimodal distribution of molecules into slowly and fast-diffusing, with the fraction of slowly diffusing molecules increasing with the increasing concentration of κ-casein.

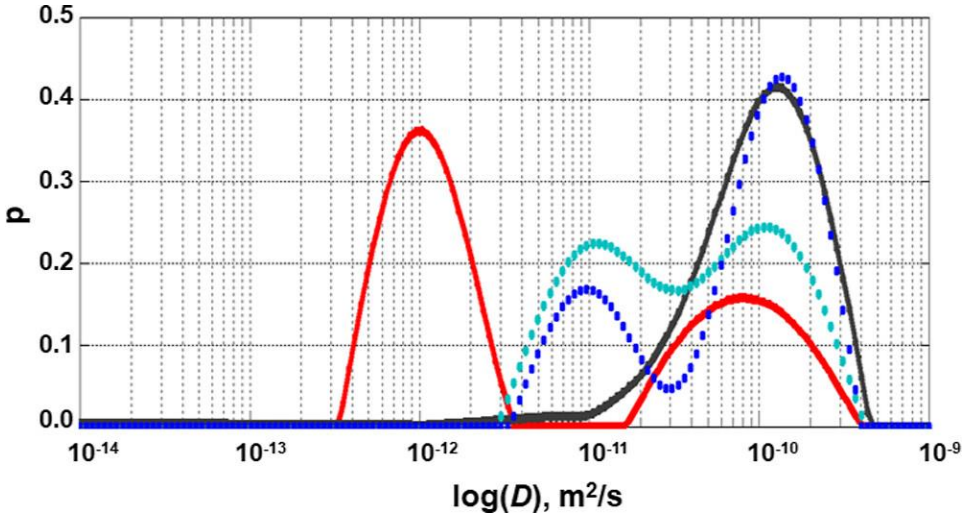

Figure 2.5 The distribution of diffusion coefficients in κ-casein solutions. The diffusion coefficient spectra are shown for 0.1% (black), 1% (blue), 4% (cyan), and 10% (red).

We noticed that incubation of κ-casein solutions with protein concentration below 10%, but not above 10%, at 4°C for more than 6 days led to visible changes characteristic of phase separation in the sample. Solutions, which were initially transparent, turned cloudy, with two layers becoming visible (Figure 2.6A). At the same time, the spin-spin relaxation exhibited two relaxation times for the protein, with the shortest time corresponding to species with the slowest diffusion coefficient, characterizing the dense phase.

Since high gradients allowed us to register diffusion coefficients in both the dilute (D_{max}) and dense phases (D_{min}), we tracked their changes depending on the sample concentration. Figure 2.6B shows that the diffusion coefficient of κ-casein in the dilute phase remains unchanged within the limits of experimental error over the entire concentration range. This indicates that as the concentration of κ-casein increases, κ-casein molecules enter the concentrated phase while the concentration of

phase depleted of κ-casein remains constant. At the same time, the relative population of the dilute phase decreases as indicated by the decrease of the population p_{max} of the component in the diffusion attenuation characterized by D_{max}. The minimum diffusion coefficient of the condensed phase decreases with increasing protein concentration, as expected, considering typical effects of concentration, such as the increase in solution viscosity. A clear transition point on the concentration dependence of D_{min} is observed, indicating the possibility of a qualitatively different diffusion regime.

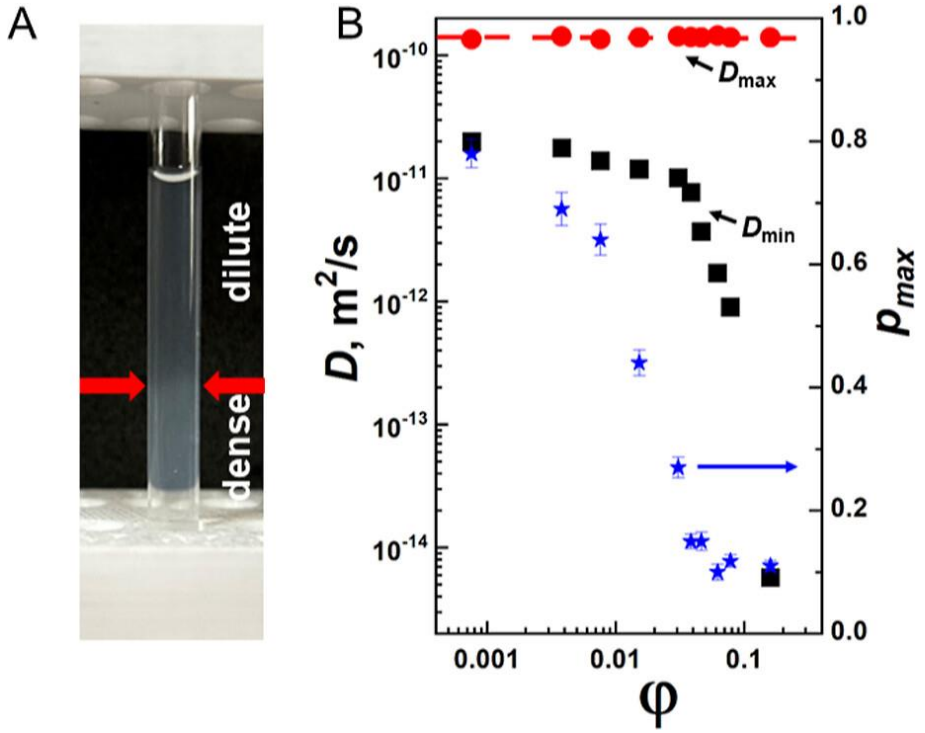

Figure 2.6 Phase separation in κ-casein solution. (A) Image of κ-casein sample, taken at ambient room temperature immediately after removal from 4 °C, shows the separation in dilute and condensed phases. (B) The dependence of D_{max} (red symbols) and D_{min} (black symbols) (left vertical axis) and p_{max} (blue symbols, right vertical axis) on κ-casein volume fraction φ.

We next investigated the translational diffusion in the 20% κ-casein solution. Figure 2.7A shows multi-exponential diffusion attenuations of the spin-echo signal recorded at different diffusion times for a 20% κ-casein solution using MODSTE pulse sequence. The initial slope of the diffusion attenuation remains unchanged, while the fraction of the slowest diffusing molecules, p_{min},

decreases with increasing diffusion time. As with the 0.5% κ-casein solution, the decrease of p_{min} with increasing diffusion time in the 20% κ-casein solution is caused by molecular exchange. However, in contrast to the 0.5% κ-casein solution, we observed the change of the diffusion coefficient D_{min} with diffusion time. The dependence of D_{min} on t_d is readily revealed by re-plotting

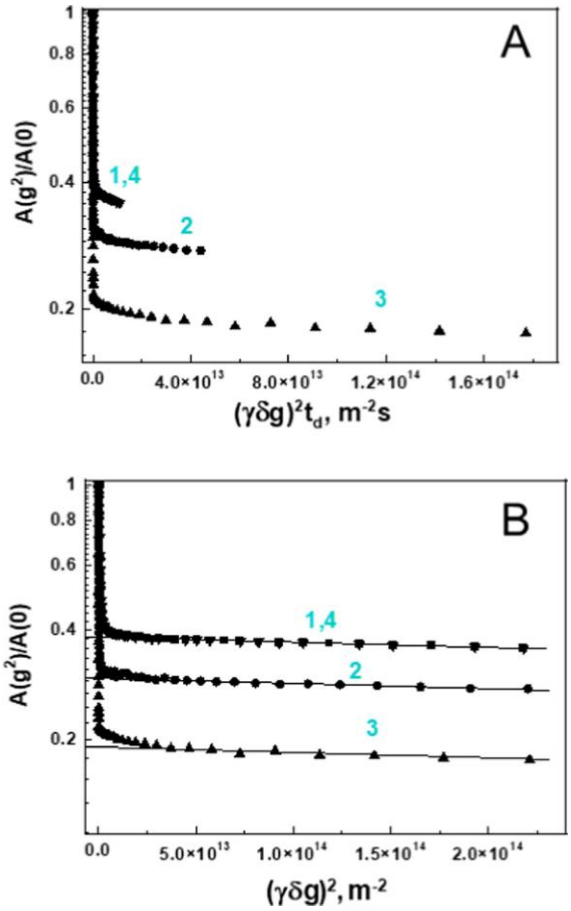

Figure 2.7 Diffusion attenuation for 20% κ-casein solution. (A) Curves 1–3 represent diffusion attenuations collected at t_d = 50, 200, and 800 ms, respectively. Curve 4 is a control. It was collected at t_d = 50 ms after the completion of experiments carried out at different values of t_d and coincides with curve 1, indicating no changes to the sample during the measurement time. (B) Curves 1–4 from panel A are replotted using coordinates $\log(A(g^2)/A(0))$ vs $(\gamma\delta g)^2$ to evaluate the dependence of the diffusion coefficient on diffusion time.

diffusion attenuations using coordinates $\log[A(g^2)/A(0)]$ vs. k·td (Figure 2.7B). In these coordinates, the slope of the slowest-diffusing component remains constant at all values of t_d, indicating that D_{min} is inversely proportional to the diffusion time. Accordingly, based on the Einstein relationship between the diffusion coefficient and the root-mean-square (RMS) displacement, $\langle r^2 \rangle = 6t_d \cdot D_{min}$, the RMS displacement of κ-casein molecules remains constant,

indicating that in the investigated t_d range in a 20% solution the molecules of κ-casein undergo anomalous, fully restricted diffusion. The estimated size of the restrictions is ≈ 42 ± 4 nm, which is about an order of magnitude greater than the hydrodynamic radius of κ-casein (3.56 nm). This value is comparable to the size of restrictions in the gel-like network of α-casein (50 ± 5 nm), indicating a geometrical similarity between their structures.

The observation of fully restricted diffusion, in the range of t_d from 50 to 800 ms, in the 20% solution of κ-casein suggests the formation of a three-dimensional gel network, in which individual κ-casein associates, already present in dilute solutions, interact with each other, losing their mobility as a whole[66, 69, 73, 74]. In this scenario, the diffusion coefficient D_{min} is associated with the movement of segments between the points of contact in the gel network. Using the dependence of p_{min} on t_d, we estimated that 34% of κ-casein molecules join the gel network, and their lifetime within the network is equal to 0.98 ± 0.08 s. This value is about twice as large as the lifetime of κ-casein molecules within associates in a 0.5% κ-casein solution. In the case of α-casein, ~93% of its molecules joined gel-like network and the lifetime in the associated gel state was about 3.5 s. Accordingly, the gel-like network formed by κ-casein is less extensive and more dynamic than that of α-casein.

Additionally, the state of 10-20% samples does not visibly change on storage time as the samples remain transparent and do not show any indication of macroscopic phase separation. Apparently, the formation of gel network restricts the κ-casein capability to phase separate, e.g. by imposing spatial restrictions.

2.4 DISCUSSION AND CONCLUSIONS

In this study, we examined the self-association behavior and translational diffusion of κ-casein in aqueous solutions across a broad concentration range, from 0.1% to 20%. Collectively, our data show that κ-casein self-associates over the entire concentration range. The associated structures are labile, as indicated by the exchange between κ-casein molecules in the associated state and the bulk solution. In the 20% κ-casein solution, these associated species further aggregate, forming a three-dimensional gel network, where κ-casein molecules remain about twice as long as in the associated states in solutions without gel formation. Additionally, in solution with concentrations below threshold of gel formation, the translational diffusion of κ-casein is unrestricted, whereas in the 20% solution, while the molecule remains in the gel network, it is fully restricted.

Only about a third of κ-casein molecules join the gel network at a time, which is the average dynamic equilibrium number of κ-casein molecules in the gel state. This allows us to compare the translational diffusion of remaining free κ-casein molecules, characterized by $\langle D \rangle$, to that of the closely related IDP α-casein[66] and to master curves for globular proteins[60] and flexible polymers[59] across different concentrations. The concentration dependence of the κ-casein diffusion coefficient differs from that of α-casein or globular proteins but follows the master curve for flexible polymers (Figure 2.2). The main difference between the master curves for flexible polymers and globular proteins is that for polymers, the transition from dilute to concentrated solutions is smoother and more gradual. The concentration dependence of the κ-casein diffusion coefficient demonstrates the same gradual transition. We attribute this behavior to the continuous self-association of κ-casein observed across the entire concentration range. Assuming that at larger concentrations, the associates of larger size are increasingly frequent, the average diffusion coefficient reflecting the whole distribution of diffusion coefficients (Eq. 4) would gradually

decrease. To explain the alignment of the κ-casein <$D(\varphi)$> dependence with the master curve of flexible polymers, we speculate that κ-casein can be equivalent to polymers that exist as heterogeneous species characterized by molecular weight (and size) distribution. Thus, the shape of the <$D(\varphi)$> dependence is sensitive to mass heterogeneity and molecular self-association. In contrast, globular proteins or α-casein do not show self-association in dilute solutions, with the onset of self-association being more discernible. Interestingly, the master curve for three high-generation poly(allylcarbosilane) dendrimers[106] follows the master curve for globular proteins over the entire concentration range evaluated, suggesting that the sharp transition from dilute to concentrated solutions is characteristic of monodisperse hydrodynamically compact molecules. Note that although the <$D(\varphi)$> dependence does not show a sharp crossover due to the effect of self-association, our data do not fully rule out the relative compactness of κ-casein molecules, even though they form many intermolecular contacts and join associates. Overall, we note, however, that applying the scaling law without a detailed investigation of the diffusion attenuation shape using ultra-high pulsed-field gradients would not enable discerning specific details of the translational diffusion in κ-casein solutions, such as gel formation or phase separation.

The behavior of the κ-casein diffusion coefficient <D> differs from that of α-casein in concentrated solutions[66]. The concentration dependence of the α-casein diffusion coefficient has an asymptotic behavior of φ^{-12}. The reason for such a strong concentration dependence of the α-casein diffusion coefficient is the formation of a three-dimensional gel network, supported by a fine balance of electrostatic repulsion and attractive hydrophobic interactions. Previously, a deviation of the concentration dependence of the diffusion coefficient from the master curve due to gel formation was also observed for several polymer systems, including gelatin-water and cellulose triacetate-benzyl alcohol[69, 73, 74]. Unlike α-casein, the diffusion coefficient of κ-casein does

not show such a strong dependence on concentration, despite the formation of a geometrically similar gel structure with comparable restriction sizes of about 50 nm. Besides the effect of continuous self-association, other factors partially explaining the observed difference may include the smaller number of κ-casein molecules joining the gel network (~34% vs. ~93% in the case of α-casein) and faster molecular exchange (the lifetime of α-casein in the bound state is 3.5 s). The translational diffusion coefficient directly reflects the size of the diffusing species, and our study clearly demonstrates that it allows insight into their self-association processes. As the experimental data on concentration dependence of different IDPs accumulate, it will become clear whether a general scaling law can be established for IDPs as it has been done for synthetic polymers, dendrimers, or globular proteins.

Considering the non-exponential shape of the spin-echo diffusion attenuation, we observe the pronounced separation of κ-casein molecules into fast and slow in terms of their translational mobility that progressively increases with time in solutions with κ-casein concentration below 10%. Macroscopic phases become visually apparent upon the incubation of κ-casein samples at 4°C for more than 6 days hours, with a clear boundary between the dilute and dense phases (Figure 2.6A). In this regard, the association observed in freshly made samples could be interpreted as the formation of microscopic LLPS before layer separation. The dense phase is characterized by extensive intermolecular contacts[107], significantly slowing down molecular translational diffusion. Previous PFG NMR studies demonstrated up to two orders of magnitude reduction in translational diffusion coefficient of a 103-residue disordered region of CAPRIN1 protein[108] or intrinsically disordered N-terminal 236 residues of the germ-granule protein Ddx4[109]. Our data show a similar level of reduction in translational diffusion coefficient of κ-casein in dense phase (Figure 2.6B).

Interestingly, we have not observed phase separation in solutions of α-casein, studied at similar experimental conditions at any concentration[66]. It is tempting to speculate that the role of κ-casein in the formation of casein micelle may be dictated by its unique association properties and tendency to form protein condensates. Future studies of different casein mixtures should clarify its role.

CHAPTER 3: DNA BINDING OF THE SLEEPING BEAUTY TRANSPOSASE[2]

3.1 INTRODUCTION

The use of DNA transposons as a means of delivering new genetic information into vertebrate organisms and cells [110, 111] relies on their ability to relocate from one DNA location to another [112, 113]. Specifically, when a transposition reaction is exploited to insert foreign genes into cells, relocation of a gene of interest typically occurs between a plasmid vector which carries the transposon and the genome of the targeted cell. DNA transposon-based technologies have become powerful tools in functional genomics [114, 115] and have been used for transgenesis and insertional mutagenesis in vertebrates [116]. Moreover, DNA transposon-based systems such as *Sleeping Beauty* (SB) or *piggyBac* have been examined in clinical trials for their potential to correct genetic diseases through the genetic modification of patient cells [117-120]. Notably, recent advances in the understanding of SB transposition mechanism have enabled the improvement of its safety for gene therapy by directing the transposon integration away from the transcriptional regulatory regions and exons of genes, thus increasing its potential utility for human applications [121].

A DNA transposon gene delivery system typically consists of two essential components: a transposon DNA containing the gene of interest flanked by terminal inverted repeats (TIRs) and a transposase enzyme that facilitates the transfer of the gene [122-128] (Figure 3.1). In the case of the SB transposon, the TIRs contain two imperfect 32-bp direct repeats (DRs). The outer DRs are situated at the ends of the transposon, while the inner DRs are located inside the transposon, about 165–166 bp away from the outer DRs [123, 129]. Each DR serves as a binding site for the transposase enzyme [130].

[2] The work presented in this chapter forms the basis of a manuscript undergoing revision at *Nucleic Acid Research*.

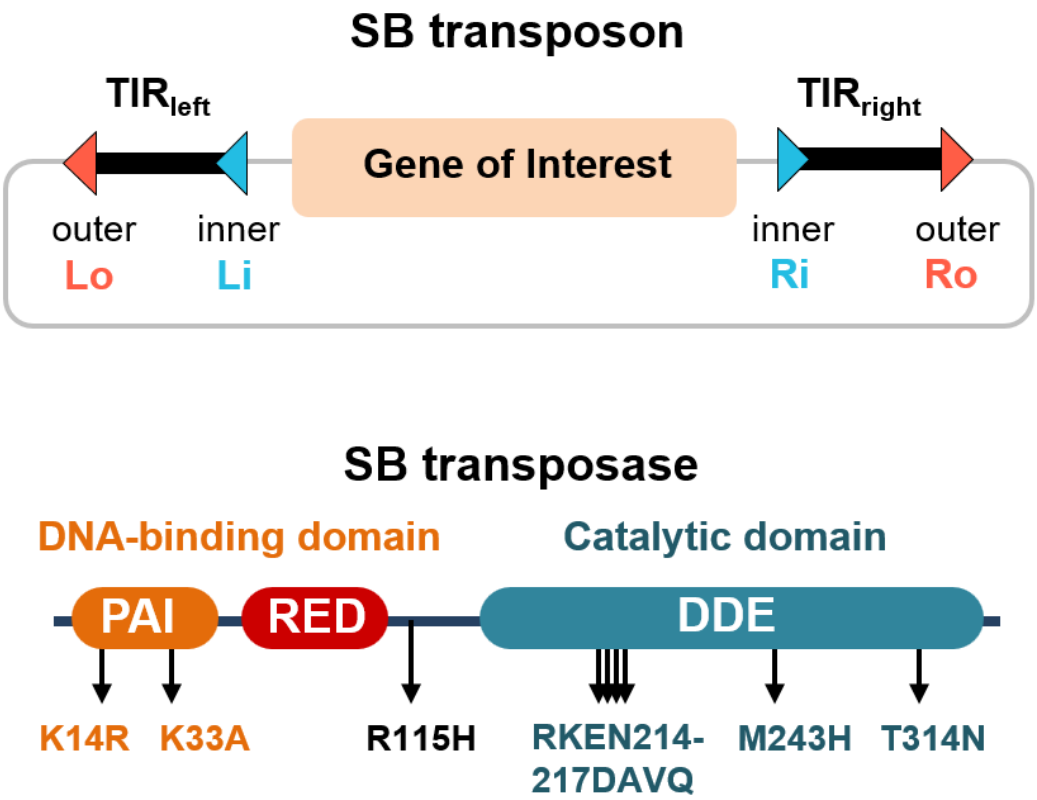

Figure 3. 1 The schematic presentation of SB transposon (top panel) and SB transposase (bottom panel) structures. SB transposon consists of the gene of interest to be delivered, flanked by terminal inverted repeats (TIR_{left} and TIR_{right}), each containing two (inner and outer) transposase binding sites. SB transposase consists of the catalytic domain and the DNA-binding domain containing two subdomains, PAI and RED. The location of the hyperactive mutations is indicated by arrows.

The efficiency of the DNA transposon system depends on how well the transposon DNA and the transposase enzyme work together, prompting modifications to both the transposon DNA and the transposase enzyme to achieve higher transposition efficiency [121, 124, 131]. Particularly, extensive efforts have been dedicated to engineering highly active transposases. Employing a molecular evolution (or genetic screening) approach, specific mutations have been identified in the *piggyBac* and SB transposases, resulting in a remarkable ~17- and ~100-fold increase of transposition rates, respectively [124, 128].

While the molecular evolution approach is a powerful tool for engineering transposases with increased activity, it does not provide insights into the underlying molecular mechanisms. Significantly, the mutations leading to hyperactivity are distributed throughout the entire protein. For instance, the SB transposase is a multidomain protein that consists of an N-terminal DNA binding domain and a C-terminal catalytic domain [123] (Figure 3.1). The DNA-binding domain is further divided into the PAI and RED subdomains (collectively called the paired domain, PAI + RED = paired), which are connected by a flexible linker [123, 132]. In the most active variant of the SB transposase, SB100X, the hyperactive mutations include K14R and K33A mutations in the PAI subdomain, the RKEN214-217DAVQ, M243H, and T314N in the catalytic domain, and R115H mutation on the linker between the catalytic and DNA-binding domains [124]. The distribution of hyperactive mutations across structurally and functionally distinct domains of the SB100X transposase suggests that multiple mutations may increase activity through different underlying mechanisms, ultimately producing a cumulative effect. Indeed, K33A was shown to have increased DNA binding activity and be hyperactive in transposition [133], whereas K14R was shown to be hyperactive in transposition, but did not change the DNA binding affinity of SB transposase [124]. The crystallographic structure of the catalytic domain of SB100X transposase (PDB code 5cr4) suggests that the substitution T314N located on the protein surface improves transposase activity by increasing its solubility [134]. Additionally, increasing the solubility of the SB100X transposase by introducing C176S and I212S surface amino acid substitutions resulted in the highly active transposase variant capable of penetrating cellular membranes autonomously, thus enabling efficient transgenesis in target cells by direct protein administration [135]. In contrast, the SB100X crystallographic structure reveals that M243H substitution likely assists in positioning the catalytic residue D244 in the active site, while the RKEN214-217DAVQ mutations

immediately follow the loop that participates in DNA-binding likely aid in shaping and optimally positioning this loop for DNA interactions [134]. Furthermore, RKEN214-217DAVQ mutations favorably alter residue dynamics and mechanical couplings within SB transposase [136].

One aspect of structure-function activity relationship in DNA transposons, particularly the SB transposon, has remained underexplored – how the DNA-binding capability of the DNA-binding domain of the transposase impacts its activity in transposition. Previously, we have solved the structure of the PAI subdomain [4] and demonstrated that it selectively binds to the transposon DNA in its folded conformation, e.g., through the conformational selection mechanism [11]. This means that PAI subdomain's interaction with DNA is contingent upon its pre-existing conformation, and only those molecules that exist in a DNA-binding-competent state can initiate binding. Building upon these findings, in this study, we further investigate the relationship between folding, DNA-binding, and transposition activity of the SB transposase. Using a rational, structural biology approach, we engineered a structurally stable variant of the PAI subdomain with enhanced DNA-binding capabilities by introducing the H19Y mutation. With this variant, we were able to determine the structure of the PAI-transposon DNA complex and gain mechanistic insights into the PAI-DNA binding process. We further introduced the H19Y mutation into the full-length, first-generation SB10 transposase and its hyperactive version SB100X, which allowed gaining new structural and mechanistic information about their DNA-binding properties. Finally, we demonstrated that increased structural stability of the PAI subdomain due to the H19Y mutation correlated with enhanced transposition activity of the SB10 transposase and with increased DNA-binding specificity of SB100X transposase.

3.2 MATERIAL AND METHODS

3.2.1 Protein expression, purification, and sample preparation

The nomenclature of proteins used in this study is as follows: PAI (the PAI subdomain from SB10 transposase without modifications), PAI-K14RK33A (PAI with mutations K14R, K33A), H19Y (PAI with three mutations K14R, H19Y, K33A), SB10 (the first-generation full-length SB10 transposase as reported in [123]), SB10-H19Y (SB10 with H19Y mutation), SB100X transposase (full-length 100-times more active SB transposase as reported in [124]), SB100X-H19Y (SB100X transposase with H19Y mutation). For reference, amino acid sequences of all proteins are provided in Appendix B Table 1. All proteins were expressed and purified following a previously reported protocol [4]. DNA plasmids encoding proteins were ordered from GenScript USA Inc. The proteins were expressed in BL21-AI *E. coli* cells. For ^{15}N and ^{13}C isotopic labeling, bacterial cells were grown in M9 medium using $^{15}NH_4Cl$ and ^{13}C-glycerol (Cambridge Isotope Laboratories) as the sole nitrogen and carbon sources, respectively. The proteins were purified by metal chelating chromatography using a Ni-NTA Agarose (Thermo Fisher Scientific). Samples with PAI subdomain or its variants were prepared in an aqueous 25 mM sodium phosphate buffer at pH 5.2 or pH 7.4. For NMR experiments, the buffer contained 5% D_2O. To assess the effect of pH on protein structure, the solution pH was increased from 4.5 to 8.5 in 0.5 increments by adding microliter quantities of NaOH. Purified SB10 and SB100X full-length transposases were prepared in an aqueous 50 mM TRIS buffer at pH of 7.5, containing 5 mM $MgCl_2$, 300 mM NaCl, 2 % glycerol, 50 mM Arg-Glu mixture, and 1 mM TCEP. For DNA binding experiments, unlabeled or Cy5-labeled DR-core, Li, and Lo sequences were synthesized by IDT (Integrated *DNA* Technologies, Inc.). Cy5 label was attached to the 5’ end of the forward DNA.

3.2.2 Circular dichroism (CD) spectroscopy

CD measurements were performed on a Jasco-715 spectropolarimeter, equipped with a Peltier temperature control system, using quartz glass cell with a path length, *l*, of 1 mm. Far-UV CD spectra were recorded in the range of 190 to 250 nm at room temperature. The corresponding buffer (25 mM sodium phosphate buffer) baseline was subtracted from each spectrum. To assess the effect of pH on protein structure, the solution pH was increased from 4.5 to 8.5 in 0.5 increments. Spectra were recorded using a 50 nm/min scan rate with a 4 s response and a 1 nm bandwidth. Reported spectra are averages of 2-5 scans and are expressed as mean-residue molar ellipticity (MRE) calculated by using the following relation:

$$MRE = \frac{M_0 \theta_\lambda}{100 \cdot C \cdot \lambda}, \qquad (1)$$

where M_0 is the mean residue molar mass, θ_λ is the measured ellipticity in degrees, and C is the total concentration of protein. The value of M_0 was obtained by dividing the molecular weight of the protein with the number of amino acid residues in it. To follow the folding of PAI subdomain and its mutants, mean residue ellipticity at 222 nm, $[\theta]_{222}$, was used to assess protein structural changes.

CD spectra were analyzed on the DichroWeb server [137, 138] to estimate the fraction of secondary structure elements. The dependence of the fraction of alpha-helical conformation, f_h, on pH was fit to a modified Henderson-Hasselbalch equation:

$$f_h = \frac{f_a + f_b\left(10^{n(pKa-pH)}\right)}{1+10^{n(pKa-pH)}}. \qquad (2)$$

In this equation, f_a is the fraction of alpha-helical conformation at acidic pH prior to transition, f_b is the fraction of alpha-helical conformation at basic pH after transition, pKa is the pH value

corresponding to an inflection point of the dependence, and the *n* value (Hill coefficient) is the slope at the inflection point, which determines the number of protons involved in the transition. The Hill coefficient was set to be a free parameter during fitting.

3.2.3 NMR spectroscopy

NMR experiments were performed on a Bruker Avance-III 950 MHz and 700 MHz spectrometers equipped with helium-cooled cryoprobes at David H. Murdock Research Institute (DHMRI) and in the Molecular Education, Technology, and Research Innovation Center (METRIC) at North Carolina State University, respectively. NMR experiments for structure determination were carried out at 5 ^{o}C, whereas DNA-binding experiments were carried out at 35 ^{o}C. Proton chemical shifts were calibrated with respect to water signal relative to DSS $(CH_3)_3Si(CH_2)_3SO_3Na$) and ^{15}N and ^{13}C chemical shifts were indirectly referenced to DSS [139]. Sequence-specific resonance assignments have been performed using 3D HNCACB, CBCA(CO)NH, HCCH-TOCSY, ^{15}N-separated TOCSY-HSQC and HSQC-NOESY, and ^{13}C-separated NOESY-HSQC experiments as described in original references [140]. Interproton distance restraints were derived from Nuclear Overhauser Enhancement (NOE) signals in ^{15}N-NOESY-HSQC and ^{13}C-NOESY-HSQC experiments, collected at 120 ms mixing time. Hydrogen bond restraints were identified from the pattern of sequential and inter-helical NOEs involving NH and CαH protons and with evidence of slow amide proton-solvent exchange, monitored with a series of 2D [^{1}H,^{15}N]-HSQC spectra recorded in 100% D_2O. All NMR spectra were processed using the NMRPipe software [141]. Linear prediction was applied for both ^{15}N and ^{13}C dimensions to double the data size and improve the digital resolution. Cosine square window function and automatic zero filling were applied to ^{1}H, ^{15}N and ^{13}C dimensions. NMR spectra were analyzed with programs CARA [142] and NMRView [143].

Three-dimensional structures of H19Y were calculated by utilizing internuclear distances from NOESY spectra, dihedral angle restraints generated from the chemical shifts using the program TALOS [144] and hydrogen bond distance restraints using the program XPLOR-NIH [145]. The 8 minimum energy structures with no restraint violations were selected from a set of 100 calculated structures as a representative ensemble based on the absence of NOE violations greater than 0.5 Å and dihedral angle violations greater than 5° and assessed for stereochemical quality using the PROCHECK program [146]. Experimental restraints and structural statistics are summarized in Appendix B Table 2. Molecules were visualized and aligned using the program PYMOL [147]. The coordinates and related information were deposited to Protein Data Bank (PDB code 6URS). Chemical shift information was deposited to Biological Magnetic Resonance Data Bank (BMRB code 30680).

3.2.4 Microscale Thermophoresis (MST)

MST experiments were performed using a Monolith NT.115 (NanoTemper) instrument. For protein-DNA binding experiments, we used Cy5-labeled DR-core, Li, or Lo sequences purchased from IDT (Integrated *DNA* Technologies, Inc.). For H19Y protein-protein interaction experiments, we used RED-NHS 2nd Generation dye that reacts with primary amines (NanoTemper) and performed labelling at a 1:1 ratio (dye molecule to protein molecule) to have approximately one dye molecule per protein and its location statistically distributed over the protein surface to avoid interference with binding. 40 nM of labelled H19Y or 30 nM of Cy5-labelled DNA was mixed with increasing amounts of unlabeled protein, and the mixtures were incubated for 30 minutes in the dark at room temperature. The experiments were done using Monolith NT.115 premium capillaries. The assay buffer contained 25 mM sodium phosphate buffer at pH 5.2 or pH 7.4, 150 mM NaCl, and 0.05% Tween-20 to prevent sample sticking to the capillaries for H19Y

experiments. For experiments with full-length SB10 and SB100X transposases, the assay buffer contained 50 mM TRIS, 5 mM $MgCl_2$, 300 mM NaCl, 2 % glycerol, 50 mM Arg-Glu mixture, 1 mM TCEP, 0.1% Triton X-100, and 0.1 mg/mL BSA to prevent sample sticking to the capillaries, prepared at pH of 7.5. Initially, resulting dose response (F_{norm}) curves obtained from normalized fluorescence were analyzed by least-squares curve fit using the Nanotemper software. F_{norm} is the normalized fluorescence signal, which corresponds to the ratio of the fluorescence value measured in the heated state to the fluorescence value measured in the cold state before the IR-laser is turned on. Subsequently, the data from different experiment repeats (>3) were analyzed together and figures were created using Origin 22 software.

3.2.5 Fluorescence Lifetime (FLT)

Fluorescence lifetime measurements were performed using a home-built time-resolved fluorimeter, equipped with a QuadraCentric sample compartment with a cuvette holder (Horiba Scientific), a passively Q-switched microchip YAG laser (SNV-20F-100, 532 nm, 20 kHz, Teem Photonics), a photomultiplier (H6779-20, Hamamatsu), and a digitizer (Acqiris DC252, Agilent). A 532/18 nm BrightLine single-band bandpass filter (Semrock) and a polarizer set at the magic angle (54.7^o) were used in the detection arm. All experiments were carried out at room temperature (21 ± 1 ^{o}C). The instrument response function (IRF) was obtained before each measurement using buffer as a scatterer and accounted for in the data analysis. The sample solution with 25nM fluorescently labeled Cy3-DNA titrated against concentration of proteins ranging from sub-nM to 400nM range was added to the observation cuvette one by one, and the fluorescence waveform was acquired by averaging fluorescence transients from one thousand laser pulses. All analyses of time-resolved fluorescence data were performed using the software package FargoFit, designed by I. Negrashov, which executes a global least-square fitting of multiple time-resolved fluorescence

waveforms using different models that allow linking fitting parameters between waveforms. The obtained waveforms of donor fluorescence were best fitted by two exponential components, convoluted with the IRF:

$$I(t) = a_1 \exp\left(-\frac{t}{\tau_1}\right) + \ a_2 \exp\left(-\frac{t}{\tau_2}\right) \quad (3)$$

The intensity weighted average lifetime <τ> was then calculated as[148]:

$$< \tau >= \frac{a_1\tau_1^2+a_2\tau_2^2}{a_1\tau_1+a_2\tau_2} \quad (4)$$

3.2.6 Fluorescence anisotropy (FA)

The same instrumental setup was used for fluorescence anisotropy measurements, with a crucial difference: instead of detection arm being set at the magic angle (54.7^o), each sample was measured once with the polarizer set at 0^o (vertical) and then with polarizer set at 90^o(horizontal). The obtained waveforms of fluorescence were fitted by two exponential components, convoluted with the IRF. The vertical and horizontal intensities, $I_{\parallel}$ and $I_{\perp}$, were then used to calculate fluorescence anisotropy, r, using the formula:

$$r = \frac{(I_{\parallel}-GI_{\perp})}{(I_{\parallel}+2GI_{\perp})} \quad (5)$$

where the grating factor G represents the instrument's bias towards horizontally polarized light compared to vertically polarized light in its emission optics system. G was calculated to be 1.03.

3.2.7 Protein-DNA docking using HADDOCK

The structure of the DR-core double helix was generated using the MAKE-NA program [149-152]. Chemical shift perturbations were used as ambiguous interaction restraints (AIRs) to drive the docking process using HADDOCK version 2.2 [153]. Residues having a weighted chemical shift perturbation upon binding to DNA greater than two standard deviations and displaying high solvent accessibility (>50%) were selected as active residues. Solvent accessibility for the active residues was calculated using the program NACCESS (31). The HADDOCK program was allowed

to define passive residues (residues near the interface that may play a role in the formation of the complex) as docking parameters automatically. The lowest energy structure of H19Y calculated with XPLOR-NIH was selected as the starting H19Y structure for the docking. All DNA bases were selected as active bases. During the rigid body energy minimization, 10 000 structures were calculated, and the 200 best solutions based on intermolecular energy were used for the semi-flexible, simulated annealing followed by an explicit water refinement. Docked structures corresponding to the 200 best solutions with lowest intermolecular energies were generated. The 200 solutions were clustered using a 1.0 Å RMSD cut-off criterion into eight different clusters. The clusters were ranked based on the averaged HADDOCK score of their top 10 structures. From these, we selected the top cluster with a Z-score of -1.3, a HADDOCK score of -145 ± 7.5, and the most favorable intermolecular energies, to represent the model of the H19Y-DR-core complex.

3.2.8 Protein-DNA docking using PD-DOCK

The PD-DOCK program [154, 155] employs a rigid protein-DNA docking procedure and searches for the optimal complex solution using a Monte Carlo algorithm and a knowledge-based, orientation potential for assessing protein-DNA interaction [155]. 200 independent docking experiments were carried out to maximize the conformational search space, resulting in a total of 200 predicted protein-DNA complex structures. The interaction energy for each protein-DNA complex was calculated, and expressed in arbitrary units, as a sum of the energies of all residue (*i*) and DNA base (*j*) interactions using our knowledge-based, distance and orientation-dependent protein-DNA interaction potential as shown in Equation 1 [155, 156]:

$$E = \sum E_{ij}^{0}(r, \varphi), \qquad (3)$$

where r and φ represent the distance between residue *i* and base *j* and the angle between residue sidechain and the base plane respectively.

These complex structures were then clustered hierarchically based on their structural similarity. The distance between two complex structures was the root-mean-square deviation of protein Cα atoms after the corresponding DNA structures were superimposed. The hierarchical clustering was carried out using a complete linkage approach, in which the distance between two clusters is represented by the largest distance among all pairwise distances between the members of two clusters. Using a bottom-up approach and an RMSD cut-off of 3 Å, we divided the structures into clusters and ranked the clusters based on the protein-DNA binding energy of the representative structure, i.e., the complex with the lowest energy within the cluster. The top cluster was selected and used for follow-up analyses.

3.2.9 EMSA experiments

The Light Shift EMSA Kit (Thermo Scientific #20148) was used as recommended by the manufacturer. Briefly, 20 µL reactions contained 2 µl of 10x Mobility Shift Buffer, 100 µM EDTA, 1 µg of poly dIdC, 0.015 pmol biotinylated, double-stranded oligonucleotide, protein, and water. The reaction was incubated for 20 min at room temperature and run on a 6 % native acrylamide gel. The double-stranded probe for the binding reactions was prepared by mixing equimolar amounts of the single-stranded oligonucleotides 5'-BIOTIN-TACAGTTGAAGTCGGAAGTTTACATACACTTAAG-3' and 5'-BIOTIN-CTTAAGTGTATGTAAACTTCCGACTTCAACTGTA-3', boiling and annealing by allowing the solution to cool down to room temperature overnight.

3.2.10 Transposition assays

Antibiotic resistance-based transposition assays were done by plating 300.000 HeLa cells per well on a 6-well plate one day before transfection and transfecting 100 ng of pT2/HB-puro and 50 ng pFV-SB10, pFV-SB10-H19Y, pFV-SB100X, or pFV-SB100X-H19Y in 250 µl Opti-MEM by

using the Mirus-LT1 transfection reagent (Fischer Scientific). 60 % of the transfected cells were seeded on a 10 cm dish 24 h post-transfection with medium containing puromycin. Puromycin-resistant cell colonies were allowed to grow on the dishes for 2 weeks, at which time point they were stained and counted.

3.3 RESULTS

3.3.1 DNA-binding affinity of the full-length Sleeping Beauty transposase

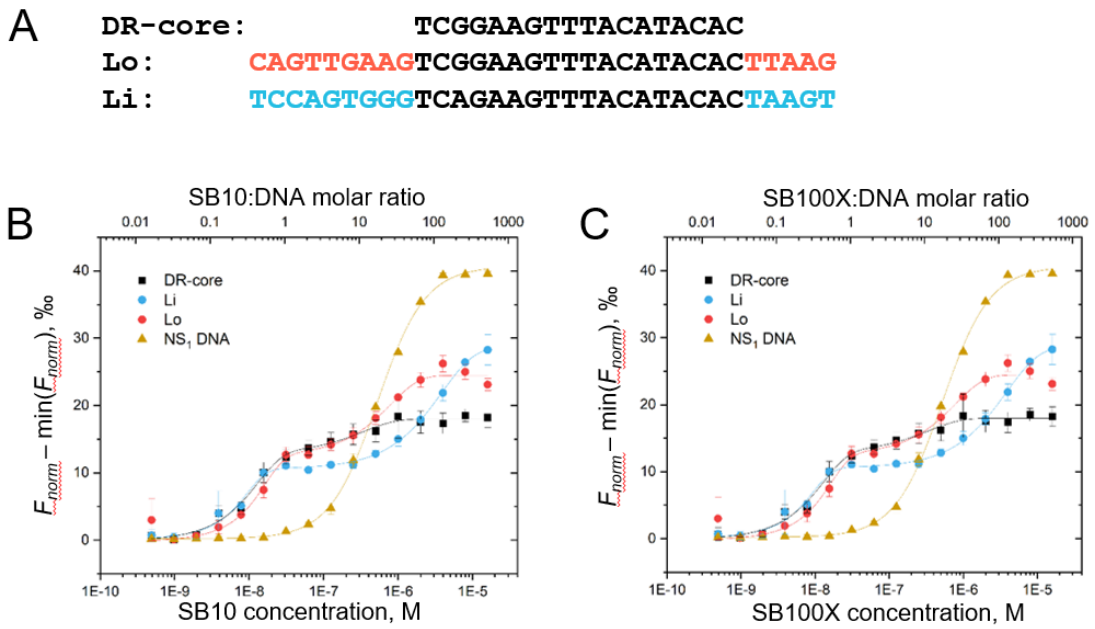

Figure 3. 2 First-generation SB10 and hyperactive SB100X transposase binding to the transposon DNA. (A) SB transposon DNA sequences of DR-core, outer (Lo), and inner (Li) transposase binding sites of the left TIR. (B-C) Binding affinity for SB10 (B) and SB100X (C) binding to DR-core, Lo, and Li was evaluated using the MST titration experiment with Cy5-labeled DNA sequences held at a constant concentration of 30 nM, to which unlabeled proteins were added at gradually increasing concentrations. As a control, a nonspecific (NS_1) DNA sequence (5'-ACCTTCCTCCGCAATACTCCCCCAGGT-3') was used. To facilitate the comparison of binding curves for different DNA sequences we show data on the same scale. For this, we subtracted the respective minimum value of F_{norm} for each curve. All data were evaluated over the T-jump time interval, e.g., within 1 s of IR-laser activation. Experimental error bars show S.E. for $n \geq 3$ separate experiments. The solid lines represent Hill fits to the experimental data. MST binding curves reveal specific and nonspecific DNA binding modes.

SB transposition begins with the site-specific binding of the SB transposase to the transposon DNA, containing four 32 bp-long transposase binding sites known as imperfect direct repeats (IR-DRs) (Figure 3.1). These IR-DRs contain a shared 18-bp DR-core sequence, while the surrounding adjacent sequences vary among the four transposase-binding sites, with this variation being significant for efficient transposition [130] (Figure 3.2A). The DNA-binding constants (KDs) of the SB transposase have not yet been reported. Therefore, we started with a broad range of protein concentrations to coarse-grain the DNA-binding affinity first, using the first-generation

reconstructed SB10 [123] and hyperactive SB100X [124] full-length transposases and the DR-core sequence and DNA sequences representing the left outer (Lo) and inner (Li) transposase binding sites (Figure 2A). Binding affinity was evaluated using the MST titration experiment with Cy5-labeled DNA sequences held at a constant concentration of 25nM-30nM, to which unlabeled proteins were added at gradually increasing concentrations. As a control, a nonspecific (NS_1) DNA sequence (5'-ACCTTCCTCCGCAATACTCCCCCAGGT-3') was used. Both SB10 and SB100X showed a biphasic binding mode (Figures 3.2B-C) to DR-core, Li, and Lo sequences, and, interestingly, bound NS_1 DNA sequence at protein concentrations close to micromolar (protein to DNA molar ratios above 20:1, top axes). The second transition in the biphasic binding to DR-core, Li, and Lo transposon DNA sequences occurred at protein concentrations similar to those for NS_1 DNA, indicating that it likely corresponds to a nonspecific DNA binding mode. The binding of SB100X to NS_1 DNA occurred at lower protein concentrations than SB10, indicating its stronger tendency to bind DNA nonspecifically. Calculated equilibrium binding constants (KDs) for NS_1 DNA were 2.60 ± 0.16 μM for SB10 and 0.53 ± 0.02 μM for SB100X.

The first transition in the biphasic binding to DR-core, Li, and Lo transposon DNA sequences, occurring with nanomolar binding affinities, is not observed in NS_1 DNA binding by SB10 and SB100X, and the MST signal plateaus above approximately a 1:1 protein to DNA molar ratio. Thus, this transition represents the specific binding mode of SB10 and SB100X to the transposon DNA sequences. To determine binding constants for specific DNA binding more accurately, we sampled the protein concentration range of the first transition in more detail. Estimated K_d values are given in Table 3.1, and titration curves are shown in Appendix B Figures 1A-F. At 0.05 level, SB10 shows stronger affinity when binding Li and Lo, but weaker affinity to DR-core.

Table 3.1. KD values (nM) for full-length SB10 and SB100X transposases binding to DR-core, Li, and Lo transposon DNA sequences determined by using MST.

DNA ↓ / Protein →	SB10	SB100X
DR-core	25.3 ± 0.9	17.3 ± 1.1
Li	17.3 ± 1.2	23.6 ± 2.2
Lo	16.6 ± 2.6	23.0 ± 1.9

Previously, we showed that the PAI subdomain of the SB transposase must be folded to bind the transposon DNA [11]. We next sought to determine the impact of the structural stability of the PAI subdomain on the DNA-binding properties of SB10 and SB100X transposases.

3.3.2 H19Y mutation promotes structural stability of the primary DNA-recognition subdomain of the Sleeping Beauty transposase

The deprotonation of the H19 sidechain drives folding of the PAI subdomain. The PAI subdomain is mostly unstructured close to physiological conditions but adopts a well-structured conformation at low temperature (5 °C) in the presence of NaCl at the concentration of ~ 600 mM and at solution pH values greater than 7.0 [4, 11]. To identify the molecular determinants of PAI subdomain folding, we monitored the pH-induced conformational transition of the PAI subdomain and its double mutant PAI-K14RK33A, as in the hyperactive SB100X transposase (Figure 3.1), by analyzing far-UV CD spectra (190-260 nm) within the pH range of 4.5 to 8.5 (Figure 3A). Basic pH induces helix formation, evidenced by a gradual increase in negative peak intensities at 208 and 222 nm, and a positive peak intensity at 192 nm. The pH dependence of the CD spectra (Figure 3A) shows an isodichroic point at ~204 nm, suggesting a two-state process. Figure 3B displays the content of alpha-helical conformation estimated for different pH values using the DichroWeb server [137, 138]. The simultaneous presence of K14A and K33A mutations does not alter the pH-dependent folding of the PAI subdomain. Fitting MRE values measured at 222 nM with a modified

Henderson-Hasselbalch equation (2), under the assumption of a two-state transition, yields the p*Ka* values of 5.98 ± 0.06 and 5.96 ± 0.05 for the PAI and PAI-K14RK33A, respectively. These pKa values coincide with the pKa value of histidine sidechain protonation [157, 158], indicating that the folding of the PAI subdomain at basic pH is likely related to the deprotonation of histidine sidechain(s).

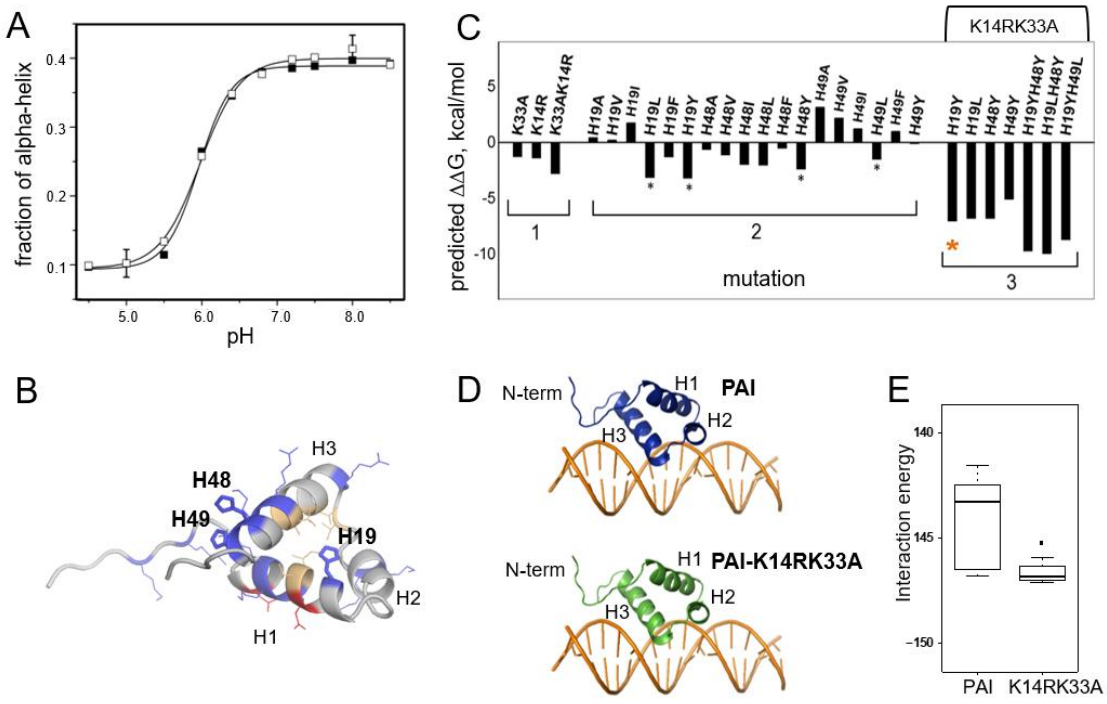

Figure 3.3 The pH-induced folding of the PAI subdomain. (A) Fraction of the helical conformation versus pH of PAI (filled squares) and PAI-K14RK33A double mutant (open squares) estimated using the DicroWeb server [137, 138]. Solid lines represent a global sigmoid dose-response fit of the data and are included as a guide to the eye. (B) Three-dimensional structure of the PAI-WT subdomain (PDB ID 2M8E). Charged residues are shown in blue (Arg, Lys, His) or red (Asp, Glu). Stick representation highlights histidine residues. Hydrophobic residues surrounding H19 are shown in light orange. Helices H1, H2, and H3 are labeled. (C) The predicted effect of mutations on the Gibbs free energy of unfolding of the PAI subdomain. Group 1 shows the effect of K14R, K33A, and K14RK33A mutations. Group 2 shows the effect of H19, H48, and H49 single mutations. Stars label the most energetically favorable substitutions. Group 3 shows the effect of H19, H48, and H49 mutants in the presence of double K14RK33A mutation. The values of $\Delta\Delta G$ were calculated using the Eris protein stability prediction server [159]. (D) Top PD-DOCK-predicted structures of protein-DNA complexes for PAI subdomain and K14RK33A mutant. (E) Interaction energies of PAI subdomain or K14RK33A mutant complexes with DR-core. The K14RK33A mutations provide increased stabilization of protein-DNA complex formation without changing its overall arrangement.

The PAI subdomain contains three histidine residues: H19, H48, and H49 (Figure 3.3C). Histidines H48 and H49 are located at the C-terminus of the PAI subdomain (on helix H3) and are surface-exposed. In contrast, H19 is located at the end of helix H1, with its sidechain oriented toward the

interior of the protein and surrounded by the hydrophobic residues I15, L25, V39, I42, and V43 (colored light orange in Figure 3.3C). Therefore, it is plausible to assume that deprotonated H19 (at a pH greater than the pKa value of 5.98 ± 0.06) promotes PAI subdomain folding. To support this assumption, we calculated the changes in the Gibbs free energy of folding, ΔΔG, for several H19, H48, and H49 mutants in the absence or presence of K14RK33A mutations, using the Eris protein stability prediction server with the flexible backbone option [159]. Group 1 in Figure 3.3D shows the individual and cumulative favorable effects of K14R and K33A mutations on PAI stability, as indicated by negative ΔΔG values. These effects are likely due to a decrease in charge (K33A) and a change in sidechain geometry (K14R), favorably affecting the local environment of these residues, as the ERIS server uses a stochastic sidechain optimization around a given structure [159, 160]. Group 2 shows histidine substitutions. Several substitutions, including H19L, H19Y, and a few H48 and H49 substitutions show increased stability (Figure 3.3D, group 2, marked with stars). Since H48 and H49 are adjacent, substituting either for a hydrophobic amino acid is electrostatically advantageous. The ERIS prediction for histidine substitutions with a hydrophobic residue agrees with our CD data on the pH-dependence of PAI and PAI-K14RK33A folding (Figures 3A-B). The effect of histidine substitutions, including H19L and H19Y, is enhanced by the presence of two hyperactive mutations K14R and K33A (group 3, Figure 3.3D). Based on the ΔΔG calculations, we selected the H19Y substitution for experimental testing. H19 was mutated to tyrosine, and the folding of the triple mutant K14RH19YK33A of the PAI subdomain (referred to hereinafter as the H19Y) was assessed by NMR spectroscopy.

The [^{1}H,^{15}N]-HSQC spectra of PAI (left) and H19Y (right), both collected under identical buffer conditions at pH 5.2 and 5 °C, are shown in Figure 3.4A. The spectrum of PAI displays a very narrow distribution of cross-peaks, with chemical shifts for many of the ^{1}H and ^{15}N resonances centered around 8 ppm and 120 ppm, indicating that the respective amino acid residues are in a random coil conformation and, therefore, the PAI subdomain is not folded [4, 11]. In contrast, H19Y reveals a drastically different [^{1}H,^{15}N]-HSQC spectrum with widely dispersed resonances,

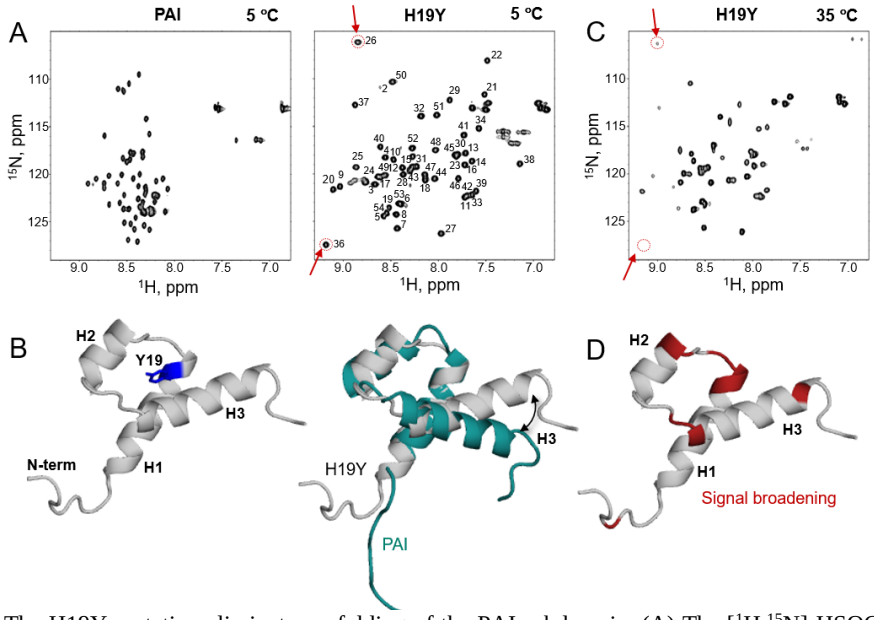

Figure 3.4 The H19Y mutation eliminates unfolding of the PAI subdomain. (A) The [^{1}H,^{15}N]-HSQC spectrum of 0.2 mM PAI (left panel) and H19Y (right panel), both collected at pH 5.2 and 5 °C. The backbone assignments for H19Y are labelled. Note that the peak numbering is made consistent with previous literature on the SB transposase and differs from our previous work [4] by two amino acids. (B) A representative structure of H19Y from the ensemble of minimal energy structures with Y19 highlighted (left) and overlaid with the PAI structure in teal (right). The superposition of structures highlights the difference in the orientation of helix H3. (C) The [^{1}H,^{15}N]-HSQC spectrum of H19Y collected at 35 °C at pH 5.2. Red arrows exemplify the observed peak broadening as compared to 5 °C in panel A. (D) The amino acid residues exhibiting significant signal broadening are colored (red) on the H19Y structure.

indicative of a folded structure. This difference is solely due to the H19Y mutation, because PAI-K14RK33A has the same folding properties as PAI (Figure 3.3B). The backbone assignments for H19Y were derived by standard triple-resonance NMR techniques using ^{15}N,^{13}C-labeled protein. Note that the peak numbering is made consistent with previous literature on the SB transposase and differs from our previous work [4] by two amino acids.

3.3.3 NMR solution structure of the H19Y resembles the PAI subdomain structure.

We determined the H19Y solution structure based on experimentally measured NOE (Nuclear Overhauser Effect) and hydrogen bond distance restraints, as well as dihedral angle restraints generated using the program TALOS [144] from chemical shift data. The left panel of Figure 3.4B shows a representative H19Y structure selected from the ensemble of lowest energy structures chosen for the analysis with Y19 substitution indicated. The coordinates for the 8 lowest energy structures with no restraint violations were deposited in the Protein Data Bank (PDB) under the accession code 6URS; the BMRB ID for this entry is 30680. Similar to PAI (PDB ID 2M8E), H19Y folds into a compact, three-helix domain. The PAI and H19Y structures can be superimposed with a root-mean-square displacement (RMSD) of 2.6-2.7 Å (Figure 3.4B, right panel). Interestingly, while the helix H1 is superimposed with a small RMSD difference, the orientation of helix H3 differs between the two structures, likely due to a better fit of the tyrosine sidechain that in contrast to histidine avoids electrostatic repulsion. However, the overall fold of the two structures is similar.

3.3.4 Structural stability of H19Y at elevated temperatures.

To determine whether H19Y maintains its structure at physiological temperatures, we collected a series of its [^{1}H,^{15}N]-HSQC spectra across a temperature range of 5 to 45 °C. As the temperature increases, the spectral quality degrades, resulting in significant signal broadening, while the peak dispersion remains unchanged (compare Figure 3.4C to the [^{1}H,^{15}N]-HSQC spectrum of H19Y at 5 °C in Figure 3.4A; a few peaks are circled and indicated by arrows as an example). [^{1}H,^{15}N]-HSQC spectra at all temperature increments are provided in Appendix B Figure S2. By color-coding the H19Y structure (Figure 3.4D), we determined that the broadened peaks primarily originate from the amino acid residues in the loop region between helices H1 and H2, as well as

in the region at the end of the loop between helices H2 and H3, leading into helix H3. This is in contrast to the regions anticipated to be involved in protein-protein interactions, as inferred from amino acid sequence analysis and the structures of closely related Mos1 or Tc3 transposases [161, 162]. We previously observed a similar temperature-dependent spectral behavior for the PAI subdomain at pH 7.0, where we demonstrated that the peak broadening arises from the folding-unfolding process of PAI rather than its dimerization or higher-order oligomerization [11]. The similar spectral behavior of H19Y and the localization of residues exhibiting broadened NMR signals to the regions outside the predicted protein-protein interaction interface suggest that the mutant exists in an equilibrium between folded and unfolded conformations, resulting in signal broadening. On the other hand, signal broadening occurs at the DNA-binding site or in its vicinity, supporting the idea that structural stability may improve the DNA-binding properties of the PAI subdomain. It is worth noting that the crystal structure was only obtained for the catalytic domain of the SB transposase, likely due to the conformational flexibility of the DNA-binding domain, particularly PAI subdomain, which precludes crystal formation.

To address the possibility of dimerization or oligomerization of H19Y, which could contribute to signal broadening, we initially attempted MST experiments to determine the dissociation constant of H19Y. The MST data collected at pH 5.2 and 7.4 suggested minimal protein association at the protein concentrations employed in our NMR experiments and up to millimolar concentrations, at which protein aggregation induced in the capillaries prevented the measurement (Appendix B Supplementary Figure S3). Subsequently, using PFG-NMR, we determined that the diffusion coefficient of the H19Y mutant at 35 °C was $1.86 \pm 0.09 \times 10^{-10}$ m^2/s. This value closely matches the theoretically predicted diffusion coefficient of the H19Y monomer calculated using the

HullRad server [163] from the H19Y structure (1.9 x 10^{-10} m^2/s for a monomer vs. 1.3 x 10^{-10} m^2/s for a dimer).

Collectively, our results show that H19Y mutation stabilizes the structure of the PAI subdomain, and H19Y and PAI adopt a similar three-helical fold. Therefore, we next utilized the H19Y mutation to gain relevant insights into the DNA-binding mechanism of the SB transposase.

3.3.5 Transposon DNA binding mechanism of the primary DNA-recognition subdomain of the Sleeping Beauty transposase

By promoting the folding of the PAI subdomain, H19Y replacement favors the formation of protein-DNA complexes. Previously, the presence of NMR signal broadening, resulting from the intermediate exchange on the chemical shift time scale between folded-unfolded and unbound-DNA-bound states of the PAI subdomain, prevented a detailed structural analysis of the PAI-DNA complex [4]. Since H19Y stabilizes the three-dimensional structure of the PAI subdomain, required for transposon DNA binding [11], we performed NMR DNA-binding experiments using H19Y in combination with the 18 bp DR-core sequence (Figure 3.2A), which represents the minimal DNA sequence necessary for transposase binding [130]. We initiated the experiments at 5 °C, since the H19Y mutant exhibited the highest structural stability and yielded the best quality [^{1}H,^{15}N]-HSQC spectrum at this temperature. However, the amide cross peaks were either missing completely or strongly broadened. As the temperature increased from 5 to 35 °C, the intensity of the peaks gradually increased. Remarkably, at 35 °C, we obtained the [^{1}H,^{15}N]-HSQC spectrum of the ^{15}N-labeled H19Y PAI subdomain, which displayed all signals, albeit still somewhat broadened (red cross peaks in Figure 3.5A).

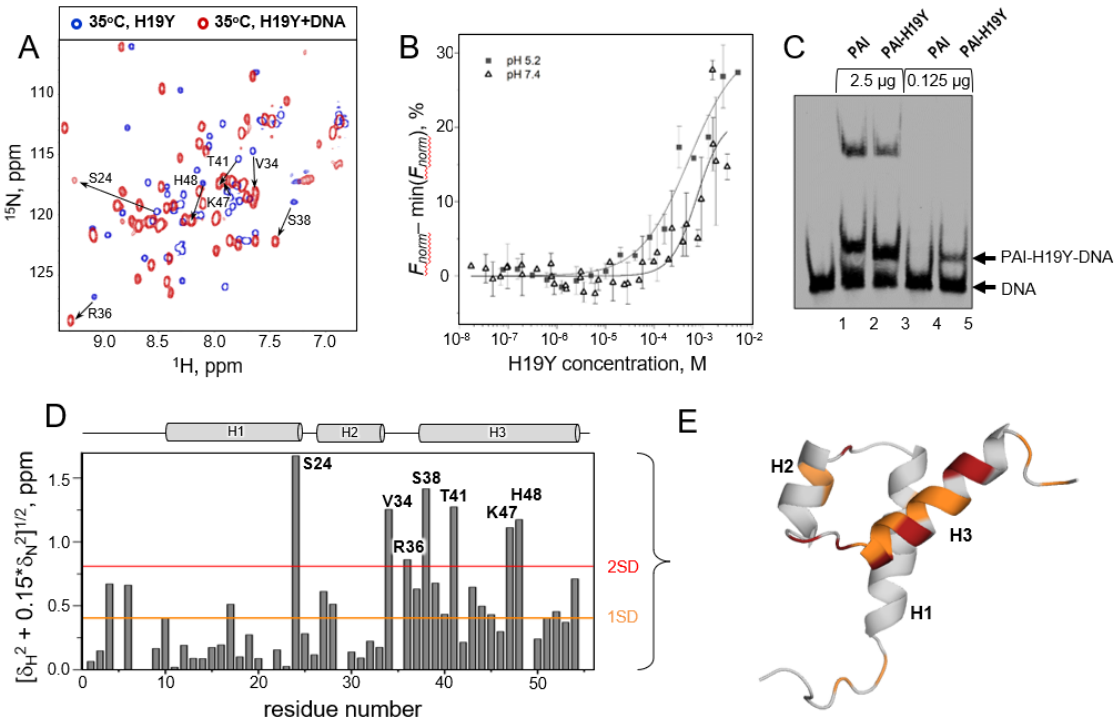

Figure 3.5 H19Y binding to the transposon DR-core sequence. (A) [^{1}H,^{15}N]-HSQC spectra of 0.085 mM ^{15}N,^{13}C-labeled H19Y are shown in the absence (*blue cross-peaks*) and presence (*red cross-peaks*) of DR-core (1:4.5 molar ratio) collected at 35 °C in an aqueous solution of 25 mM sodium-phosphate buffer at pH 5.2. Arrows exemplify chemical shift changes caused by the addition of DNA. (B) MST binding curbes for H19Y to DR-core collected at pH values of 5.2 or 7.4. Time region for shown data corresponded to the 5.0 s data collection interval. Experimental error bars show S.E. for n ≥ 3 separate experiments. The solid lines represent Hill fits to the experimental data. (C) Electrophoretic mobility shift analysis (EMSA) of the PAI subdomain and PAI-H19Y mutant produced in E. coli. Proteins were incubated with a biotinylated, double-stranded DNA oligonucleotide representing the Lo of the SB transposase. Lane 1: no protein; lane 2: SB10 2.5 µg; lane 3: SB10-H19Y 2.5 µg; lane 4: SB10 0.125 µg; lane 5: SB-H19Y 0.125 µg. D**)** ^{1}H and ^{15}N chemical shift differences of H19Y NMR signals from A due to DR-core binding, weighted according to $((\Delta\delta(^{1}H))^2 + 0.15(\Delta\delta(^{15}N))^2)^{1/2}$**.** Orange and red lines represent one and two standard deviations for the data, respectively. (E) ^{1}H and ^{15}N chemical shift differences are colored orange and red according to the magnitude of change (above one or two standard deviations, respectively) on the H19Y three-dimensional structure.

These data suggest that by shifting the folding-unfolding equilibrium towards a more structured conformation, the structure-stabilizing H19Y mutation also shifts the equilibrium between different species in solution towards the formation of protein-DNA complexes in line with the PAI-DNA binding via conformational selection [11]. The presence of all peaks in the [^{1}H,^{15}N]-HSQC spectrum corresponds to a DNA-bound conformation of H19Y. Signal broadening likely indicates that the protein-DNA interaction occurs with intermediate-to-fast exchange on the NMR chemical shift time scale as previously observed for the PAI subdomain [4] and can also be partially attributed to the reduced T2 relaxation time of H19Y due to the formation of the H19Y-

DR-core complex. Additionally, considering that DNA transposition takes place within a nucleoprotein complex involving at least two transposase enzymes [112, 164], at this point, we cannot completely exclude the possibility of dimerization or oligomerization of H19Y-DR-core complexes leading to signal broadening.

Using MST titration experiment, we quantitatively assessed the affinity of the H19Y binding to DR-core (Figure 3.5B). The concentration of Cy5-labeled DR-core sequence was constant and the concentration of unlabeled H19Y gradually increased. By detecting the fluorescence signal of the DR-core, we directly observed H19Y binding to DNA, irrespective of its folding-unfolding dynamics that would complicate the analysis of binding by NMR. KD values reveal a sub millimolar binding affinity of H19Y to DR-core, specifically 0.51 ± 0.29 mM at pH 5.2 and a weaker binding affinity 0.74 ± 0.26 mM at pH 7.4. Regarding the intermediate-to-fast exchange regime, these KD values align with the NMR data presented in Figure 3.5A.

3.3.6 Structural model of H19Y-DR-core complex.

Despite the overall similarity of peak distribution between the [^{1}H,^{15}N]-HSQC spectra of H19Y alone and in complex with DNA, there are substantial changes in chemical shifts for many peaks. To interpret the changes induced by DNA binding unambiguously, we repeated NMR resonance assignments for the ^{15}N^{13}C-labeled H19Y in the presence of DNA. The calculated weighted chemical shift perturbations of backbone amide resonances in H19Y upon addition of the DR-core sequence are depicted as a bar graph (Figure 3.5D) and mapped onto the H19Y structure (Figure 3.5E). Note that the observed differences in the behavior of NMR signal broadening at different temperatures for the DNA-free and DNA-bound [^{1}H,^{15}N]-HSQC spectra of H19Y necessitates the comparison of spectra collected at temperatures of 5 and 35 °C, with the rationale of comparing two stable and distinct structural states of the H19Y. There could be two reasons for such behavior.

First, the exchange regime between DNA-bound and unbound states likely shifts from fast-to-intermediate at 35 °C to intermediate as the temperature decreases, leading to a significant signal loss due to broadening at 5 °C. In addition, at 5 °C, large protein-DNA complexes may form, additionally decreasing signal due to the shorter transverse relaxation time. Notably, the backbone amide resonances exhibiting significant chemical shift perturbations, exceeding one (orange) or two (red, labelled in Figures 5D and 5E) standard deviations, are primarily localized to helix H3 and the loop connecting helices H2 and H3, which verifies that helix H3 serves as the DNA-recognition helix of the PAI subdomain.

To gain atomic-level insights into DNA recognition by the PAI subdomain, we employed a molecular docking approach to generate structural models of the H19Y-DR-core complex. We did not observe any intermolecular NOEs for the H19Y-DR-core complex, likely due to the fast-to-intermediate exchange regime observed in NMR experiments. Therefore, we utilized the chemical shift data to construct a structural model of the H19Y-DR-core complex using the HADDOCK program [153]. Figure 3.6 shows the predicted top H19Y-DR-core complex, where the third helix H3 interacts with the DNA major groove and helix H2 forms contacts with the DNA minor groove. Helix H1 is positioned away from the DNA, allowing for potential protein-protein interactions. Residues Y19, S23-K30, A33-R36, Q40, T41, R44, K45, K47-H49, and T52-H55 (shown in red color in Figure 4) contribute to the interactions with the DNA molecule. Moreover, the sidechains

of Y19, Q40, R44, K45, K47, H48, H49, and H54 form hydrogen bonds with the DNA bases, represented by blue dotted lines.

We also constructed structural models for complexes formed by the PAI subdomain or PAI-K14RK33A mutant with the DR-core sequence. These models were generated using the PD-DOCK program shown to produce reliable data without experimental input [154, 155, 165]. Figure 3.6 displays the predicted top PAI-DR-core and PAI-K14RK33A-DR-core complexes. In both

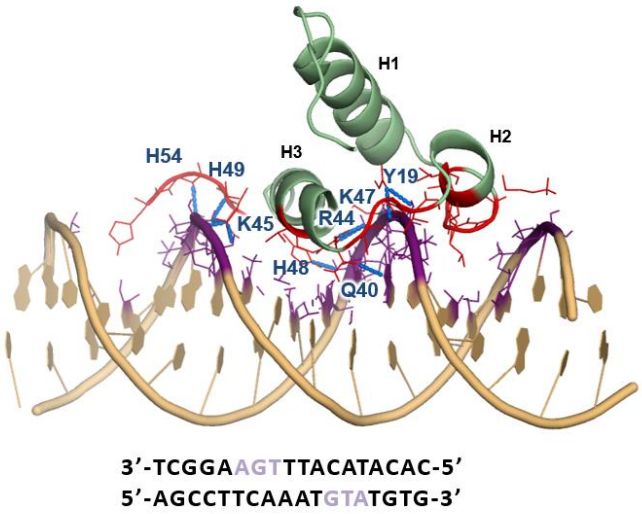

Figure 3.6 Structural model of H19Y in complex with DR-core constructed using the HADDOCK program. H19Y amino acid residues involved in contact with DR-core are labelled and all interface atoms within 5Å from the interacting partner are colored in red and purple for H19Y and DR-core, respectively. Dashed blue lines show hydrogen bonds formed between the protein and the DNA. The DR-core sequence is shown with base pairs involved in the interaction with H19Y colored blue.

cases, the third helix H3 interacts with the DNA major groove, whereas helix H1 is positioned away from the DNA. Residues involved in DNA interaction, shared between PAI and PAI-K14RK33A, are located in helix H3 (R36, S37, Q40, T41, R44, K45). In regards of helix H3 binding in the major groove, the DR-core binding mode of PAI and PAI-K14RK33A is the same as of H19Y, whose DR-core complex structure was based on experimental data, validating the prediction. The binding energy of the top predicted protein-DNA complex for PAI-K14RK33A is lower than for the PAI subdomain, indicating a more favorable protein-DNA interaction.

We also examined the binding of H19Y to Li and Lo binding sites (Appendix B Figure S5). As with DR-core, we obtained the [^{1}H,^{15}N]-HSQC spectra of the ^{15}N-labeled H19Y PAI subdomain also displaying all signals, corresponding to a DNA-bound conformation of H19Y. The spectra demonstrate changes in chemical shifts and varying degrees of signal broadening upon the addition of either the Li or Lo sequence. The addition of Li DNA sequence induces chemical shift changes distributed throughout the molecule, primarily localized to helices H2 and H3, while a few residues at the end of helix H1 are also affected. The addition of the Lo DNA sequence leads to significant chemical shift changes in a smaller subset of residues compared to Li, predominantly concentrated in helices H2 and H3, but accompanied by a more pronounced signal broadening. These observations resemble our previous results for a full-length DNA-binding domain of the SB transposase [4].

The observation of affected residues on helix H1 upon the addition of the Li sequence and the broadening of NMR signals in NMR H19Y-DNA binding experiments prompted us to investigate the potential dimerization or oligomerization of H19Y-DNA complexes. We performed PFG-NMR diffusion coefficient measurements for the H19Y-Li and H19Y-Lo mixtures. At high DNA concentrations, the protein signal was obscured by the DNA signal. For 1 mM Li or Lo addition, the diffusion coefficients were measured as $1.42 \pm 0.08 \times 10^{-10}$ m^2/s and $1.38 \pm 0.06 \times 10^{-10}$ m^2/s, respectively, supporting the formation of monomeric H19Y-Li or H19Y-Lo complexes [166]. By extension, these findings suggest that the H19Y-DR-core complex is also monomeric, and the observed NMR signal broadening arises from the exchange process between bound and unbound H19Y conformations, as well as equilibrium between folded and unfolded states. These data agree with the results of EMSA experiments showing that PAI subdomain binds to the transposon DNA as a monomer [167].

3.3.7 The structure stabilizing H19Y mutation in the PAI subdomain of the full-length transposase improves DNA-binding and results in its hyperactivity

Our data clearly show that the H19Y mutation improves DNA-binding properties of the PAI subdomain by increasing the number of molecules in the DNA-binding competent conformation. However, considering the multi-domain structure of SB transposase, other factors may affect its DNA binding capability, e.g., other domains contribution to DNA binding or interdomain interactions. Therefore, we investigated the impact of the H19Y-induced stabilization of the PAI subdomain's folded conformation on the DNA-binding of the full-length SB10 transposase using three independent techniques, MST, fluorescence anisotropy (FA), and fluorescence lifetime (FLT). Estimated binding constants are given in Table 3.2, and binding curves are provided in Appendix B Figures S6-S8. The three methods consistently indicate that the H19Y mutation enhances the binding of SB transposase to Li DNA sequence. The binding to Lo sequence was not significantly improved at 0.05 level, while the binding to the DR-core sequence was weakened insignificantly (MST) or significantly (FA and FLT).

To evaluate the impact of the H19Y replacement on the transposition activity of the SB transposase, relative transposition efficiencies were assessed by a colony-forming transposition assay, monitoring the integration of antibiotic resistance gene-containing SB transposons from donor plasmids into the chromosomes of transfected cells, leading to the generation of antibiotic-resistant cell clones. Human HeLa cells were co-transfected with a puromycin resistance gene (puro)-tagged transposon, alongside plasmids expressing either the SB10 or SB10-H19Y transposase. Analysis of the colony formation revealed a greater than 5-fold increase in the numbers of antibiotic-resistant cell colonies obtained with SB10-H19Y compared to SB10 (Figure

3.7 A). Based on these results, we conclude that the enhanced binding ability of SB10-H19Y to the transposon Lo sequence leads to an increase in transposition activity.

Table 3.2. KD values (nM) for full-length SB10, SB10-H19Y, SB100X, and SB100X-H19Y transposases binding to DR-core, Li, and Lo transposon DNA sequences determined by using MST, FA, and FLT.

Method	DNA sequence	SB10	SB10-H19Y	SB100X	SB100X-H19Y
MST	DR-Core	25.3 ± 0.9	31.1 ± 2.3	17.3 ± 1.1	31.2 ± 11.1
	Li	17.3 ± 1.2	10.2 ± 0.9	23.6 ± 2.2	10.3 ± 0.8
	Lo	16.6 ± 2.6	11.6 ± 1.8	23.0 ± 1.9	29.7 ± 3.5
FA	DR-Core	37.0 ± 4.0	51.3 ± 8.3	35.1 ± 6.6	104 ± 16
	Li	19.3 ± 2.8	18.2 ± 2.2	27.6 ± 4.9	16.0 ± 3.2
	Lo	23.9 ± 2.8	9.73 ± 2.5	13.7 ± 2.8	17.3 ± 2.2
FLT	DR-Core	29.4 ± 2.0	45.4 ± 4.8	31.6 ± 3.4	117 ± 6.7
	Li	17.2 ± 0.8	16.0 ± 0.7	25.9 ± 2.3	16.2 ± 2.3
	Lo	23.0 ± 2.2	13.2 ± 1.1	16.0 ± 2.6	20.6 ± 0.8

3.3.8 H19Y mutation reveals mechanistic differences between SB10 and SB100X transposases

We subsequently investigated the impact of the H19Y mutation on the DNA-binding properties of the SB100X hyperactive transposase (Figure 3.7B). Since SB100X showed a stronger tendency for nonspecific binding of DNA than SB10, we first tested the effect of H19Y mutation on nonspecific DNA binding using two different nonspecific DNA sequences: NS_1 (5'-ACCTTCCTCCGCAATACTCCCCCAGGT-3') DNA and NS_2 DNA (5'-CGGTCTTTCCGTCTT-3'). As shown in Figure 3.7B (cyan and purple vs. the black curve), there is a significant, about 6-fold decrease in the binding affinity of SB100X-H19Y transposase towards both the NS_1 and NS_2 DNA sequences. This indicates that the H19Y mutation enhances the specificity of SB100X transposase binding and underscores that the binding to the transposon

DNA follows the conformational selection mechanism at the level of the PAI subdomain. The determined KD values are 2.71 ± 0.2 μM and 3.43 ± 0.2 μM for SB100X-H19Y binding to NS_1

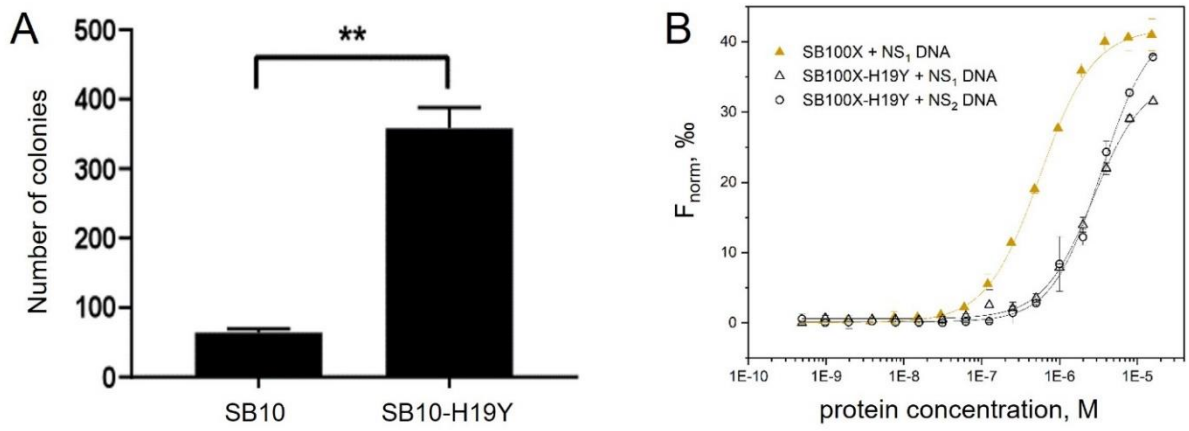

Figure 3.7 (A). A greater than 5-fold increase in the numbers of antibiotic-resistant cell colonies obtained with SB10-H19Y compared to SB10. (B) A 6-fold decrease in the binding affinity of SB100X-H19Y transposase towards both the NS1 and NS2 DNA sequences.

DNA and NS_2 DNA, respectively, as compared to 0.53 ± 0.02 μM for SB100X binding to NS_1 DNA.

We then investigated the effect of H19Y mutation on the SB100X transposase binding to the transposon DNA sequences DR-core, Li, and Lo (Table 3.2 and Appendix B). SB100X-H19Y exhibited stronger binding to Li sequence than SB100X. However, the binding to the DR-core and Lo sequences became weaker.

We have also evaluated the transposition efficiency of SB100X-H19Y transposase using a colony-forming transposition assay, however, we did not see an increase in the transposition activity.

3.4 DISCUSSION

In this study, we investigated the relationship between the folding, DNA binding, and transposition activity of the SB transposase. Previously, we demonstrated that the PAI subdomain of the SB transposase DNA-binding domain must adopt a folded conformation prior to binding the transposon DNA [11]. This suggested a conformational selection mechanism for PAI-DNA binding, e.g., only the PAI-subdomain in the pre-existing binding-competent conformation would bind to the DNA. However, the PAI subdomain adopts a stable folded conformation only under non-physiological environmental conditions, such as low temperature (5 ^{o}C), high salt concentration (more than 600 mM NaCl), and pH values above 7.0 [4], limiting the number of transposase molecules with its PAI subdomain in the binding-competent conformation at physiological conditions. Accordingly, it is reasonable to assume that by stabilizing the structure of the PAI subdomain under physiological conditions, where the presence of the transposase is a limiting factor of the transposition reaction, a stable structure-enabling DNA binding would promote the entire process of transposition.

The PAI subdomain carries a net overall positive charge, which suggests that the folded conformation may be electrostatically unfavorable and destabilized due to repulsion among positively charged amino acid residues. To replicate the folded conformation of the PAI subdomain, we sought to identify specific amino acid substitution(s) that would enhance its structural stability. The presence of two hyperactive mutations, K14R and K33A, located within the PAI subdomain [124], did not affect the folding properties of the PAI subdomain. However, our data revealed that the midpoint of pH-induced transition to the folded state of the PAI subdomain occurred at pH values around 6.0, at which the protonation state of the imidazole sidechain of histidine changed from cationic to neutral [157, 158]. Based on this finding, we

surveyed several histidine substitutions and found that H19Y had a stabilizing effect on the structure of the PAI subdomain, and the effect was enhanced in the presence of K14R and K33A hyperactive mutations. We generated a triple mutant of the PAI subdomain with K14R, H19Y, and K33A mutations, named H19Y, and experimentally confirmed its structural stability compared to the PAI subdomain. Consistent with enhanced structural stability, the H19Y mutation enables more robust formation of DNA-protein complexes *in vitro* (Figure 3.5C), suggesting that a higher fraction of the protein is in the conformation compatible with engaging transposon DNA.

The three-dimensional structures formed by the PAI subdomain [4] and H19Y are similar. Consequently, H19Y mutation can serve as a tool to investigate the DNA-binding properties of the PAI subdomain or of the full-length SB transposase. Indeed, in contrast to PAI, H19Y produced tractable NMR [^{1}H,^{15}N]-HSQC spectra in the presence of DR-core sequence with all signals present, enabling the generation of experimental constraints for building the structural model of the H19Y-DR-core complex. The structure of the complex showed that helices H3 and H2 and the turn connecting these helices participate in the binding to the transposon DNA with helix H3 being the primary DNA-recognition helix, substantiating prior predictions [4, 123, 132]. Furthermore, the contribution of L25 and R36 to DNA binding was previously observed, further verifying our model [133].

By utilizing H19Y, we assessed the DNA-binding affinity of the PAI subdomain. Surprisingly, our data revealed a rather weak binding affinity between the PAI subdomain and the DR-core DNA sequence with KD values within the sub-millimolar range. The experiments using longer DNA sequences, both Li and Lo, suggest a somewhat stronger binding affinity as the observed NMR signal broadening is increased, shifting the exchange regime from fast towards intermediate. However, full-length transposase is necessary for stronger DNA binding. Indeed, our data show

nanomolar DNA-binding affinity for both SB10 (Figure 3.2, Table 3.1) and SB100X (Figure 3.9), likely because a modular organization of the SB transposase allows for the formation of stable and structurally optimal nucleoprotein complex, needed for successful transposition.

At the protein concentrations used and a temperature of 35 ºC the observed exchange broadening of H19Y NMR signals in the presence of DNA does not represent dimerization or higher-order oligomerization of the H19Y-DNA complexes but stems from the exchange between DNA-bound and unbound H19Y conformations, as well as between alternative conformational H19Y states. The latter includes the ensemble of conformations sampled by the DNA-free H19Y and the conformational adjustment in H19Y induced by DNA binding. Indeed, the PAI subdomain contains a cluster of positively charged residues (K30, R36, R44, K45, K47-H49, and H55) at the DNA interface. Upon binding to DNA, the interactions with the negatively charged DNA molecule contribute to stabilizing the structure of the PAI subdomain, particularly the relative orientation of its three helices compared to the free state. This is supported by the reduction in NMR signal broadening, which mainly occurs in the loops connecting the helices in the absence of DNA (compare Figure 3.4C and 3.5A, red cross-peaks). The stabilization of α-helices coupled to complex formation was previously observed for example in helix H3 of the Pax5 protein [168] that shares structural and amino acid sequence similarity to the PAI subdomain [169]. The observed conformational selection mechanism and the conformational plasticity of the PAI subdomain do not contradict the established fact that the PAI subdomain plays a more dominant role in transposon DNA recognition [4, 132]. In fact, a positive correlation between larger conformational changes upon DNA-binding of transcription factors and increased DNA-binding specificity was previously found [170]. The folded conformation of H19Y facilitates the initial

DNA binding of the SB transposase. Therefore, we observe more DNA-bound complexes in the case of the H19Y mutant.

The comparison of SB10 and SB10-H19Y full-length transposases binding to DR-core, Li, and Lo sequences showed that H19Y mutation increased its DNA-binding affinity to Li sequence where the transposon DNA is excised, confirming our predictions that structural stability of the PAI subdomain enhances DNA-binding properties of SB transposase. Furthermore, we find that SB transposase carrying the H19Y mutation is hyperactive as measured by the efficiency of transposon integration into the genome (Figure 3.8C), suggesting that protein folding properties are major determinants of transposase activity.

While investigating full-length SB transposase binding to the transposon DNA binding sites DR-core, Li, and Lo, we discovered several important transposase properties providing novel mechanistic insights regarding the initial step of SB transposition. First, we observed the bimodal DNA binding of both SB10 and SB100X transposases, with the second binding mode corresponding to non-specific DNA binding. Non-specific binding can be expected based on the high positive charge of SB transposase – it has 50 (SB10) and 46 (SB100X) positively charged residues in excess over negatively charged residues. However, it is surprising that it occurs with the low micromolar affinity, while the affinity of the isolated PAI-subdomain is in sub-millimolar range. It is possible that the relatively weak DNA binding affinity of the primary DNA-binding domain of SB transposase is an indication of a cooperation among multiple transposase domains for enhanced DNA binding affinity, as well as between transposase molecules (via protein-protein interactions) required to form and maintain a stable transpositional complex during the multi-step reaction of transposition. It could be that the PAI subdomain facilitates precise identification of and latching onto specific binding sites on the transposon DNA by sampling different DNA sites

until the most favorable interaction of the entire transposase molecule with the transposon DNA is established.

It seems plausible that non-specific DNA binding of SB transposase with low micromolar affinity contributes to the decrease in transpositional activity when the transposase concentration increases [171]. For the SB transposon, the optimal ratio of the transposase-expressing plasmid to the transposon donor plasmid is a 1:10 to 1:25-fold [131, 172-174]. Our data (Figures 3.2B-C and 3.9A) show that the start of the nonspecific DNA-binding approximately corresponds to greater than 1:20 DNA:transposase molar ratios. Thus, increasing the number of transposase causes it to bind DNA randomly, still leaving the number of specific transposase:transposon DNA complexes limited for carrying out the productive transposition reaction. This observation is in line with the hypothesis that highly active (or hyperactive) transposases are counter-selected in nature because efficient transposition is expected to exert a negative impact on overall organismal fitness, and the underlying mechanism in the case of SB transposase could be its non-specific DNA binding activity caused by its high positive charge.

Low micromolar affinity of nonspecific DNA binding of SB transposase is compatible with its close-to-random integration into genome [175, 176]. This observation is also compatible with the tethering mechanism of SB transposase activity that involves the interaction of the SB transpositional nucleoprotein complex with chromatin-bound excess transposase molecules [177]. While we observed a significant increase in DNA-binding specificity of SB100X transposase with the H19Y structurally stabilized PAI subdomain to the Li DNA sequence, we observed weakened affinity to the DR-core and Lo DNA sequences. We also did not observe an increase in the transposition activity as with SB10 transposase. It is possible that the effect is masked by an already high, 100-fold increase in the transposition activity of SB100X transposase via other

mechanisms, including the increased DNA-binding affinity due to the K33A mutation or the need for stronger binding to the Lo sequences [133]. This observation signifies the dependence of SB transposition on several factors, with DNA-binding being only one of them, and explains the effectiveness of molecular evolution approach, which optimizes function by simultaneously screening mutations with different underlying molecular mechanisms, to design hyperactive transposase [124]. Nonetheless, in the case of SB100X transposase, the H19Y mutation significantly improved its DNA-binding specificity (Figure 9A), which agrees with the idea of conformational selection specific transposon DNA binding.

Furthermore, introducing the H19Y mutation enabled a mechanistic insight into the nucleoprotein complex formation by SB100X transposase. Our data show that SB100X transposase binds as a dimer or forms a dimer when bound to DR-core and Lo transposon DNA-sequence, at which the cleavage of the transposon occurs, but to Li (Figure 3.9B-C). In the context of the ordered transpososome assembly, our data agree with the model proposed by Ochmann and Ivics [164]. SB100X transposase, bound as a monomer to the inner and as a dimer to the outer binding sites on the transposon TIRs, engages in protein-protein interactions to form a transpososome, in part via the interaction of the PAI subdomain's helix H1. The formation of SB100X dimer at the outer binding site is thought to be mediated by the RED subdomain [167]. For SB10 transposase, dimerization upon DNA binding was not observed for Li and Lo binding sites. We hypothesize that dimerization might occur at higher concentrations of transposase, but this is obscured by the second DNA binding mode in our experiments. It is possible that the formation of a dimer by SB100X at lower concentrations in comparison to SB10 correlates with higher transposition activity of SB100X.

To the best of our knowledge, we present the first quantitative estimation of the DNA-binding constants for the SB transposase. KD values determined for the full-length SB transposase are comparable to those reported for another mariner transposase SETMAR, 85 nM determined by EMSA titration [178] and 53 ± 4 nM by fluorescence anisotropy assays [179]. We anticipate that the insight into the DNA-binding properties of SB10 and SB100X transposases provided by our data may facilitate the search for optimal experimental conditions for crystallographic or cryo-EM structural studies of the SB nucleoprotein complex.

The property of being disordered in the DNA-unbound form and folding upon DNA binding is commonly observed in protein-nucleic acid interactions. In the absence of a preformed interface, a protein can rapidly scan the surfaces of the DNA target and adopt a folded conformation upon binding to a specific DNA sequence [180, 181]. In the case of SB transposase, its PAI subdomain binds the transposon DNA via a pre-folded conformation. This suggests that, in the rational engineering of efficient transposases, modifications to the PAI subdomain should follow the general principle of increasing its structural stability. We also propose that increasing the selectivity and specificity of DNA binding for targeted gene delivery can be applied to the SB transposase, despite first direct fusions of target-specific DNA-binding domains to SB transposase diminished the SB transposition activity [182, 183]. To achieve this, the selection of the target DNA sequence should begin with the DNA site cleaved by the catalytic domain and ensuring an optimal spacing to the site where the engineered PAI subdomain variant would bind specifically. This parallels earlier observations that highlight the paramount importance of selecting an appropriate target site for successful SB transposition [184].

CHAPTER 4: FRET-BASED EXPERIMENTAL VERIFICATION OF THE COMPUTATIONALLY PREDICTED STRUCTURAL MODEL OF THE SLEEPING BEAUTY PAIRED-END COMPLEX

4.1 INTRODUCTION

DNA transposons are mobile DNA elements that can move (transpose) from one location within a genome to another [113]. Due to this ability, DNA transposons can be used as DNA transfer vehicles in applications that require a coordinated delivery of genetic information, such as functional genomics and genetic engineering/gene therapy [115]. Several DNA transposon systems are currently being developed and used in basic, preclinical, and clinical research, including Sleeping Beauty (SB), piggyBac (PB), and Tol2 [185]. SB transposon, due to its synthetic origin and subsequent developments [121, 123, 124, 130, 135], is the most active DNA transposon in vertebrate animal cells. It has been extensively characterized and extensively used to generate transgenic cell lines, induce pluripotent stem cell reprogramming, and in insertional mutagenesis screens [111, 118, 186, 187].

There are several advantages to using DNA transposons for genetic applications. First, DNA transposition ensures that the mobilized genetic sequence is precisely defined, and relatively large DNA sequences intended for delivery can be incorporated into the transposon [125]. Second, in contrast to viruses that often have highly immunogenic protein coats, transposons are composed only of DNA, thus avoiding immune and other defense mechanisms that cells use to prevent the integration of foreign DNA [188, 189]. Furthermore, for human genetic applications, the integration site preference of the vehicle used to deliver the new genetic information must be considered. SB transposon inserts at the TA dinucleotides, exhibiting a close-to-random genome-

wide profile [190, 191]. In this regard, it surpasses PB and Tol2 transposons as well as viral vectors, which have a higher probability of disrupting transcriptional and regulatory genes and potentially can disrupt tumor suppressor genes or activate oncogenes [192, 193]. At the same time, a non-zero risk associated with the low specificity of integration remains [193, 194]. Therefore, a molecular-level understanding of the interactions between the transposase enzyme and the transposon DNA during the reaction of DNA transposition would aid in designing better DNA transposons with controlled insertion site selection, thus avoiding genotoxic effects. However, the structural basis of the SB transposition and the mechanism of site selection by SB transposon remains unknown due to the lack of high-resolution structural information on SB transposase interaction with the transposon DNA.

In the case of SB transposase, experimental structures have been individually solved only for the PAI and RED subdomains by NMR spectroscopy [4, 195] and for the catalytic domain by x-ray crystallography [134]. None of these subdomain/domain structures have been solved in complex with DNA, but they were used to model the SB transposase/transposon end/target (TCC) DNA complex [134, 196] by homology to the Mos1 PEC complex [161]. The SB TCC model shows the two transposon DNA ends arranged in parallel and held together by two transposase enzymes. The SB transposase enzymes bind the transposon DNA in trans, where the DNA-binding domain of one SB transposase enzyme binds to one transposon DNA end whereas the catalytic domain of the same transposase enzyme binds to (and cleaves) the other transposon DNA end. The SB TCC model was instrumental in generating hypotheses for rational improvements of the SB transposase [121, 135]. However, it is important to note that Mos1, showing very little amino acid sequence similarity to SB transposase, is the only transposase from Tc1/*mariner* family, for which the experimental structure of the full-length protein was solved. Thus, given a limited number of

protein templates for structural modeling, an experimental verification of the structural model for the full-length SB transposase-transposon DNA complex is still required.

Förster Resonance Energy Transfer (FRET) experiments can provide site-specific distance information for restrained structural modeling of complex biomolecular assemblies [197]. In this work, we predicted the initial structural model of the SB transposase complex with Lo DNA using AlphaFold 3. We used distance restraints gathered from FRET experiments to verify the predicted structural model of the SB transposase-transposon DNA complex.

4.2 MATERIALS AND METHODS

4.2.1 Protein expression and purification

DNA plasmids encoding SB transposase cysteine mutants were ordered from GenScript USA Inc. The proteins were expressed in BL21-AI *E. coli* cells (Novagen). The nomenclature of SB transposase mutants and DNA constructs used in this study is given in Appendix C. An overnight bacterial culture (5 ml) was added to 0.5 L of LB medium supplemented with 100 μg/mL ampicillin and grown at 30 °C. When the culture reached the OD_{600} of 0.6-0.8, the protein expression was induced for 4 hours by the addition of IPTG (isopropyl β-D-1-thiogalactopyranoside) and L-arabinose to the final concentration of 1 mM and 0.2 % w/v, respectively. Cells were harvested by centrifugation at 5,000 g for 20 min at 4ºC. Cell pellet was resuspended in the 50 mM Tris-HCl buffer prepared at pH 7.5 and containing 5 mM $MgCl_2$, 5 mM $CaCl_2$, lysozyme at the concentration of 5 μg/mL, and 200 μM PMSF (1 g of pellet per 15 mL of buffer) and incubated for 45 min. with agitation at room temperature. Subsequently, DNase I was added at 30 units per 1 g of cell pellet, and the suspension was incubated for an additional 45 min. with agitation at room temperature. The cell lysate was centrifuged at 30,000 g for 45 min. at 4°C and the pellet containing inclusion bodies was collected. The inclusion bodies were sequentially washed once with 1 % Tween-20 in buffer A (50 mM Tris-HCl, 0.6 M NaCl, 5 mM $MgCl_2$, pH 7.5) to remove membrane proteins, twice in buffer A containing 3 M NaCl to remove DNA, and one more time in buffer A. After washing, inclusion bodies were solubilized in buffer B (50 mM Tris-HCl, 0.6 M NaCl, 5 mM $MgCl_2$, 5 mM imidazole, 2 % glycerol, 6 M GuHCl, 1 mM TCEP, pH 7.5) by stirring overnight at 4 °C. The solubilized protein was sonicated on ice and centrifuged at 85,000 g for 25 min. The supernatant was incubated at room temperature for 20 min. with Ni-Penta resin (Marvelgent Biosciences Inc., 0.5 mL of resin per 1 g of cell pellet) primed with 5 mM

imidazole, loaded onto a gravity column, and washed with 5 mM imidazole in buffer B until no absorbance signal at 280 nm was detected. SB transposase mutants were eluted with 100 mM imidazole in buffer C (50 mM Tris, 0.6 M NaCl, 5 mM $MgCl_2$, 2.5% glycerol, 6 M urea, pH 7.5) and refolded by stepwise dialysis out of urea using buffer D (50 mM Tris-HCl, 300 mM NaCl, 5 mM $MgCl_2$, 2% glycerol, 1 mM TCEP, 50 mM 1:1 arginine-glutamate mixture, pH 7.5). The steps involved decreasing the concentration of urea from 6 M to 4 M, 2 M, and 0 M, using dialysis in 1:50 volume ratio at each step. The first two dialysis steps were performed for two hours each at room temperature, the final dialysis step was performed overnight at 4^{o}C. The samples were stored at 4^{o}C, and all experiments were performed within 96 hours of the final refolding step.

4.2.2 Protein and DNA labeling

The Lo DNA was purchased from IDT (Integrated *DNA* Technologies, Inc.) labeled at the 3' end. Cy5 (sulfo-Cyanine5) and TMR (tetramethylrhodamine) maleimide fluorophores (Lumiprobe) were prepared as stock solutions in DMSO at concentrations 10 mM and 7.2 mM, respectively. SB transposase cysteine mutants were labeled with Cy5 or TMR fluorophores by incubating the dye with the protein at a 2:1 dye:protein molar ratio at 4°C overnight, in the presence of TCEP (tris(2-carboxyethyl)phosphine), which was added at a 2:1 TCEP:protein molar ratio for optimal results. Excess dye was removed using PD-10 desalting columns, and the final buffer contained 50 mM Tris, 300 mM NaCl, 5 mM MgCl2, 2% glycerol, and a 50 mM arginine-glutamate mixture. Protein and DNA concentrations and concentrations of fluorescent probes were measured using UV-Vis spectrophotometer Evolution 60S. The degree of labeling was estimated from the light absorption measurements.

The degree of labeling *d* was determined using the following formula:

$$Degree\ of\ Labeling = \frac{A_{dye}^{max}}{\varepsilon_{dye}^{max} \times C} \times 100\%, \quad (1)$$

where A_{dye}^{max} is the absorbance of labeled protein measured at the absorbance maximum of the dye, ε_{dye}^{max} is the extinction coefficient of the dye, and C is the molar concentration of the labeled protein. C was calculated using the following formula:

$$C = \frac{A_{280} - \left(A_{dye}^{max} \times CF_{dye,280}\right)}{\varepsilon_{protein}}, \quad (2)$$

where A_{280} is the absorbance of the labeled protein measured at 280 nM, $CF_{dye,280}$ is the correction factor of dye (to account for the absorbance contribution of the dye to the measured protein absorption) at 280 nM, and $\varepsilon_{protein}$ is the extinction coefficient of the protein (61000 $M^{-1}cm^{-1}$). To calculate the degree of labeling for DNA, A_{280}, $\varepsilon_{protein}$, and $CF_{dye,280}$ were substituted with A_{260}, ε_{DNA}, and $CF_{dye,260}$, respectively. The labeling efficiencies for DNA and proteins ranged from 40% to 70%.

4.2.3 MST experiments

MST experiments were performed using a Monolith NT.115 (NanoTemper) instrument. For protein-DNA binding experiments, we used Cy5-labeled Lo sequences. The experiments were done using Monolith NT.115 premium capillaries. The assay buffer contained 50 mM TRIS-HCl at pH 7.5, 5 mM $MgCl_2$, 300 mM NaCl, 2 % glycerol, 50 mM Arg-Glu mixture, 1 mM TCEP, 0.1% Triton X-100, and 0.1 mg/mL BSA to prevent sample sticking to the capillaries. Initially, resulting dose response (F_{norm}) curves obtained from normalized fluorescence were analyzed by least-squares curve fit using the Nanotemper software. Subsequently, the data from different experiment repeats (>3) were analyzed together and figures were created using Origin 22 software.

4.2.4 Steady state fluorescence and fluorescence lifetime measurements

Steady state fluorescence measurements were performed using a PTI QuantaMaster fluorometer (Horiba Scientific). Fluorescence of the donor-labeled protein or DNA (TMR or Cy3) was excited at 500nm and detected in the range of 550 to 750 nm. Fluorescence of the acceptor-labeled protein or DNA (Cy5) was excited at 570 nm and detected in the range of 620 to 750 nm. All experiments were carried out at room temperature (21 ± 1 ºC).

Fluorescence lifetime measurements were performed using a home-built time-resolved fluorimeter, equipped with a QuadraCentric sample compartment with a cuvette holder (Horiba Scientific), a passively Q-switched microchip YAG laser (SNV-20F-100, 532 nm, 20 kHz, Teem Photonics), a photomultiplier (H6779-20, Hamamatsu), and a digitizer (Acqiris DC252, Agilent). A 532/18 nm BrightLine single-band bandpass filter (Semrock) and a polarizer set at the magic angle were used in the detection arm. All experiments were carried out at room temperature (21 ± 1 ºC). The instrument response function (IRF) was obtained before each measurement using buffer as a scatterer and accounted for in the data analysis. The solution of the fluorescently labeled protein was added to the observation cuvette, and the fluorescence waveform was acquired by averaging fluorescence transients from one thousand laser pulses. All analyses of time-resolved fluorescence data were performed using the software package FargoFit, designed by I. Negrashov, which executes a global least-square fitting of multiple time-resolved fluorescence waveforms using different models that allow linking fitting parameters between waveforms. The obtained waveforms of donor fluorescence were best fitted by two exponential components, convoluted with the IRF.

4.2.5 FRET efficiency and donor-acceptor distance calculation

FRET efficiencies, *E*, were determined using two complementary approaches, by measuring the decreased fluorescence quantum yield of donor, and the decrease in donor fluorescence lifetime [198]. For all measurements, we used protein-DNA samples, in which one entity (protein or DNA) was donor-labeled, and the other (DNA or protein, respectively) was either unlabeled or labeled with acceptor. Controls included sample buffers and donor-labeled or acceptor-labeled entities alone.

To calculate FRET efficiency from decreased fluorescence quantum yield of donor, we used the following equation:

$$E = 1 - \frac{DA_{500}}{D_{500}} \qquad (4)$$

where DA_{500} and D_{500} are the integral emission intensities of donor in the mixture of donor-acceptor or donor-unlabeled entity respectively, excited at 500 nm. To obtain the pure donor fluorescence of the DA_{500} spectrum, we acquired fluorescence spectrum of acceptor, scaled it to the acceptor portion of the DA_{500} spectrum, and then subtracted the spectrum of acceptor.

To calculate FRET efficiency from the measurements of donor fluorescence lifetime, we used the following equation:

$$E = 1 - \frac{\tau_{DA}}{\tau_D}, \qquad (5)$$

where τ_{DA} is the donor fluorescence lifetime in the presence of acceptor and τ_D is the donor fluorescence lifetime in the presence of the unlabeled entity added to the same total concentration of protein and DNA in the sample as with donor and acceptor labeled entities.

The donor-acceptor distances were obtained from FRET efficiencies by using Förster's equation:

$$R = R_0\left(E^{-1} - 1\right)^{1/6}, \qquad (6)$$

where R_0, the Förster distance, was calculated according to the formula [148]:

$$R_0 = 9786\left[J(\lambda)k^2\eta^4 Q_D\right]^{1/6}, \qquad (7)$$

where λ is the wavelength, $J(\lambda)$ is the spectral overlap integral between the normalized donor emission spectrum $F_D(\lambda)$ and the acceptor absorption spectrum $\varepsilon_A(\lambda)$, k^2 = 2/3 is the probes orientation factor, η =1.4 is the refraction index of the medium, and Q_D is the quantum yield of donor-only labeled protein (Q_D = 0.11). Q_D was estimated by the comparison to the quantum yield of quinine sulfate in 0.05 M H_2SO_4 at λ_{ex} = 347.5 nm (Q_S=0.31 [199] [200]), according to the equation:

$$Q_D = Q_S \frac{F_D(\lambda)/A_D(\lambda)}{F_S(\lambda)/A_S(\lambda)}, \qquad (8)$$

where $F(\lambda)$ is the integral emission and $A(\lambda)$ is the absorbance at the excitation wavelength of donor-labeled protein or the standard. Value of R_0 = 53 Å for the TMR-Cy5 pair and 46.7 Å for the Cy3-Cy5 pair were obtained.

4.3 RESULTS

4.3.1 The selection of labeling sites on SB transposase and transposon DNA

The selection of labeling sites on SB transposase for gathering experimental distance constraints from FRET experiments was based on the following three criteria. First, the sites should be away from the catalytic site and should not involve functionally important residues, on which comprehensive functional data are available [196]. Second, the sites should be as distinct as possible from each other. Third, labeling site pairs should produce distances close to the Förster radius, where the sensitivity of the measurements is highest. To predict the initial structural model of the SB transposase complex with Lo DNA, we used AlphaFold 3 (AF3) [201]. The overall arrangement of the individual molecules in the protein-DNA complex in the AF3-predicted model (Figure 4.1) resembles the model of the SB-target DNA capture complex [134] as well as the structure of the paired-end complex of a closely related Mos1 transposase [161]. This suggests that our initial model provides a reasonable representation of the molecular interactions and spatial organization expected in the SB transposase-Lo DNA complex. Based on AF3-predicted model, T51C, T94C, R126C, and T295C residues were selected for cysteine substitutions. We also labeled

Lo DNA at both ends (Figure 4.1). We used Crystallography & NMR Systems (CNS)[202] to computationally attach labels to the predicted structure and modeled their average positions as pseudo atoms (PSDO) (Appendix C Figure S1). The distances between PSDOs in the AF3-predicted structure were compared to the distances determined experimentally using FRET.

4.3.2 The formation of paired-end SB complex

We used Microscale Thermophoresis (MST) and fluorescence lifetime (LT) experiments to

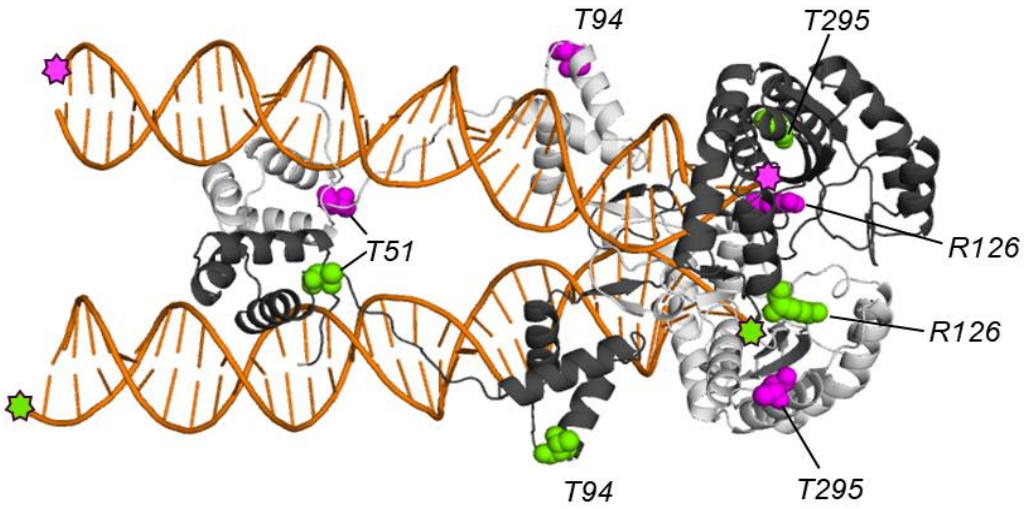

Figure 4.1 Positions of labeling sites on SB transposase and Lo DNA. The cartoon representation of the structural model of the SB transposase-transposon DNA paired-end complex, predicted using AF3, demonstrates the two transposase monomers bound to two transposon Lo DNA sequences. The two transposase monomers are colored in black and white, respectively. Residues T51, T94, R126, and T295, which were selected for labeling, are shown as spheres and are colored differently to indicate their locations on different transposase monomers. Stars indicate the positions of labels on the DNAs.

confirm that SB transposase mutants generated for FRET experiments bind to the transposon DNA and to evaluate their binding affinity. Figures 2A and 2B show binding curves for all mutants used in this study. All mutants demonstrated a strong, nanomolar affinity for Lo DNA, and the binding constants (*KDs*) estimated from7 MST and fluorescence lifetime experiments agreed with each other (Table 4.1).

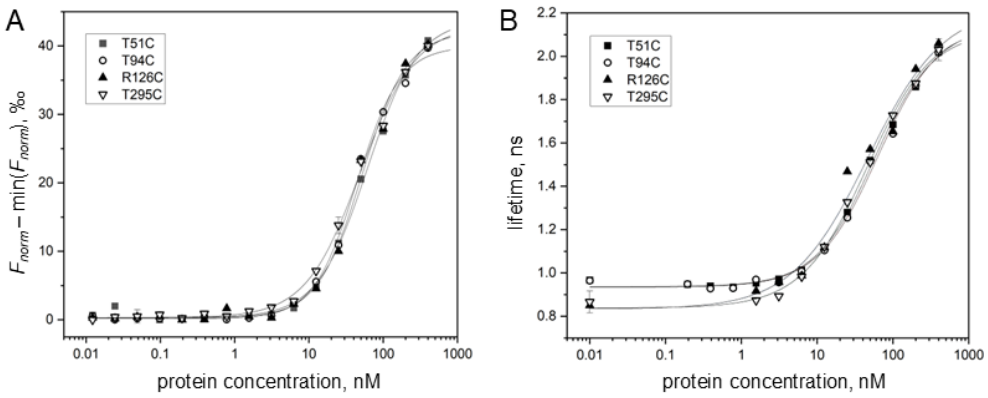

Figure 4.2 Binding affinity of SB transposase mutants to Lo DNA. (A) MST binding curves for SB transposase T51C, T94C, R126C, and T295C mutants binding to Lo DNA. The concentration of Cy5-Lo DNA was kept constant at 25nM and the protein concentration varied. To facilitate the comparison of binding curves for different mutants, we show data on the same scale. For this, we subtracted the respective minimum value of Fnorm for each curve. (B) LT binding curves obtained for similar samples with Cy3-Lo DNA at 25nM and varying protein concentrations.

Table 4.1. The binding affinity of single-cysteine SB transposase mutants to Lo DNA.

Mutant	*KD*, nM	
	MST	lifetime
T51C	61 ± 5	57 ± 10
T94C	44 ± 3	64 ± 12
R126C	52 ± 5	47 ± 10
T295C	50 ± 4	46 ± 8

Next, we verified the formation of the SB protein-DNA paired-end complex. For these experiments, we used two Lo DNA sequences, one labeled with Cy3 fluorophore acting as donor and another labeled with Cy5 fluorophore acting as acceptor. Labels were attached to the ends of Lo DNAs that enter the catalytic site of SB transposase. Lo-Cy3 and Lo-Cy5 were mixed at a 1:5 molar ratio, with acceptor-labeled DNA (Lo-Cy5) added in excess to increase the likelihood of donor-acceptor pairing in the paired-end complex, thereby enhancing the probability of observing the desired FRET effect. Without SB transposase, for an excitation at 500 nm, the fluorescence spectrum of Cy3-Lo/Cy5-Lo mixture showed only one strong emission peak at 566 nm,

corresponding to Cy3 fluorophore emission (Figure 4.3, black line), and no FRET signal was observed between Lo-Cy3 and Lo-Cy5 DNAs. When unlabeled protein was added to the Cy3-Lo/Cy5-Lo DNA mixture at a molar ratio 1:1, a strong FRET effect, i.e., the decrease in donor emission peak intensity and an increase in acceptor emission peak intensity at 670 nm was observed, indicating that the two DNA sequences were brought together (Figure 4.3, red curve). The distance between Lo-Cy3 and Lo-Cy5 fluorophores estimated from FRET efficiency was equal to 43 Å, which was consistent with the model predictions (42.7 Å), indicating the formation of the SB protein-DNA paired-end complex and validating our approach.

We further verified the SB protein-DNA paired-end complex formation using SB transposase labeled with TMR (donor) or Cy5 (acceptor) and unlabeled Lo DNA. In the absence of DNA, no

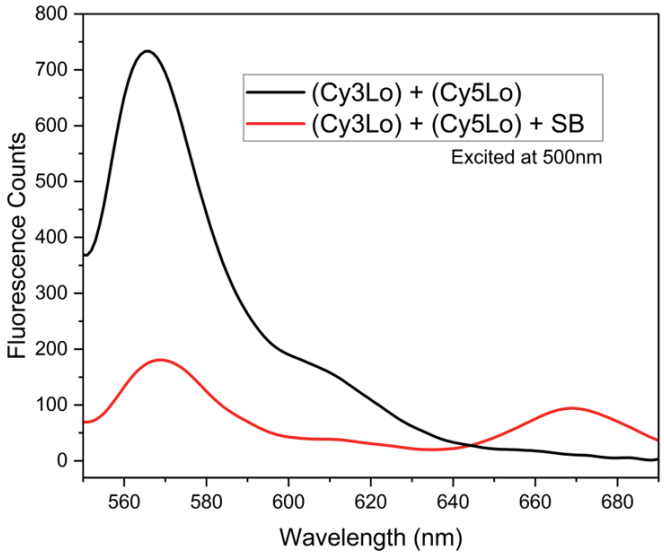

Figure 4.3 Strong FRET effect observed on adding unlabeled SB to a solution containing Cy3-Lo and Cy5-Lo, evidence of complex formation.

FRET signal was observed between SB-TMR and SB-Cy5 mixed at 1:1 molar ratio at protein concentration of 5 µM, indicating that even at micromolar concentrations SB transposase does not form dimers or higher order oligomers in the absence of DNA (Appendix C Figure S2). In contrast, a strong FRET effect was observed when unlabeled Lo DNA was added to SB-TMR/SB-Cy5

mixture, indicating the SB protein-DNA paired-end complex formation. This result is exemplified in Figure 4.4 for SB100-T295 mutant. The distance between labeled SB transposase evaluated from FRET efficiency was 54 Å, which was inconsistent with the distance determined with model predictions (65.1Å).

4.3.3 Distance mapping

The presence of energy transfer was readily seen either as a decrease in donor fluorescence or an increase in acceptor fluorescence (Figure 4.3 and Figure 4.4). To obtain distances between

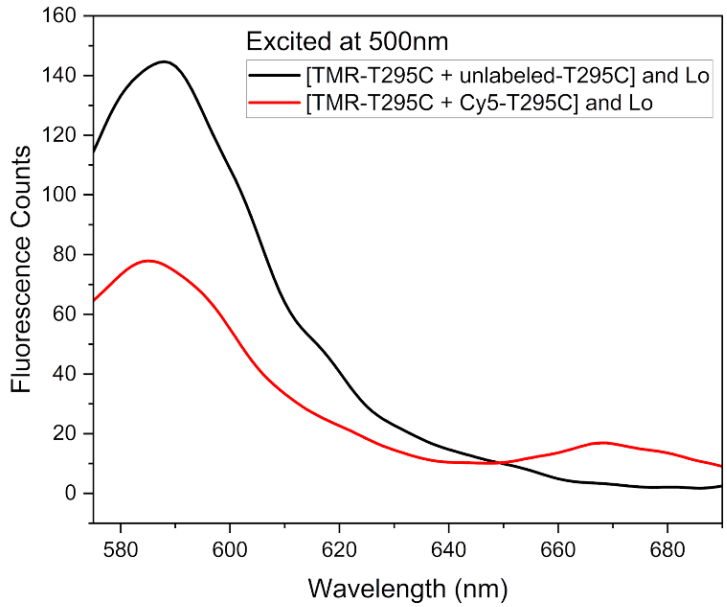

Figure 4.4 Strong FRET effect observed on adding unlabeled Lo to a solution containing TMR-T295 and Cy5-T295, evidence of complex formation.

fluorophores, we used the two complementary approaches described above. The calculated FRET efficiencies were in the 0.3–0.8 range enabling reliable detection of differences in the distance. Table 4.2 lists distances determined for all FRET pairs.

Table 4.2 The AF-predicted and FRET-determined distances in single-cysteine SB transposase complexed with Lo.

Fluorophore pair	**Distance, Å**		
Labeled protein	**Predicted**	**DQ**	**LT**
TMR-T51C and Cy5-T51C	44.0	60.4	
TMR-T94C and Cy5-T94C	83.0	60.8	
TMR-T94C and Cy5-R126C	73.6	48.1	46.0±25.97
TMR-R126C and Cy5-R126C	55.0	59.0	
TMR-R126C and Cy5-T295C	23.4	73.8	74.1±1.4
TMR-T295C and Cy5-T295C	65.1	54.0	
Labeled DNA			
Inside catalytic site (3'—3')	43.0	43.0	

DISCUSSION AND FUTURE DIRECTIONS

SB transposase has proven to be a challenging protein for structural studies. This is partly due to the disordered nature of its DNA-binding domain and partly due to the tendency of the full-length SB transposase to aggregate at the concentrations required for most biophysical experiments for structure determination. As the results, a divide-and-concur approach was utilized to gain structural insight into the structural properties of SB transposase. Solution structures of the two subdomains of the DNA-binding domain of SB transposase, PAI and RED, were obtained by NMR spectroscopy [4, 195] and the structure of the catalytic domain was obtained by x-ray crystallography [134]. Starting from the AF3-predicted structure, we took advantage of high sensitivity and, hence, small protein concentration requirement of fluorescence measurements to verify the solution-state of the full-length SB transposase dimer in complex with DNA representing the outer binding site on the left transposon end. Additionally, we incorporated the structure-stabilizing substitution H19Y in the PAI subdomain, which facilitated protein-DNA complex formation.

We found significant discrepancies between the inter-residue distances predicted by AF3 and measured experimentally using FRET. This indicates that the predicted structure of SB paired-end complex needs to be refined using these restraints. We propose an integrative computational modeling approach with FRET-derived restraints incorporated in the model to solve the structure of the SB paired-end complex.

These experiments to complete the distance mapping of the SB nucleoprotein complex are currently underway. We will use computational modeling to build the experimental FRET-based structure of the SB transpososome complex. Briefly: OpenMM (https://openmm.org/) will be utilized to perform a series of molecular dynamics simulations where the fluorophore molecules

(defined by a central carbon reference point) will be orientated to match the FRET measured distances using a slow driving force. Once the fluorophores are in an initial state which corresponds to the experimental measurements, a production run of a few 10s of nanosceonds will be performed where the dye-pair distances, observed over time, will be contrasted with the experimental values to give a structure of SB complex in agreement with the FRET experiments.

REFERENCES

1. Tanner, J.E., *Use of Stimulated Echo in Nmr-Diffusion Studies.* Journal of Chemical Physics, 1970. **52**(5): p. 2523-&.
2. Melnikova, D.L., et al., *On Complex Formation between 5-Fluorouracil and β-Cyclodextrin in Solution and in the Solid State: IR Markers and Detection of Short-Lived Complexes by Diffusion NMR.* Molecules, 2020. **25**(23).
3. Buchan, D.W.A. and D.T. Jones, *The PSIPRED Protein Analysis Workbench: 20 years on.* Nucleic Acids Res, 2019. **47**(W1): p. W402-W407.
4. Carpentier, C.E., et al., *NMR structural analysis of Sleeping Beauty transposase binding to DNA.* Protein Sci, 2014. **23**(1): p. 23-33.
5. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nat Rev Mol Cell Biol, 2005. **6**(3): p. 197-208.
6. Habchi, J., et al., *Introducing protein intrinsic disorder.* Chem Rev, 2014. **114**(13): p. 6561-88.
7. Mao, A.H., et al., *Net charge per residue modulates conformational ensembles of intrinsically disordered proteins.* Proc Natl Acad Sci U S A, 2010. **107**(18): p. 8183-8.
8. Uversky, V.N., *Unusual biophysics of intrinsically disordered proteins.* Biochim Biophys Acta, 2013. **1834**(5): p. 932-51.
9. Dehner, A. and H. Kessler, *Diffusion NMR spectroscopy: folding and aggregation of domains in p53.* Chembiochem, 2005. **6**(9): p. 1550-65.
10. Jones, J.A., et al., *Characterisation of protein unfolding by NMR diffusion measurements.* Journal of Biomolecular Nmr, 1997. **10**(2): p. 199-203.
11. Leighton, G.O., et al., *The folding of the specific DNA recognition subdomain of the sleeping beauty transposase is temperature-dependent and is required for its binding to the transposon DNA.* PLoS One, 2014. **9**(11): p. e112114.
12. Pan, H., G. Barany, and C. Woodward, *Reduced BPTI is collapsed. A pulsed field gradient NMR study of unfolded and partially folded bovine pancreatic trypsin inhibitor.* Protein Sci, 1997. **6**(9): p. 1985-92.
13. Penkett, C.J., et al., *Structural and dynamical characterization of a biologically active unfolded fibronectin-binding protein from Staphylococcus aureus.* Biochemistry, 1998. **37**(48): p. 17054-67.
14. Wang, Y., et al., *Disordered Protein Diffusion under Crowded Conditions.* J Phys Chem Lett, 2012. **3**(18): p. 2703-2706.
15. Schmidt, D.G. and P. Both, *LOCATION OF ALPHA-SI-CASEIN,BETA-CASEIN AND K-CASEIN IN ARTIFICIAL CASEIN MICELLES.* Milchwissenschaft-Milk Science International, 1982. **37**(6): p. 336-337.
16. Kunz, C. and B. Lonnerdal, *HUMAN-MILK PROTEINS - ANALYSIS OF CASEIN AND CASEIN SUBUNITS BY ANION-EXCHANGE CHROMATOGRAPHY, GEL-ELECTROPHORESIS, AND SPECIFIC STAINING METHODS.* American Journal of Clinical Nutrition, 1990. **51**(1): p. 37-46.
17. Redwan, E.M., et al., *Disorder in Milk Proteins: Caseins, Intrinsically Disordered Colloids.* Current Protein & Peptide Science, 2015. **16**(3): p. 228-242.
18. Dunker, A.K., et al., *Intrinsically disordered protein.* Journal of Molecular Graphics & Modelling, 2001. **19**(1): p. 26-59.
19. Uversky, V.N., *Intrinsically disordered proteins from A to Z.* International Journal of Biochemistry & Cell Biology, 2011. **43**(8): p. 1090-1103.
20. Uversky, V.N., *Intrinsically Disordered Proteins and Their Environment: Effects of Strong Denaturants, Temperature, pH, Counter Ions, Membranes, Binding Partners, Osmolytes, and Macromolecular Crowding.* Protein Journal, 2009. **28**(7-8): p. 305-325.
21. Pepper, L. and H.M. Farrell, *INTERACTIONS LEADING TO FORMATION OF CASEIN SUBMICELLES.* Journal of Dairy Science, 1982. **65**(12): p. 2259-2266.
22. Doi, H., F. Ibuki, and M. Kanamori, *INTERACTIONS OF KAPPA-CASEIN COMPONENTS WITH ALPHA-S1-CASEINS AND BETA-CASEINS.* Agricultural and Biological Chemistry, 1979. **43**(6): p. 1301-1308.
23. Fonin, A., et al., *Biological soft matter: intrinsically disordered proteins in liquid-liquid phase separation and biomolecular condensates.* ESSAYS IN BIOCHEMISTRY, 2022. **66**: p. 831-847.

24. Melnikova, D.L., et al., *Translational Diffusion and Self-Association of an Intrinsically Disordered Protein κ-Casein Using NMR with Ultra-High Pulsed-Field Gradient and Time-Resolved FRET.* Journal of Physical Chemistry B, 2024. **128**(32): p. 7781-7791.
25. Nesmelova, I.V., et al., *Translational diffusion of unfolded and intrinsically disordered proteins,* in *Dancing Protein Clouds: Intrinsically Disordered Proteins in Health and Disease, Pt A*, V.N. Uversky, Editor. 2019. p. 85-108.
26. Izsvák, Z., Z. Ivics, and R.H. Plasterk, *Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates.* Journal of Molecular Biology, 2000. **302**(1): p. 93-102.
27. Harris, J.W., et al., *Construction of a Tc1-like transposon Sleeping Beauty-based gene transfer plasmid vector for generation of stable transgenic mammalian cell clones.* Analytical Biochemistry, 2002. **310**(1): p. 15-26.
28. Liu, H.Z. and G.A. Visner, *Applications of Sleeping Beauty transposons for nonviral gene therapy.* Iubmb Life, 2007. **59**(6): p. 374-379.
29. Aronovich, E.L., et al., *Systemic Correction of Storage Disease in MPS I NOD/SCID Mice Using the Sleeping Beauty Transposon System*. Molecular Therapy, 2009. **17**(7): p. 1136-1144.
30. Huls, M.H., et al., *Clinical Application of Sleeping Beauty and Artificial Antigen Presenting Cells to Genetically Modify T Cells from Peripheral and Umbilical Cord Blood.* Jove-Journal of Visualized Experiments, 2013(72).
31. Negrao, M.V., et al., *First-in-human phase 1/2 study of autologous T cells engineered using the Sleeping Beauty System transposon/transposase to express T-cell receptors (TCRs) reactive against cancer-specific mutations in patients with advanced solid tumors.* Journal of Clinical Oncology, 2022. **40**(16).
32. Ivics, Z., et al., *Molecular reconstruction of Sleeping beauty, a Tc1-like transposon from fish, and its transposition in human cells.* Cell, 1997. **91**(4): p. 501-510.
33. Mizuuchi, M., et al., *Control of transposase activity within a transpososome by the configuration of the flanking DNA segment of the transposon.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(37): p. 14622-14627.
34. Bouuaert, C.C., et al., *Crosstalk between transposase subunits during cleavage of the mariner transposon.* Nucleic Acids Research, 2014. **42**(9): p. 5799-5808.
35. Dornan, J., H. Grey, and J.M. Richardson, *Structural role of the flanking DNA in mariner transposon excision.* Nucleic Acids Research, 2015. **43**(4): p. 2424-2432.
36. Ochmann, M.T. and Z. Ivics, *Jumping Ahead with Sleeping Beauty: Mechanistic Insights into Cut-and-Paste Transposition.* Viruses-Basel, 2021. **13**(1).
37. Liu, G.Y., et al., *Excision of Sleeping Beauty transposons:: parameters and applications to gene therapy.* Journal of Gene Medicine, 2004. **6**(5): p. 574-583.
38. Park, J., S.R. Yant, and M.A. Kay, *The altered binding properties of sleeping beauty transposase hyperactive mutants may explain their enhanced efficacy.* Molecular Therapy, 2004. **9**: p. S57-S57.
39. Yant, S.R., et al., *Hyperactive transposase mutants of the Sleeping Beauty transposon.* Molecular Therapy, 2004. **9**: p. S310-S310.
40. Zayed, H., et al., *Development of hyperactive Sleeping Beauty transposon vectors by mutational analysis.* Molecular Therapy, 2004. **9**(2): p. 292-304.
41. Baus, J., et al., *Hyperactive transposase mutants of the Sleeping Beauty transposon.* Molecular Therapy, 2005. **12**(6): p. 1148-1156.
42. Kovac, A., C. Miskey, and Z. Ivics, *Sleeping Beauty Transposon Insertions into Nucleolar DNA by an Engineered Transposase Localized in the Nucleolus.* International Journal of Molecular Sciences, 2023. **24**(19).
43. Konnova, T.A., C.M. Singer, and I.V. Nesmelova, *NMR solution structure of the RED subdomain of the Sleeping Beauty transposase.* Protein Science, 2017. **26**(6): p. 1171-1181.
44. Voigt, F., et al., *Sleeping Beauty transposase structure allows rational design of hyperactive variants for genetic engineering.* Nature Communications, 2016. **7**.
45. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(4096): p. 223-30.
46. Bellissent-Funel, M.C., et al., *Water Determines the Structure and Dynamics of Proteins.* Chem Rev, 2016. **116**(13): p. 7673-97.
47. Kauzmann, W., *Some Factors in the Interpretation of Protein Denaturation*, in *Advances in Protein Chemistry Volume 14*, C.B. Anfinsen, et al., Editors. 1959, Academic Press. p. 1-63.
48. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins, 2000. **41**(3): p. 415-27.

49. Mao, A.H., N. Lyle, and R.V. Pappu, *Describing sequence-ensemble relationships for intrinsically disordered proteins.* Biochem J, 2013. **449**(2): p. 307-18.
50. Weathers, E.A., et al., *Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein.* FEBS Lett, 2004. **576**(3): p. 348-52.
51. Wright, P.E. and H.J. Dyson, *Intrinsically disordered proteins in cellular signalling and regulation.* Nat Rev Mol Cell Biol, 2015. **16**(1): p. 18-29.
52. Tompa, P., et al., *Intrinsically disordered proteins: emerging interaction specialists.* Curr Opin Struct Biol, 2015. **35**: p. 49-59.
53. Piersimoni, L., et al., *Lighting up Nobel Prize-winning studies with protein intrinsic disorder.* Cell Mol Life Sci, 2022. **79**(8): p. 449.
54. Denning, D.P., et al., *Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded.* Proc Natl Acad Sci U S A, 2003. **100**(5): p. 2450-5.
55. Han, T.W., et al., *Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies.* Cell, 2012. **149**(4): p. 768-79.
56. Rout, M.P., et al., *The yeast nuclear pore complex: composition, architecture, and transport mechanism.* J Cell Biol, 2000. **148**(4): p. 635-51.
57. Nesmelova, I.V., et al., *Translational diffusion of unfolded and intrinsically disordered proteins.* Prog Mol Biol Transl Sci, 2019. **166**: p. 85-108.
58. de Gennes, P.G., *Scaling concepts in polymer physics*. 1979, Ithaca, N.Y.: Cornell University Press. 324 p.
59. Skirda, V.D., et al., *On the Generalized Concentration and Molecular Mass Dependencies of Macromolecular Self-Diffusion in Polymer-Solutions.* Polymer, 1988. **29**(7): p. 1294-1300.
60. Nesmelova, I.V., V.D. Skirda, and V.D. Fedotov, *Generalized concentration dependence of globular protein self-diffusion coefficients in aqueous solutions.* Biopolymers, 2002. **63**(2): p. 132-40.
61. Tokuyama, M. and I.I. Oppenheim, *Dynamics of hard-sphere suspensions.* Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics, 1994. **50**(1): p. R16-R19.
62. Marsh, J.A. and J.D. Forman-Kay, *Sequence determinants of compaction in intrinsically disordered proteins.* Biophys J, 2010. **98**(10): p. 2383-90.
63. Dudas, E.F. and A. Bodor, *Quantitative, Diffusion NMR Based Analytical Tool To Distinguish Folded, Disordered, and Denatured Biomolecules.* Anal Chem, 2019. **91**(8): p. 4929-4933.
64. Uversky, V.N., et al., *Length-dependent compaction of intrinsically disordered proteins.* FEBS Lett, 2012. **586**(1): p. 70-3.
65. Flory, P.J., *Principles of Polymer Chemistry*. 1953: Cornell University Press.
66. Melnikova, D.L., V.D. Skirda, and I.V. Nesmelova, *Effect of Intrinsic Disorder and Self-Association on the Translational Diffusion of Proteins: The Case of alpha-Casein.* J Phys Chem B, 2017. **121**(14): p. 2980-2988.
67. Das, R.K. and R.V. Pappu, *Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues.* Proc Natl Acad Sci U S A, 2013. **110**(33): p. 13392-7.
68. Schuler, B., et al., *Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins.* Annu Rev Biophys, 2016. **45**: p. 207-31.
69. Skirda, V.D., et al., *Translational Mobility of Macromolecules in Networks.* Vysokomolekulyarnye Soedineniya Seriya B, 1988. **30**(4): p. 313-314.
70. Nesmelova, I.V. and V.D. Fedotov, *Self-diffusion and self-association of lysozyme molecules in solution.* Biochim Biophys Acta, 1998. **1383**(2): p. 311-6.
71. Price, W.S., F. Tsuchiya, and Y. Arata, *Lysozyme aggregation and solution properties studied using PGSE NMR diffusion measurements.* Journal of the American Chemical Society, 1999. **121**(49): p. 11503-11512.
72. Ermolina, I.V., V.D. Fedotov, and Y.D. Feldman, *Structure and dynamic behavior of protein molecules in solution.* Physica A, 1998. **249**(1-4): p. 347-352.
73. Gafurov, I.R., et al., *NMR study of the structure of aqueous gelatine gels and the process of their formation.* Polymer Science U.S.S.R., 1989. **31**(2): p. 292-300.
74. Gafurov, I.R., et al., *Self-diffusion and gelation in benzyl alcohol solutions of cellulose triacetate.* Polymer Science U.S.S.R., 1988. **30**(7): p. 1639-1644.
75. Fonin, A.V., et al., *Biological soft matter: intrinsically disordered proteins in liquid-liquid phase separation and biomolecular condensates.* Essays Biochem, 2022. **66**(7): p. 831-847.
76. Redwan, E.M., et al., *Disorder in milk proteins: caseins, intrinsically disordered colloids.* Curr Protein Pept Sci, 2015. **16**(3): p. 228-42.

77. Sawyer, L. and C. Holt, *The secondary structure of milk proteins and their biological function.* J Dairy Sci, 1993. **76**(10): p. 3062-78.
78. Syme, C.D., et al., *A Raman optical activity study of rheomorphism in caseins, synucleins and tau. New insight into the structure and behaviour of natively unfolded proteins.* Eur J Biochem, 2002. **269**(1): p. 148-56.
79. Fox, P.F. and A. Brodkorb, *The casein micelle: Historical aspects, current concepts and significance.* International Dairy Journal, 2008. **18**(7): p. 677-684.
80. Huppertz, T., P.F. Fox, and A.L. Kelly, *The caseins: Structure, stability, and functionality*, in *Proteins in Food Processing*, R.Y. Yada, Editor. 2018, Woodhead Publishing. p. 49-92.
81. Carver, J.A. and C. Holt, *Functional and dysfunctional folding, association and aggregation of caseins.* Adv Protein Chem Struct Biol, 2019. **118**: p. 163-216.
82. Swaisgood, H.E., *Chemistry of the Caseins*, in *Advanced Dairy Chemistry—1 Proteins*, P.F. Fox and P.L.H. McSweeney, Editors. 2003, Springer US: Boston, MA. p. 139-201.
83. Holt, C., et al., *Invited review: Caseins and the casein micelle: their biological functions, structures, and behavior in foods.* J Dairy Sci, 2013. **96**(10): p. 6127-46.
84. Horne, D.S., *Casein interactions: Casting light on the black boxes, the structure in dairy products.* International Dairy Journal, 1998. **8**(3): p. 171-177.
85. Lucey, J.A. and D.S. Horne, *Perspectives on casein interactions.* International Dairy Journal, 2018. **85**: p. 56-65.
86. Ecroyd, H., et al., *The dissociated form of kappa-casein is the precursor to its amyloid fibril formation.* Biochem J, 2010. **429**(2): p. 251-60.
87. Farrell, H.M., Jr., et al., *Environmental influences on bovine kappa-casein: reduction and conversion to fibrillar (amyloid) structures.* J Protein Chem, 2003. **22**(3): p. 259-73.
88. Leonil, J., et al., *Kinetics of fibril formation of bovine kappa-casein indicate a conformational rearrangement as a critical step in the process.* Journal of Molecular Biology, 2008. **381**(5): p. 1267-1280.
89. Thorn, D.C., et al., *Amyloid fibril formation by bovine milk kappa-casein and its inhibition by the molecular chaperones alphaS- and beta-casein.* Biochemistry, 2005. **44**(51): p. 17027-36.
90. Thorn, D.C., et al., *Casein structures in the context of unfolded proteins.* International Dairy Journal, 2015. **46**: p. 2-11.
91. deKruif, C.G. and E.B. Zhulina, *kappa-casein as a polyelectrolyte brush on the surface of casein micelles.* Colloids and Surfaces a-Physicochemical and Engineering Aspects, 1996. **117**(1-2): p. 151-159.
92. Price, W.S., *NMR studies of translational motion*. Cambridge molecular science. 2009, Cambridge ; New York: Cambridge University Press. xxii, 393 p.
93. Melnikova, D.L., V.D. Skirda, and I.V. Nesmelova, *Effect of Reducing Agent TCEP on Translational Diffusion and Supramolecular Assembly in Aqueous Solutions of alpha-Casein.* J Phys Chem B, 2019. **123**(10): p. 2305-2315.
94. McMeekin, T.L., M.L. Groves, and N.J. Hipp, *Apparent Specific Volume of α-Casein and β-Casein and the Relationship of Specific Volume to Amino Acid Composition.* Journal of the American Chemical Society, 2002. **71**(10): p. 3298-3300.
95. Sindhu, J.S. and S. Arora, *Milk | Buffalo Milk*, in *Encyclopedia of Dairy Sciences*, J.W. Fuquay, Editor. 2011, Academic Press: San Diego. p. 503-511.
96. Tanner, J.E., *Use of Stimulated Echo in NMR-Diffusion Studies.* Journal of Chemical Physics, 1970. **52**(5): p. 2523-2526.
97. Doroginitskii, M.M. and A.S. Ivanov, *Analysis of spin-echo diffusion decay data files.* 2024, Certificate No. 2024660253 of the Russian Federation. Kazan Federal University.
98. Tikhonov, A.N. and V.I.A. Arsenin, *Solutions of Ill-posed Problems*. 1977: Winston.
99. Hildebrandt, N., *How to Apply FRET: From Experimental Design to Data Analysis*, in *FRET – Förster Resonance Energy Transfer*. 2013. p. 105-163.
100. Rasmussen, L.K., et al., *Disulphide-linked caseins and casein micelles.* International Dairy Journal, 1999. **9**(3-6): p. 215-218.
101. Marchesseau, S., et al., *Casein interactions studied by the surface plasmon resonance technique.* J Dairy Sci, 2002. **85**(11): p. 2711-21.
102. Vreeman, H.J., J.A. Brinkhuis, and C.A. Vanderspek, *Some Association Properties of Bovine Sh-K-Casein.* Biophysical Chemistry, 1981. **14**(2): p. 185-193.
103. Millero, F.J., R. Dexter, and E. Hoff, *Density and Viscosity of Deuterium Oxide Solutions from 5-70 Degrees C.* Journal of Chemical and Engineering Data, 1971. **16**(1): p. 85-&.

104. Jaeger, H.M. and S.R. Nagel, *Physics of the Granular State.* Science, 1992. **255**(5051): p. 1523-1531.
105. Ossowski, S., et al., *Aggregation behavior of bovine kappa- and beta-casein studied with small angle neutron scattering, light scattering, and cryogenic transmission electron microscopy.* Langmuir, 2012. **28**(38): p. 13577-89.
106. Sagidullin, A.I., et al., *Generalized concentration dependence of self-diffusion coefficients in poly(allylcarbosilane) dendrimer solutions.* Macromolecules, 2002. **35**(25): p. 9472-9479.
107. Guseva, S., et al., *Liquid-Liquid Phase Separation Modifies the Dynamic Properties of Intrinsically Disordered Proteins.* J Am Chem Soc, 2023. **145**(19): p. 10548-10563.
108. Wong, L.E., et al., *NMR Experiments for Studies of Dilute and Condensed Protein Phases: Application to the Phase-Separating Protein CAPRIN1.* J Am Chem Soc, 2020. **142**(5): p. 2471-2489.
109. Brady, J.P., et al., *Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation.* Proc Natl Acad Sci U S A, 2017. **114**(39): p. E8194-E8203.
110. VandenDriessche, T., et al., *Emerging potential of transposons for gene therapy and generation of induced pluripotent stem cells.* Blood, 2009. **114**(8): p. 1461-8.
111. Hackett, P.B., D.A. Largaespada, and L.J. Cooper, *A transposon and transposase system for human application.* Mol Ther, 2010. **18**(4): p. 674-83.
112. Hickman, A.B. and F. Dyda, *DNA Transposition at Work.* Chemical Reviews, 2016. **116**(20): p. 12758-12784.
113. Curcio, M.J. and K.M. Derbyshire, *The outs and ins of transposition: from mu to kangaroo.* Nat Rev Mol Cell Biol, 2003. **4**(11): p. 865-77.
114. Miskey, C., et al., *DNA transposons in vertebrate functional genomics.* Cell Mol Life Sci, 2005. **62**(6): p. 629-41.
115. Kawakami, K., D.A. Largaespada, and Z. Ivics, *Transposons As Tools for Functional Genomics in Vertebrate Models.* Trends Genet, 2017. **33**(11): p. 784-801.
116. Ivics, Z., et al., *Transposon-mediated genome manipulation in vertebrates (vol 6, pg 415, 2009).* Nature Methods, 2009. **6**(7): p. 546-546.
117. Hudecek, M., et al., *Going non-viral: the Sleeping Beauty transposon system breaks on through to the clinical side.* Crit Rev Biochem Mol Biol, 2017. **52**(4): p. 355-380.
118. Kebriaei, P., et al., *Gene Therapy with the Sleeping Beauty Transposon System.* Trends Genet, 2017. **33**(11): p. 852-870.
119. Woodard, L.E. and M.H. Wilson, *piggyBac-ing models and new therapeutic strategies.* Trends Biotechnol, 2015. **33**(9): p. 525-33.
120. Kebriaei, P., et al., *Phase I trials using Sleeping Beauty to generate CD19-specific CAR T cells.* J Clin Invest, 2016. **126**(9): p. 3363-76.
121. Miskey, C., et al., *Engineered Sleeping Beauty transposase redirects transposon integration away from genes.* Nucleic Acids Res, 2022. **50**(5): p. 2807-2825.
122. Hackett, P.B., et al., *Sleeping beauty transposon-mediated gene therapy for prolonged expression.* Adv Genet, 2005. **54**: p. 189-232.
123. Ivics, Z., et al., *Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells.* Cell, 1997. **91**(4): p. 501-10.
124. Mates, L., et al., *Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates.* Nat Genet, 2009. **41**(6): p. 753-61.
125. Balciunas, D., et al., *Harnessing a high cargo-capacity transposon for genetic applications in vertebrates.* PLoS Genet, 2006. **2**(11): p. e169.
126. Miskey, C., et al., *The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells.* Nucleic Acids Res, 2003. **31**(23): p. 6873-81.
127. Ding, S., et al., *Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice.* Cell, 2005. **122**(3): p. 473-83.
128. Yusa, K., et al., *A hyperactive piggyBac transposase for mammalian applications.* Proc Natl Acad Sci U S A, 2011. **108**(4): p. 1531-6.
129. Ivics, Z., et al., *Identification of functional domains and evolution of Tc1-like transposable elements.* Proc Natl Acad Sci U S A, 1996. **93**(10): p. 5008-13.
130. Cui, Z., et al., *Structure-function analysis of the inverted terminal repeats of the sleeping beauty transposon.* J Mol Biol, 2002. **318**(5): p. 1221-35.
131. Zayed, H., et al., *Development of hyperactive sleeping beauty transposon vectors by mutational analysis.* Mol Ther, 2004. **9**(2): p. 292-304.

132. Izsvak, Z., et al., *Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition.* J Biol Chem, 2002. **277**(37): p. 34581-8.
133. Yant, S.R., et al., *Mutational analysis of the N-terminal DNA-binding domain of sleeping beauty transposase: critical residues for DNA binding and hyperactivity in mammalian cells.* Mol Cell Biol, 2004. **24**(20): p. 9239-47.
134. Voigt, F., et al., *Sleeping Beauty transposase structure allows rational design of hyperactive variants for genetic engineering.* Nat Commun, 2016. **7**: p. 11126.
135. Querques, I., et al., *A highly soluble Sleeping Beauty transposase improves control of gene insertion.* Nat Biotechnol, 2019. **37**(12): p. 1502-1512.
136. Singer, C.M., et al., *Rigidity and flexibility characteristics of DD[E/D]-transposases Mos1 and Sleeping Beauty.* Proteins, 2019. **87**(4): p. 313-325.
137. Whitmore, L. and B.A. Wallace, *DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W668-73.
138. Whitmore, L. and B.A. Wallace, *Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases.* Biopolymers, 2008. **89**(5): p. 392-400.
139. Live, D.H., et al., *Long-Range Hydrogen-Bond Mediated Effects in Peptides - N-15 Nmr-Study of Gramicidin-S in Water and Organic-Solvents.* Journal of the American Chemical Society, 1984. **106**(7): p. 1939-1941.
140. Muhandiram, D.R. and L.E. Kay, *Gradient-Enhanced Triple-Resonance Three-Dimensional NMR Experiments with Improved Sensitivity.* J. Magn. Res., 1994. **B103**: p. 203-216.
141. Delaglio, F., et al., *NMRPipe: a multidimensional spectral processing system based on UNIX pipes.* J Biomol NMR, 1995. **6**(3): p. 277-93.
142. Keller, R.L.J., *Keller, R.L.J. (2004). The computer aided resonance assignment tutorial. (Cantina Verlag).* 2004.
143. Johnson, B.A. and R.A. Blevins, *NMRView: A computer program for the visualization and analysis of NMR data.* J. Biomol. NMR, 1994. **4**: p. 603-614.
144. Shen, Y., et al., *TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts.* J Biomol NMR, 2009. **44**(4): p. 213-23.
145. Schwieters, C.D., et al., *The Xplor-NIH NMR molecular structure determination package.* J Magn Reson, 2003. **160**(1): p. 65-73.
146. Laskowski, R.A., et al., *Procheck - a Program to Check the Stereochemical Quality of Protein Structures.* Journal of Applied Crystallography, 1993. **26**: p. 283-291.
147. PyMOL, *The PyMOL Molecular Graphics System, Version 1.5.0.4.* Schrödinger, LLC. , 2012.
148. Lakowicz, J.R., *Principles of frequency-domain fluorescence spectroscopy and applications to cell membranes.* Subcell Biochem, 1988. **13**: p. 89-126.
149. Arnott, S., et al., *Left-handed DNA helices.* Nature, 1980. **283**(5749): p. 743-5.
150. Arnott, S., et al., *Structures of synthetic polynucleotides in the A-RNA and A'-RNA conformations: x-ray diffraction analyses of the molecular conformations of polyadenylic acid--polyuridylic acid and polyinosinic acid--polycytidylic acid.* J Mol Biol, 1973. **81**(2): p. 107-22.
151. Fuller, W., et al., *The Molecular Configuration of Deoxyribonucleic Acid. Iv. X-Ray Diffraction Study of the a Form.* J Mol Biol, 1965. **12**: p. 60-76.
152. Lakshiminarayanan, A.V. and V. Sasisekharan, *Stereochemistry of nucleic acids and polynucleotides. II. Allowed conformations of the monomer unit for different ribose puckerings.* Biochim Biophys Acta, 1970. **204**(1): p. 49-59.
153. van Zundert, G.C.P., et al., *The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes.* J Mol Biol, 2016. **428**(4): p. 720-725.
154. Liu, Z., et al., *Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach.* Proteins, 2008. **72**(4): p. 1114-24.
155. Takeda, T., R.I. Corona, and J.T. Guo, *A knowledge-based orientation potential for transcription factor-DNA docking.* Bioinformatics, 2013. **29**(3): p. 322-30.
156. Liu, Z., et al., *Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential.* Nucleic Acids Res, 2005. **33**(2): p. 546-58.
157. Grimsley, G.R., J.M. Scholtz, and C.N. Pace, *A summary of the measured pK values of the ionizable groups in folded proteins.* Protein Sci, 2009. **18**(1): p. 247-51.
158. Markley, J.L., *Observation of histidine residues in proteins by nuclear magnetic resonance spectroscopy.* Accounts of Chemical Research, 1975. **8**(2): p. 70-80.

159. Yin, S., F. Ding, and N.V. Dokholyan, *Eris: an automated estimator of protein stability.* Nat Methods, 2007. **4**(6): p. 466-7.
160. Ding, F. and N.V. Dokholyan, *Emergence of protein fold families through rational design.* PLoS Comput Biol, 2006. **2**(7): p. e85.
161. Richardson, J.M., et al., *Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote.* Cell, 2009. **138**(6): p. 1096-108.
162. Watkins, S., G. van Pouderoyen, and T.K. Sixma, *Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA.* Nucleic Acids Res, 2004. **32**(14): p. 4306-12.
163. Fleming, P.J. and K.G. Fleming, *HullRad: Fast Calculations of Folded and Disordered Protein and Nucleic Acid Hydrodynamic Properties.* Biophys J, 2018. **114**(4): p. 856-869.
164. Ochmann, M.T. and Z. Ivics, *Jumping Ahead with Sleeping Beauty: Mechanistic Insights into Cut-and-Paste Transposition.* Viruses, 2021. **13**(1).
165. Wu, J., et al., *High performance transcription factor-DNA docking with GPU computing.* Proteome Sci, 2012. **10 Suppl 1**(Suppl 1): p. S17.
166. Lapham, J., et al., *Measurement of diffusion constants for nucleic acids by NMR.* J Biomol NMR, 1997. **10**(3): p. 255-62.
167. Wang, Y., et al., *Regulated complex assembly safeguards the fidelity of Sleeping Beauty transposition.* Nucleic Acids Res, 2017. **45**(1): p. 311-326.
168. Perez-Borrajero, C., et al., *Conformational Plasticity and DNA-Binding Specificity of the Eukaryotic Transcription Factor Pax5.* Biochemistry, 2021. **60**(2): p. 104-117.
169. Czerny, T., G. Schaffner, and M. Busslinger, *DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site.* Genes Dev, 1993. **7**(10): p. 2048-61.
170. Corona, R.I. and J.T. Guo, *Statistical analysis of structural determinants for protein-DNA-binding specificity.* Proteins, 2016. **84**(8): p. 1147-61.
171. Lohe, A.R. and D.L. Hartl, *Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation.* Mol Biol Evol, 1996. **13**(4): p. 549-55.
172. Yant, S.R., et al., *Somatic integration and long-term transgene expression in normal and haemophilic mice using a DNA transposon system.* Nat Genet, 2000. **25**(1): p. 35-41.
173. Geurts, A.M., et al., *Gene transfer into genomes of human cells by the sleeping beauty transposon system.* Mol Ther, 2003. **8**(1): p. 108-17.
174. Harmening, N., et al., *Enhanced Biosafety of the Sleeping Beauty Transposon System by Using mRNA as Source of Transposase to Efficiently and Stably Transfect Retinal Pigment Epithelial Cells.* Biomolecules, 2023. **13**(4).
175. Yant, S.R., et al., *High-resolution genome-wide mapping of transposon integration in mammals.* Mol Cell Biol, 2005. **25**(6): p. 2085-94.
176. Moldt, B., et al., *Comparative genomic integration profiling of Sleeping Beauty transposons mobilized with high efficacy from integrase-defective lentiviral vectors in primary human cells.* Mol Ther, 2011. **19**(8): p. 1499-510.
177. Gogol-Doring, A., et al., *Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4(+) T Cells.* Mol Ther, 2016. **24**(3): p. 592-606.
178. Roman, Y., et al., *Biochemical characterization of a SET and transposase fusion protein, Metnase: its DNA binding and DNA cleavage activity.* Biochemistry, 2007. **46**(40): p. 11369-76.
179. Chen, Q., et al., *Structural and genome-wide analyses suggest that transposon-derived protein SETMAR alters transcription and splicing.* J Biol Chem, 2022. **298**(5): p. 101894.
180. Spolar, R.S. and M.T. Record, Jr., *Coupling of local folding to site-specific binding of proteins to DNA.* Science, 1994. **263**(5148): p. 777-84.
181. Hard, T., *NMR studies of protein-nucleic acid complexes: structures, solvation, dynamics and coupled protein folding.* Q Rev Biophys, 1999. **32**(1): p. 57-98.
182. Ivics, Z., et al., *Targeted Sleeping Beauty transposition in human cells.* Mol Ther, 2007. **15**(6): p. 1137-44.
183. Yant, S.R., et al., *Site-directed transposon integration in human cells.* Nucleic Acids Res, 2007. **35**(7): p. e50.
184. Voigt, K., et al., *Retargeting sleeping beauty transposon insertions by engineered zinc finger DNA-binding domains.* Mol Ther, 2012. **20**(10): p. 1852-62.

185. Sandoval-Villegas, N., et al., *Contemporary Transposon Tools: A Review and Guide through Mechanisms and Applications of Sleeping Beauty, piggyBac and Tol2 for Genome Engineering.* Int J Mol Sci, 2021. **22**(10).
186. Narayanavari, S.A., et al., *Sleeping Beauty transposition: from biology to applications.* Crit Rev Biochem Mol Biol, 2017. **52**(1): p. 18-44.
187. Amberger, M. and Z. Ivics, *Latest Advances for the Sleeping Beauty Transposon System: 23 Years of Insomnia but Prettier than Ever: Refinement and Recent Innovations of the Sleeping Beauty Transposon System Enabling Novel, Nonviral Genetic Engineering Applications.* Bioessays, 2020. **42**(11): p. e2000136.
188. Aronovich, E.L., R.S. McIvor, and P.B. Hackett, *The Sleeping Beauty transposon system: a non-viral vector for gene therapy.* Hum Mol Genet, 2011. **20**(R1): p. R14-20.
189. Irving, M., et al., *Choosing the Right Tool for Genetic Engineering: Clinical Lessons from Chimeric Antigen Receptor-T Cells.* Hum Gene Ther, 2021. **32**(19-20): p. 1044-1058.
190. Liu, G., et al., *Target-site preferences of Sleeping Beauty transposons.* J Mol Biol, 2005. **346**(1): p. 161-73.
191. Hackett, C.S., A.M. Geurts, and P.B. Hackett, *Predicting preferential DNA vector insertion sites: implications for functional genomics and gene therapy.* Genome Biol, 2007. **8 Suppl 1**: p. S12.
192. Hacein-Bey-Abina, S., et al., *Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1.* J Clin Invest, 2008. **118**(9): p. 3132-42.
193. Sultana, T., et al., *Integration site selection by retroviruses and transposable elements in eukaryotes.* Nat Rev Genet, 2017. **18**(5): p. 292-308.
194. Hackett, P.B., et al., *Evaluating risks of insertional mutagenesis by DNA transposons in gene therapy.* Transl Res, 2013. **161**(4): p. 265-83.
195. Konnova, T.A., C.M. Singer, and I.V. Nesmelova, *NMR solution structure of the RED subdomain of the Sleeping Beauty transposase.* Protein Sci, 2017. **26**(6): p. 1171-1181.
196. Abrusan, G., et al., *Structural Determinants of Sleeping Beauty Transposase Activity.* Mol Ther, 2016. **24**(8): p. 1369-77.
197. Dimura, M., et al., *Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems.* Curr Opin Struct Biol, 2016. **40**: p. 163-185.
198. Clegg, R.M., *[18] Fluorescence resonance energy transfer and nucleic acids*, in *Methods in Enzymology*. 1992, Academic Press. p. 353-388.
199. Magde, D., G.E. Rojas, and P.G. Seybold, *Solvent dependence of the fluorescence lifetimes of xanthene dyes.* Photochemistry and Photobiology, 1999. **70**(5): p. 737-744.
200. Velapoldi, R.A. and H.H. Tonnesen, *Corrected emission spectra and quantum yields for a series of fluorescent compounds in the visible spectral region.* J Fluoresc, 2004. **14**(4): p. 465-72.
201. Abramson, J., et al., *Accurate structure prediction of biomolecular interactions with AlphaFold 3.* Nature, 2024. **630**(8016): p. 493-500.
202. Brunger, A.T., et al., *Crystallography & NMR system:: A new software suite for macromolecular structure determination.* Acta Crystallographica Section D-Biological Crystallography, 1998. **54**: p. 905-921.

APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 2

SEQUENCE COMPARISON OF CASEINS

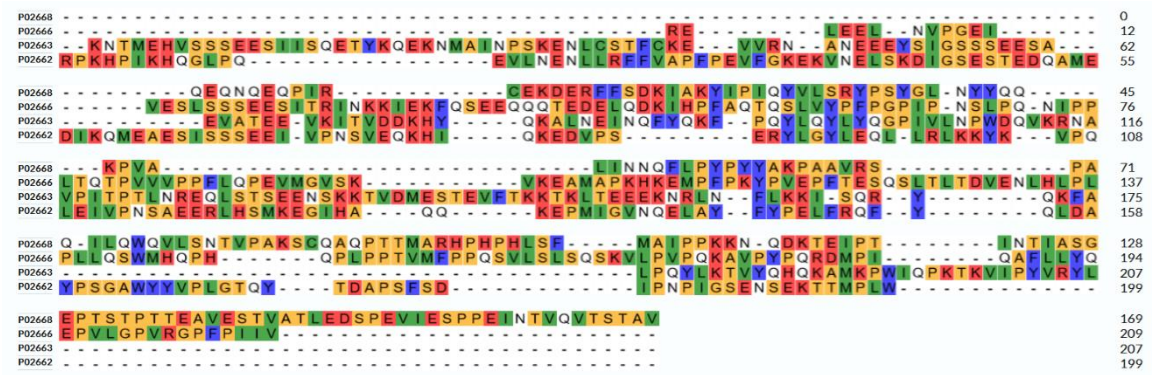

Figure S1. Amino acid sequence alignment with residues properties highlighted according to CLUSTAL color scheme: red – charged, blue – aromatic, green – aliphatic, orange – S, T, A, G, P.

Amino acid sequences of the four bovine caseins, α_{s1}, α_{s2}, β, and κ-casein:

P02668 · CASK_BOVIN Kappa-casein
QEQNQEQPIRCEKDERFFSDKIAKYIPIQYVLSRYPSYGLNYYQQKPVALINNQFLPYPYYAKPAAVRSPAQI
LQWQVLSNTVPAKSCQAQPTTMARHPHPHLSFMAIPPKKNQDKTEIPTINTIASGEPTSTPTTEAVESTVATL
EDSPEVIESPPEINTVQVTSTAV

P02663 · CASA2_BOVIN Alpha-S2-casein
KNTMEHVSSSEESIISQETYKQEKNMAINPSKENLCSTFCKEVVRNANEEEYSIGSSSEESAEVATEEVKITV
DDKHYQKALNEINQFYQKFPQYLQYLYQGPIVLNPWDQVKRNAVPITPTLNREQLSTSEENSKKTVDMEST
EVFTKKTKLTEEEKNRLNFLKKISQRYQKFALPQYLKTVYQHQKAMKPWIQPKTKVIPYVRYL

P02662 · CASA1_BOVIN Alpha-S1-casein
RPKHPIKHQGLPQEVLNENLLRFFVAPFPEVFGKEKVNELSKDIGSESTEDQAMEDIKQMEAESISSSEEIVP
NSVEQKHIQKEDVPSERYLGYLEQLLRLKKYKVPQLEIVPNSAEERLHSMKEGIHAQQKEPMIGVNQELAY
FYPELFRQFYQLDAYPSGAWYYVPLGTQYTDAPSFSDIPNPIGSENSEKTTMPLW

P02666 · CASB_BOVIN Beta-casein
RELEELNVPGEIVESLSSSEESITRINKKIEKFQSEEQQQTEDELQDKIHPFAQTQSLVYPFPGPIPNSLPQNIPP
LTQTPVVVPPFLQPEVMGVSKVKEAMAPKHKEMPFPKYPVEPFTESQSLTLTDVENLHLPLPLLQSWMHQP
HQPLPPTVMFPPQSVLSLSQSKVLPVPQKAVPYPQRDMPIQAFLLYQEPVLGPVRGPFPIIV

GEL ELECTROPHORESIS OF κ-CASEIN

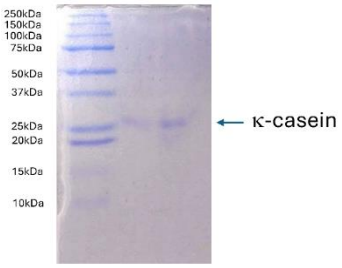

Figure S2. Denaturing SDS-PAGE analysis shows that k-casein is a pure monodisperse species.

CIRCULAR DICHROISM (CD) SPECTROSCOPY OF κ-CASEIN

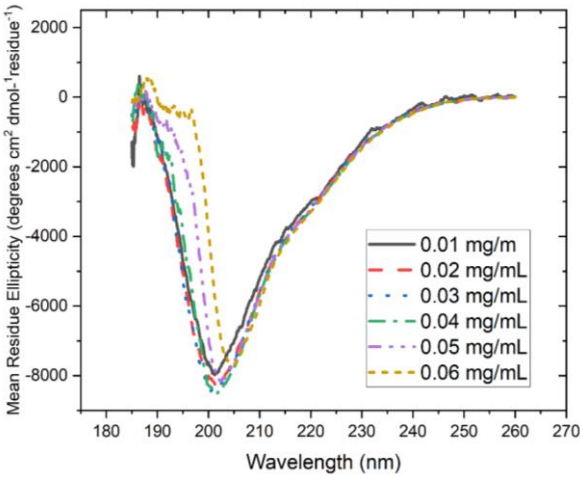

Figure S3 Circular Cichroism (CC) spectroscopy of kappa-casein.

For the far UV CD spectroscopic measurements, the lyophilized powder of κ-casein was dissolved in H_2O at the concentrations of 0.001 to 0.006% (0.01 to 0.06 mg/mL). CD measurements were performed using a Jasco-1500 spectropolarimeter, equipped with a Peltier temperature control system. CD spectra were recorded using a 50 nm/min scan rate, a 4 s D.I.T. response, and a 1 nm bandwidth. Spectra were recorded in the range of 185–260 nm using a quartz glass cell with a path length, *l*, of 1 mm. The corresponding buffer baseline was subtracted from the spectra. Reported spectra are averages of 3–5 scans and are expressed as mean-residue molar ellipticity, [θ], calculated according to the following formula:

$$[\theta] = \frac{M_0 \theta_\lambda}{100 \cdot C \cdot l},$$

where M_0 is the mean residue molar mass, θ_λ is the measured ellipticity in degrees, and C is the protein concentration. CD spectra demonstrate that κ-casein does not have significant secondary structure at our experimental conditions.

STIMULATED ECHO PULSE SEQUENCES USED FOR DIFFUSION MEASUREMENTS

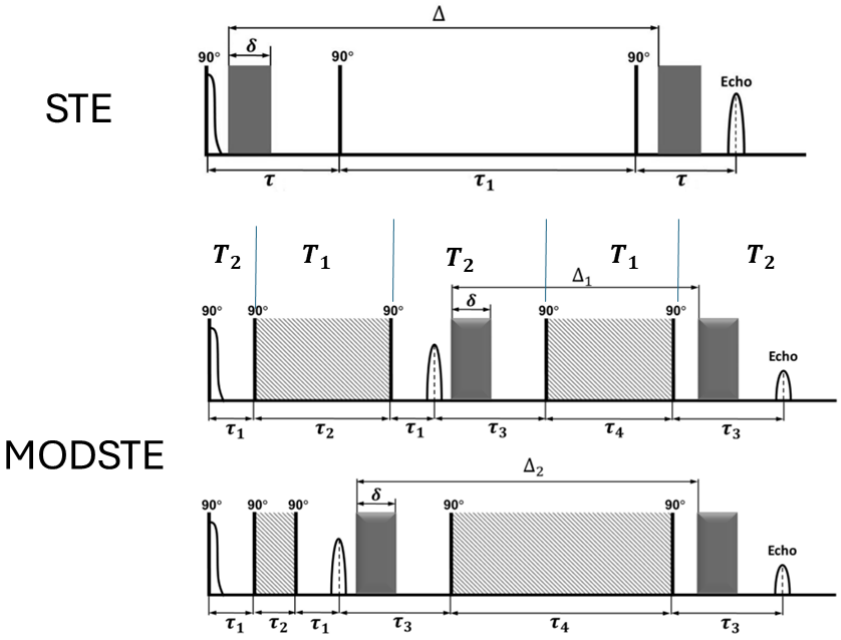

Figure S4. Stimulated echo (STE)[1] and the modified double-stimulated echo (MODSTE)[2] pulse sequences.

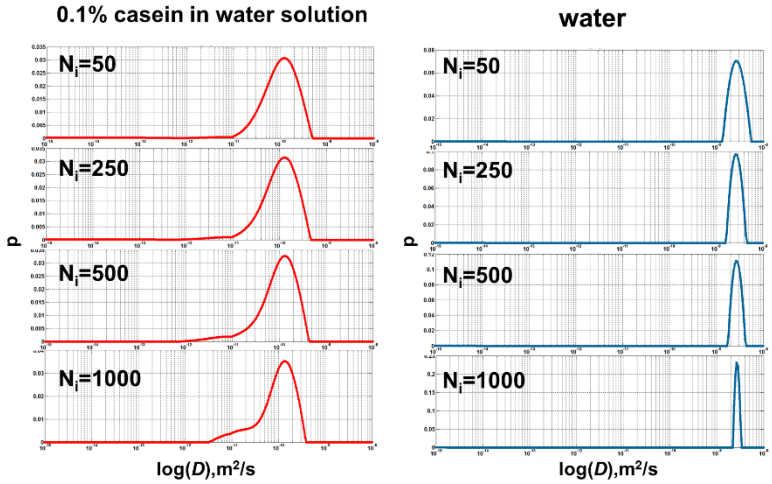

Figure S5. Spectra of diffusion coefficients of κ-casein (red, non-exponential diffusion attenuation) and water (blue, exponential diffusion attenuation) for an aqueous solution of κ-casein at a protein concentration of 0.1% as a function of the numbers of iterations N_i

REVERSIBILITY TEST

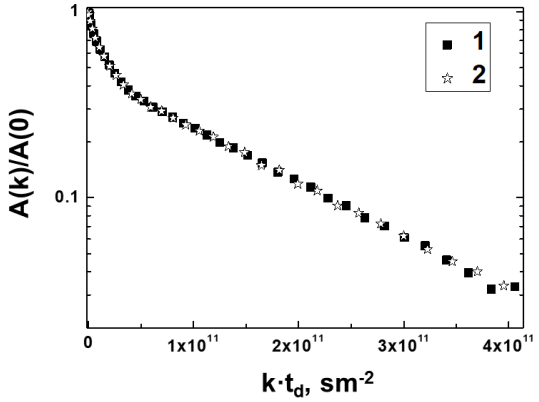

Figure S6. Diffusion attenuations recorded in κ-casein solution with protein concentration of 5%. Curve 1 corresponds to the 5 % κ-casein sample obtained by dissolving a 20 % κ-casein solution. Curve 2 corresponds to the 5 % κ-casein sample prepared directly at this concentration. $k = (\gamma\delta g)^2$, where g is the magnitude of pulsed-field gradient, $A(0)$ is the spin-echo amplitude at $g = 0$, γ is the gyromagnetic ratio for protons, δ is the gradient pulse duration and $t_d = \Delta - \delta/3$ is the diffusion time.

THE FÖRSTER DISTANCE CALCULATION

The Förster distance, R_0 = 2.6 nm, was calculated according to the formula[148]:

$$R_0 = 9786\left[J(\lambda) k^2 \eta^4 Q_D \right]^{1/6}, \quad (1)$$

where λ is the wavelength, $J(\lambda)$ is the spectral overlap integral between the normalized donor emission spectrum $F_D(\lambda)$ and the acceptor absorption spectrum $\varepsilon_A(\lambda)$, $k^2 = 2/3$ is the probes orientation factor, η =1.4 is the refraction index of the medium, and Q_D is the quantum yield of donor-only labeled protein (Q_D = 0.11). Q_D was estimated by the comparison to the quantum yield of quinine sulfate in 0.05 M H_2SO_4 at λ_{ex} = 347.5 nm (Q_S=0.51[200]), according to the equation:

$$Q_D = Q_S \frac{F_D(\lambda)/A_D(\lambda)}{F_S(\lambda)/A_S(\lambda)}, \quad (2)$$

where $F(\lambda)$ is the integral emission and $A(\lambda)$ is the absorbance at the excitation wavelength of donor-labeled protein or quinine sulfate. The analysis of time-resolved fluorescence data was performed using the software package FargoFit, designed by I.V. Negrashov, executing the global least-square fitting of multiple time-resolved luminescence waveforms using different models with ability to link fitting parameters between waveforms.

MOLECULAR DIMENSIONS OF κ-CASEIN

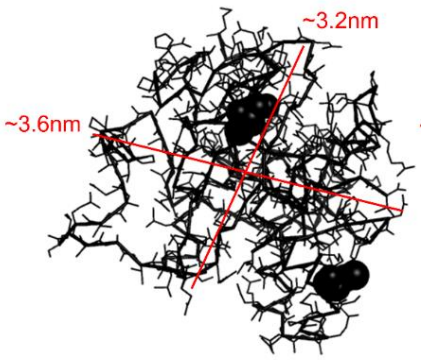

Figure S7. The linear dimensions of a κ-casein molecule were obtained using PSIPRED workbench[3], a secondary structure prediction method that incorporates two feed-forward neural networks which perform an analysis on output obtained from PSIBLAST (Position Specific Iterated – BLAST).

1. Tanner, J. E., Use of Stimulated Echo in Nmr-Diffusion Studies. *J Chem Phys* **1970,** *52* (5), 2523-&.
2. Melnikova, D. L.; Badrieva, Z. E.; Kostin, M. A.; Maller, C.; Stas, M.; Buczek, A.; Broda, M. A.; Kupka, T.; Kelterer, A. M.; Tolstoy, P. M.; Skirda, V. D., On Complex Formation between 5-Fluorouracil and β-Cyclodextrin in Solution and in the Solid State: IR Markers and Detection of Short-Lived Complexes by Diffusion NMR. *Molecules* **2020,** *25* (23).
3. Lakowicz, J. R., Principles of frequency-domain fluorescence spectroscopy and applications to cell membranes. *Subcell Biochem* **1988,** *13*, 89-126.
4. Velapoldi, R. A.; Tonnesen, H. H., Corrected emission spectra and quantum yields for a series of fluorescent compounds in the visible spectral region. *J Fluoresc* **2004,** *14* (4), 465-72.
5. Buchan, D. W. A.; Jones, D. T., The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res* **2019,** *47* (W1), W402-W407.

APPENDIX B: SUPPLEMENTARY MATERIAL FOR CHAPTER 3

Table S1. Nomenclature and amino acid sequences of SB transposase variants used in this study with H19Y mutation highlighted in orange and hyperactive mutations of SB100X highlighted in blue.

Name	**Mutations**	**Amino acid sequence**
PAI		MGKSKEISQDLRKKIVDLHKSGSSLGAISKRLKVPRSSVQT IVRKYKHHGTTQHH
PAI-K14RK33A	K14R K33A	MGKSKEISQDLRK**R**IVDLHKSGSSLGAISKRL**A**VPRSSVQT IVRKYKHHGTTQHH
PAI-H19Y	H19Y	MGKSKEISQDLRKKIVDL**Y**KSGSSLGAISKRLKVPRSSVQT IVRKYKHHGTTQHH
H19Y	K14R H19Y K33A	MGKSKEISQDLRK**R**IVDL**Y**KSGSSLGAISKRL**A**VPRSSVQT IVRKYKHHGTTQHH
SB10		MGKSKEISQDLRKKIVDLHKSGSSLGAISKRLKVPRSSVQT IVRKYKHHGTTQPSYRSGRRRVLSPRDERTLVRKVQINPR TTAKDLVKMLEETGTKVSISTVKRVLYRHNLKGRSARKK PLLQNRHKKARLRFATAHGDKDRTFWRNVLWSDETKIEL FGHNDHRYVWRKKGEACKPKNTIPTVKHGGGSIMLWGC FAAGGTGALHKIDGIMRKENYVDILKQHLKTSVRKLKLG RKWVFQMDNDPKHTSKVVAKWLKDNKVKVLEWPSQSP DLNPIENLWAELKKRVRARRPTNLTQLHQLCQEEWAKIH PTYCGKLVEGYPKRLTQVKQFKGNATKY
SB10-H19Y	H19Y	MGKSKEISQDLRKKIVDL**Y**KSGSSLGAISKRLKVPRSSVQT IVRKYKHHGTTQPSYRSGRRRVLSPRDERTLVRKVQINPR TTAKDLVKMLEETGTKVSISTVKRVLYRHNLKGRSARKK

		PLLQNRHKKARLRFATAHGDKDRTFWRNVLWSDETKIEL FGHNDHRYVWRKKGEACKPKNTIPTVKHGGGSIMLWGC FAAGGTGALHKIDGIMRKENYVDILKQHLKTSVRKLKLG RKWVFQMDNDPKHTSKVVAKWLKDNKVKVLEWPSQSP DLNPIENLWAELKKRVRARRPTNLTQLHQLCQEEWAKIH PTYCGKLVEGYPKRLTQVKQFKGNATKY
SB100X	K14R K33A R115H RKEN214 DAVQ M243H T314N	MGKSKEISQDLRK**R**IVDLHKSGSSLGAISKRL**A**VPRSSVQT IVRKYKHHGTTQPSYRSGRRRVLSPRDERTLVRKVQINPR TTAKDLVKMLEETGTKVSISTVKRVLYRHNLKG**H**SARKK PLLQNRHKKARLRFATAHGDKDRTFWRNVLWSDETKIEL FGHNDHRYVWRKKGEACKPKNTIPTVKHGGGSIMLWGC FAAGGTGALHKIDGIM**DAVQ**YVDILKQHLKTSVRKLKLG RKWVFQ**H**DNDPKHTSKVVAKWLKDNKVKVLEWPSQSP DLNPIENLWAELKKRVRARRPTNLTQLHQLCQEEWAKIH P**N**YCGKLVEGYPKRLTQVKQFKGNATKY
SB100X-H19Y	K14R K33A R115H RKEN214 DAVQ M243H T314N	MGKSKEISQDLRK**R**IVDL**Y**KSGSSLGAISKRL**A**VPRSSVQT IVRKYKHHGTTQPSYRSGRRRVLSPRDERTLVRKVQINPR TTAKDLVKMLEETGTKVSISTVKRVLYRHNLKGHSARKK PLLQNRHKKARLRFATAHGDKDRTFWRNVLWSDETKIEL FGHNDHRYVWRKKGEACKPKNTIPTVKHGGGSIMLWGC FAAGGTGALHKIDGIM**DAVQ**YVDILKQHLKTSVRKLKLG RKWVFQ**H**DNDPKHTSKVVAKWLKDNKVKVLEWPSQSP DLNPIENLWAELKKRVRARRPTNLTQLHQLCQEEWAKIH P**N**YCGKLVEGYPKRLTQVKQFKGNATKY

Table S2. Structural statistics for the NMR structure of the H19Y

Restraints and statistics	
Restraints	
NOE distance restraints (total)	557
intra-residue (j–i = 0)	302
sequential (j–i = 1)	145
medium range (j–i = 2)	22
medium range (j–i = 3)	50
medium range (j–i = 4)	15
long range (j–i ≥ 5)	23
Hydrogen bonds	28
TALOS derived dihedral angle restraints	66
Violations	
NOE distance violations >0.3 Å	0
Dihedral angle violations >5°	0
RMS deviation from mean structure (Å)	
Backbone atoms	0.75
Ramachandran statistics (backbone, ordered residues, ensemble of 10 structures)	
Analyzed	352/456 (77%)
Favored	330 (94%)
Allowed region	22 (6%)
Outliers	0 (0%)

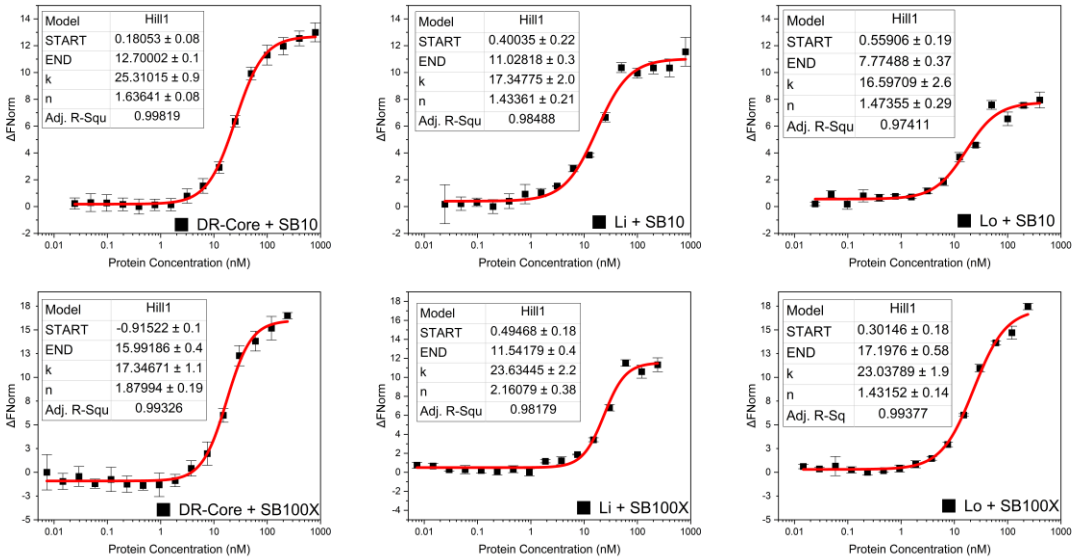

Figure S1. The MST data for SB10 and SB100X full-length transposases binding to DNA. The solid lines represent dose-response fits of the experimental data using the Hill function. The curves are averaged over n ≥ 3 independent experiments, with error bars representing S.D.

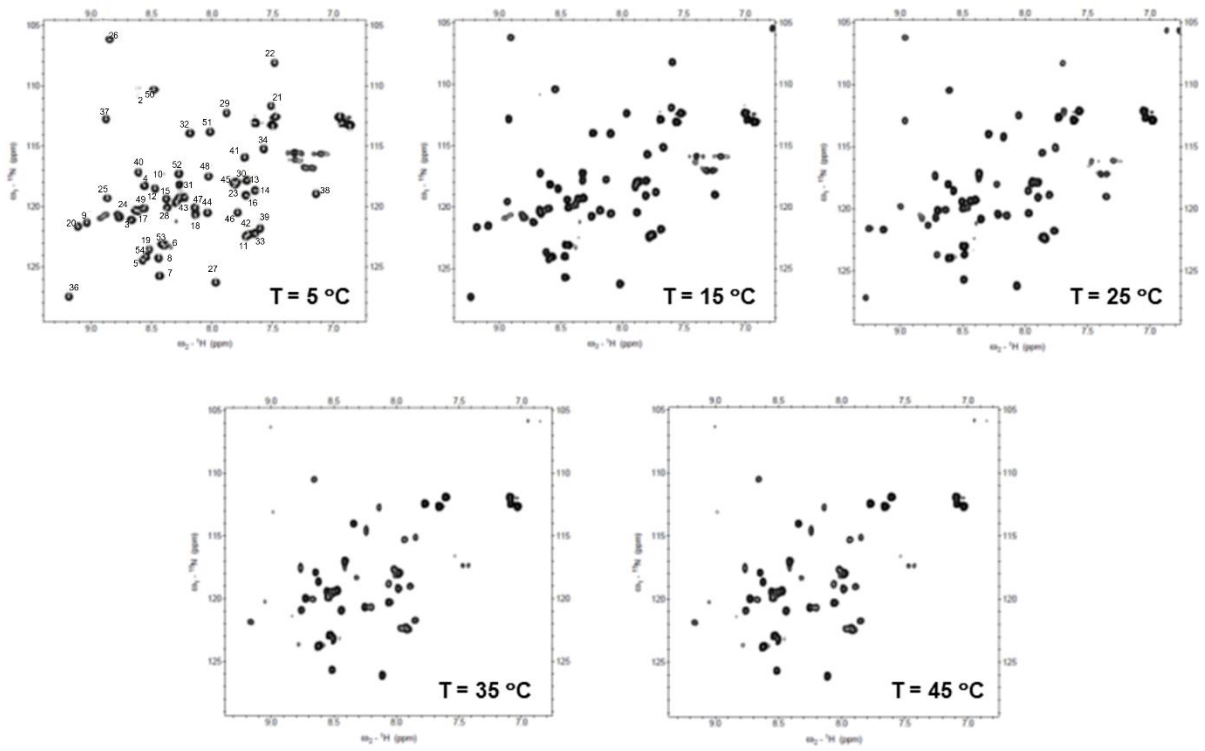

Figure S2. The [^{1}H,^{15}N]-HSQC spectra of the H19Y mutant collected at 5, 15, 25, 35, and 45 °C at pH 5.2.

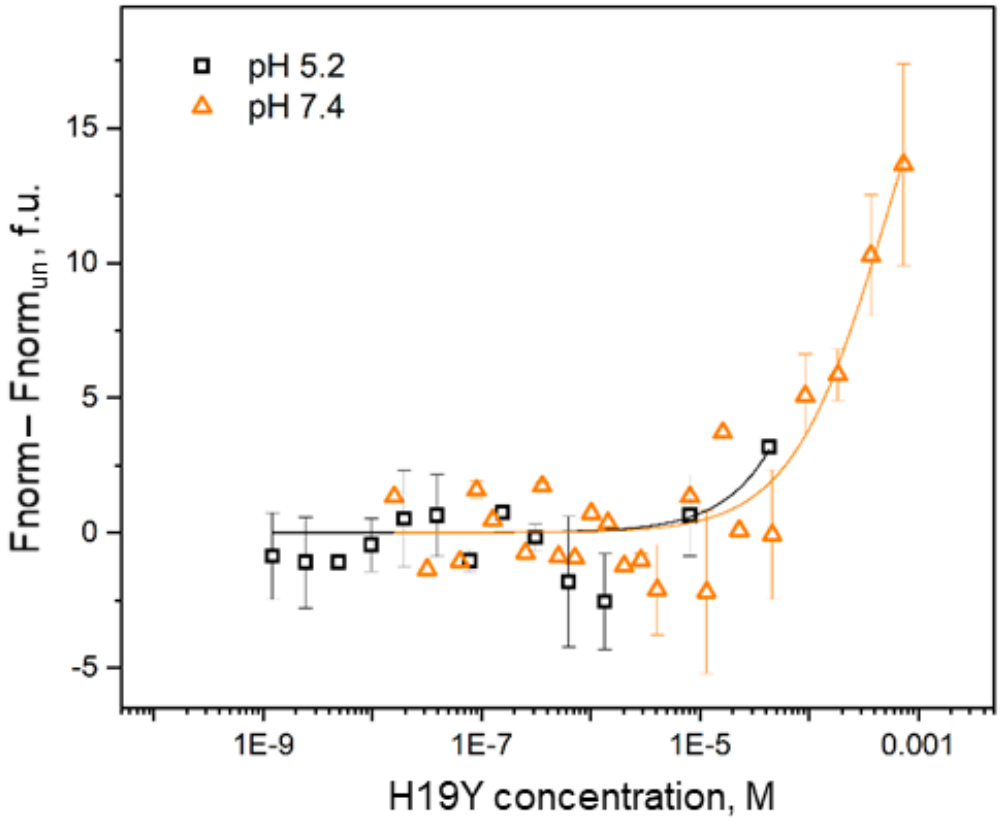

Figure S3. The MST data for the H19Y mutant were obtained at pH 5.2 and 7.4 and a temperature of 35 °C. The solid lines depict dose-response fits of the experimental data using the Hill function. The high-concentration plateau could not be reached due to induced inside the capillaries significant sample aggregation at millimolar concentrations, revealed by the bumps on the MST traces. This aggregation effect was more pronounced at pH 5.2. Our observations indicate that we did not observe H19Y dimerization or higher order oligomerization at the concentrations utilized in NMR experiments. Furthermore, the dimerization constant of the H19Y mutant is estimated to be ~ 0.5 ± 0.3 mM or greater at pH 7.4.

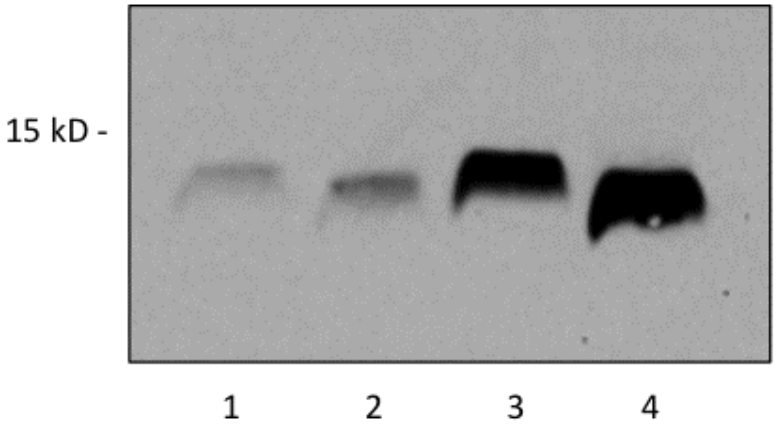

Figure S4. Western blot analysis of the PAI and PAI-H19Y amounts used for EMSA. PAI and PAI-H19Y were expressed at 30 ºC in *E. coli*. Total bacterial protein was run on a 15 % SDS polyacrylamide gel, blotted on a nitrocellulose membrane and hybridized with an anti-SB (R&D Systems, AF2798) polyclonal goat IgG primary antibody at a 1:5000 dilution for 2 h at room temperature in 1 % milk in TBST (Tris-buffered saline with 0.1 % Tween 20) followed by hybridization with a secondary rabbit anti-goat IgG antibody conjugated to horseradish peroxidase at a 1:20000 dilution for 30 min at room temperature in 1 % milk in TBST. Lane 1: 1 µg PAI; lane 2: 1 µg PAI-H19Y; lane 3: 5 µg PAI; lane 4: 5 µg PAI-H19Y.

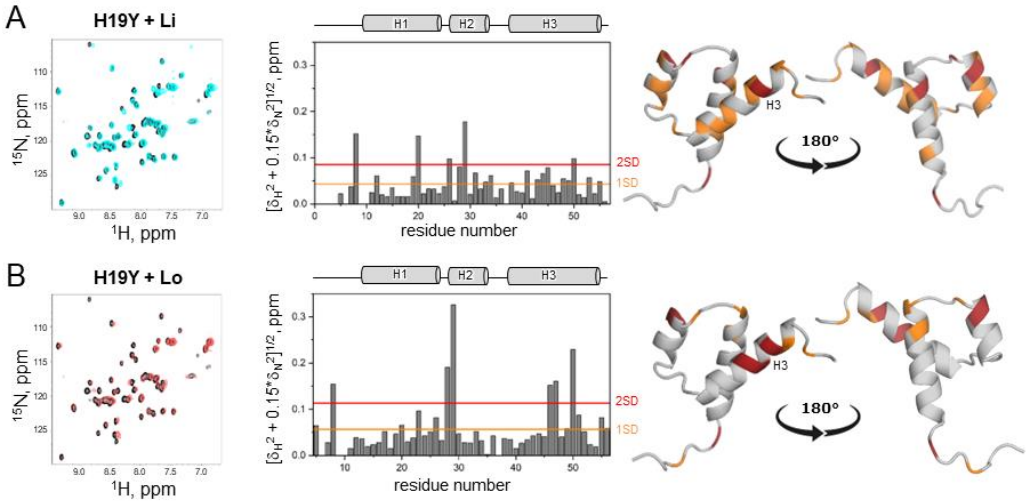

Figure S5. H19Y binding to the transposon Li and Lo sequences. (A) DNA binding to Li sequence. [^{1}H,^{15}N]-HSQC spectra of 0.085 mM ^{15}N,^{13}C-labeled H19Y is shown in the absence (black cross-peaks) and presence (cyan cross-peaks) of Li (1:5 molar ratio) collected at 35 °C in an aqueous solution of 25 mM sodium-phosphate buffer at pH 5.2. (C) DNA binding to Lo sequence. [^{1}H,^{15}N]-HSQC spectra at pH 5.2 of 0.085 mM ^{15}N,^{13}C-labeled H19Y is shown in the absence (black cross-peaks) and presence (red cross-peaks) of Lo (1:5 molar ratio) collected at the same conditions as Li.

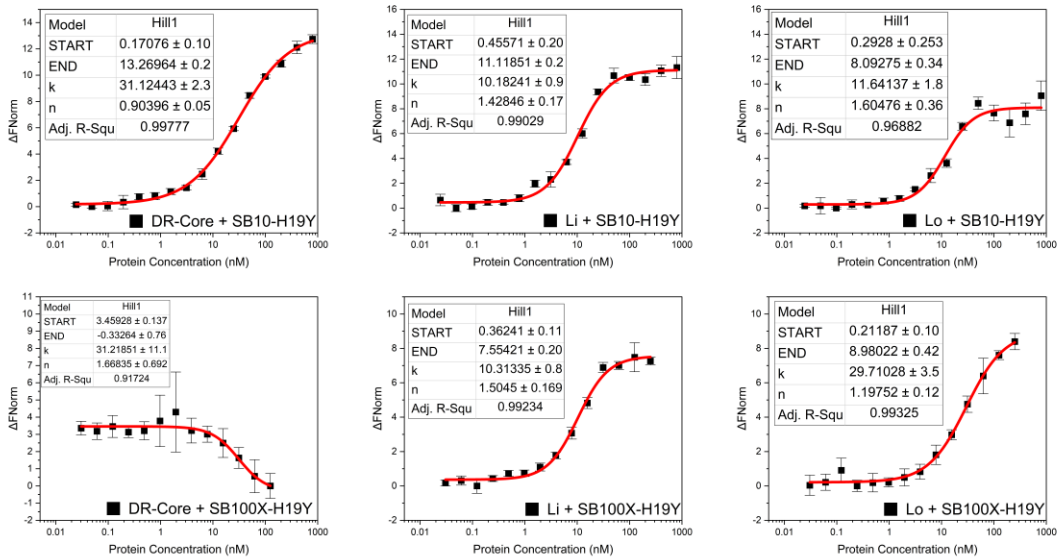

Figure S6. The MST data for the SB10-H19Y and SB100X-H19Y binding to DR-core, Li, and Lo. The data were evaluated over the T-jump time interval, e.g., within 1 s of IR-laser activation. Experimental error bars show S.E. for $n \geq 3$ separate experiments. The solid lines represent Hill fits to the experimental data.

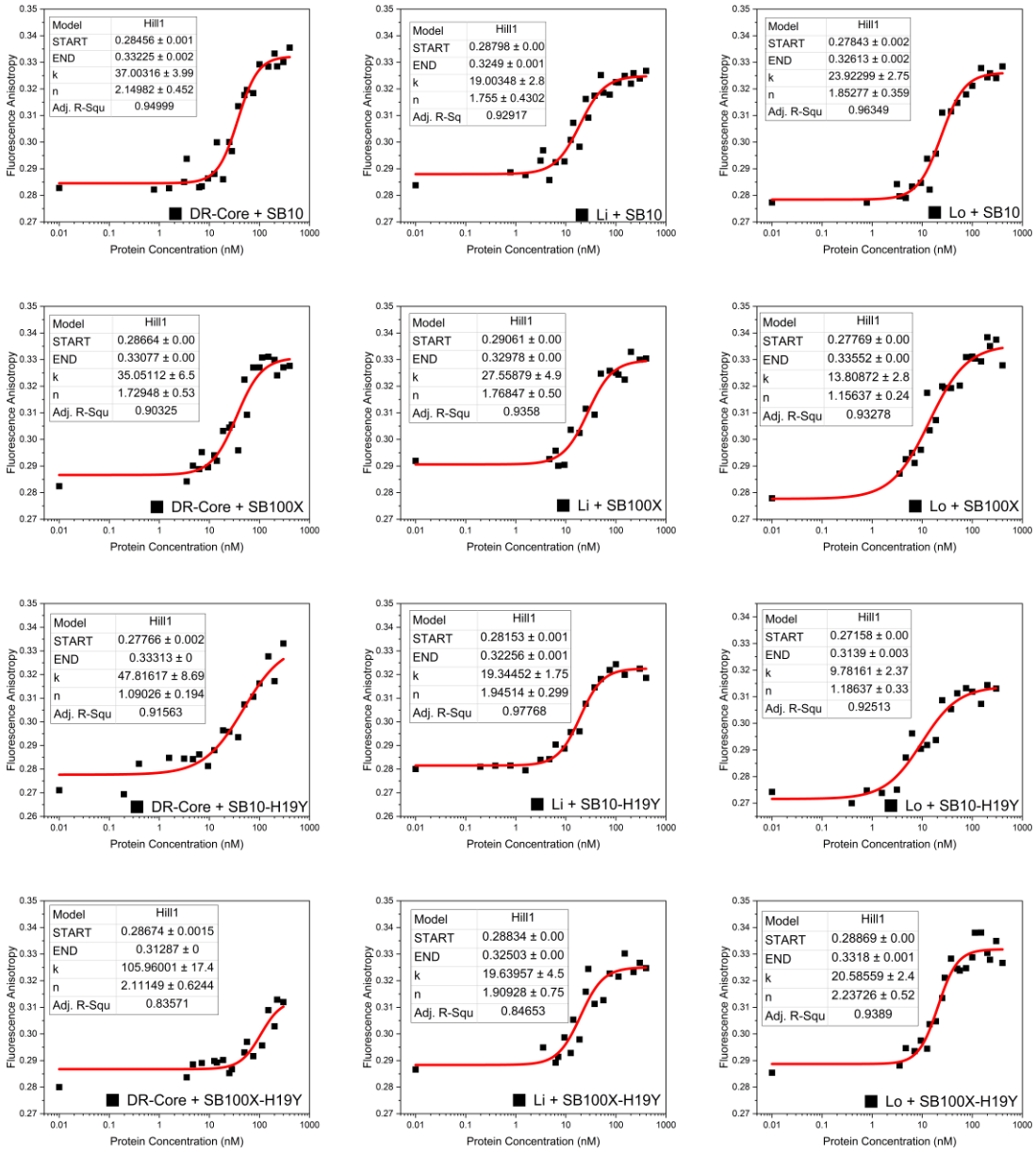

Figure S7. The fluorescence anisotropy data for SB10, SB10-H19Y, SB100X, and SB100X-H19Y full-length transposases binding to DNA. The solid lines represent dose-response fits of the experimental data using the Hill function.

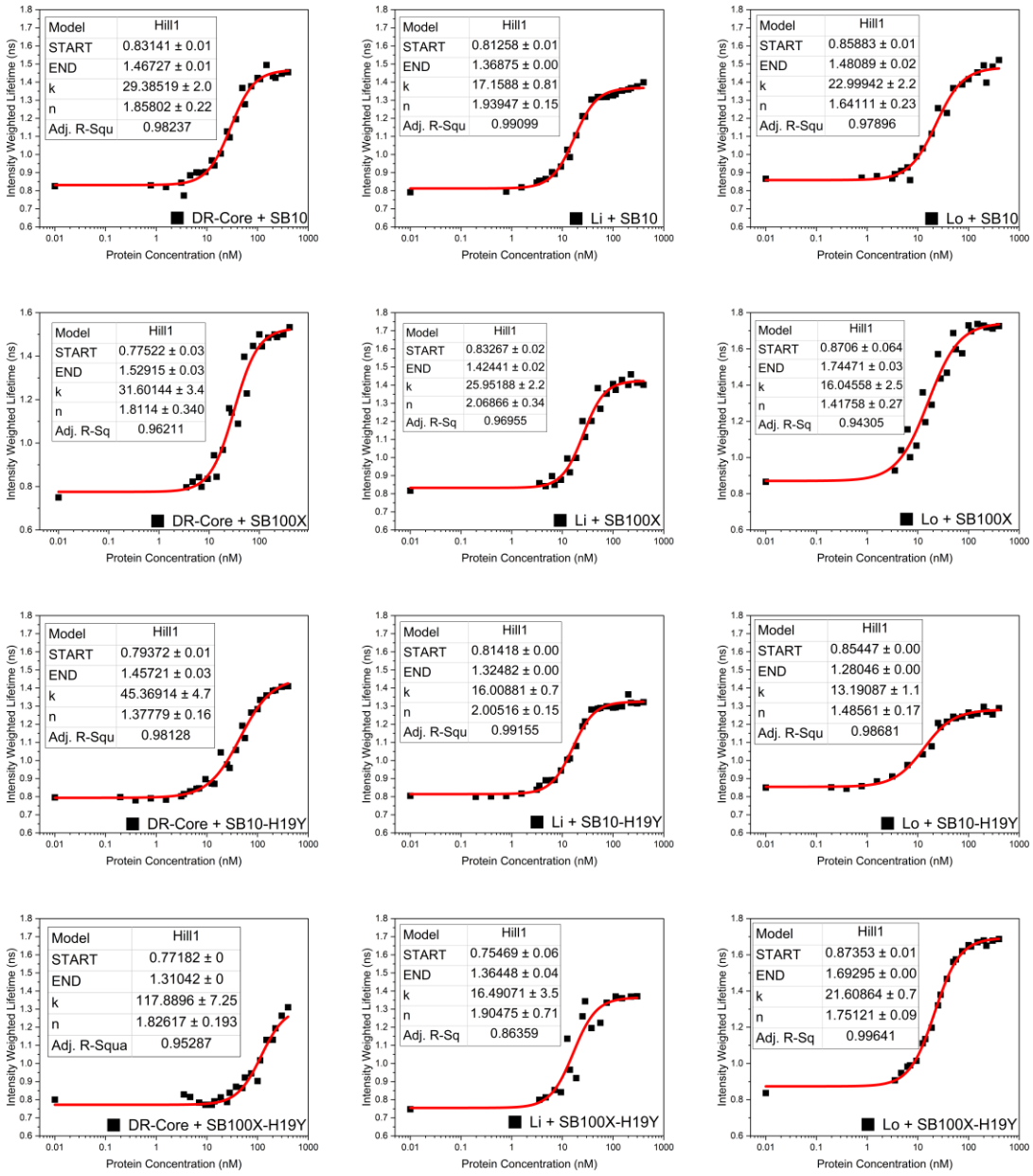

Figure S8. Representative fluorescence lifetime binding curves for SB10, SB10-H19Y, SB100X, and SB100X-H19Y full-length transposases binding to DNA. The solid lines represent dose-response fits of the experimental data using the Hill function.

APPENDIX C: SUPPLEMENTARY MATERIAL FOR CHAPTER 4

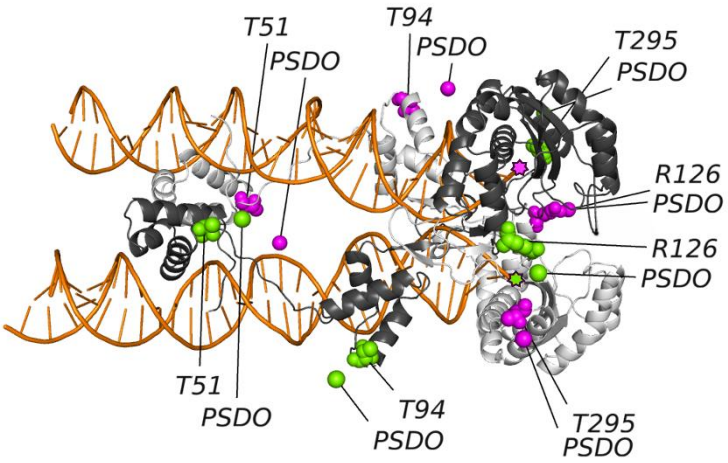

Figure S1 The average positions of fluorescent labels, represented here as pseudo atoms (PSDO), as calculated using CNS.

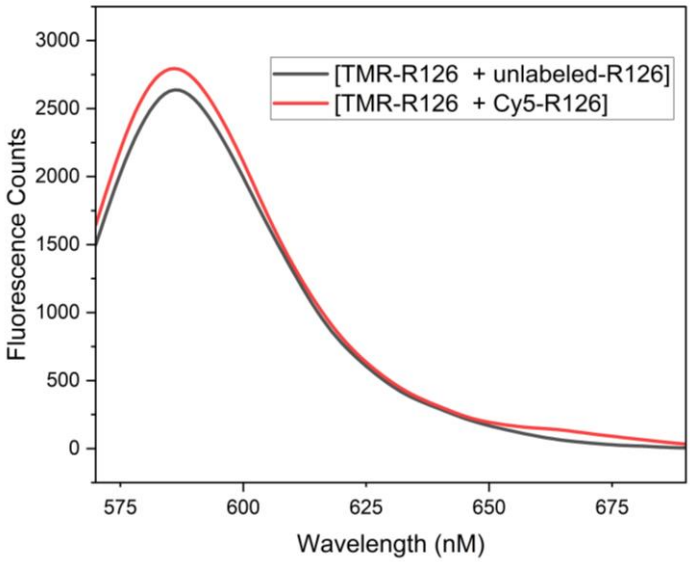

Figure S2: No FRET effect observed even at 5uM concentration of protein without DNA. The black curve represents TMR-R126 and unlabeled R126 added in 1:1 ratio with the final concentration of 5uM, while the red curve represents TMR-R126 and Cy5-R126 added in 1:1 ratio with the final concentration of 5uM

Table S1 SB and DNA Constructs used in this study:

T51C	MGKSKEISQDLRKRIVDLYKSGSSLGAISKRLAVPRSSVQTIVRKYKHHG**C**TQPSYRSGRRRVLSPRDERT LVRKVQINPRTTAKDLVKMLEETGTKVSISTVKRVLYRHNLKGHSARKKPLLQNRHKKARLRFATAHGD KDRTFWRNVLWSDETKIELFGHNDHRYVWRKKGEASKPKNTIPTVKHGGGSIMLWGCFAAGGTGALHK IDGIMDAVQYVDILKQHLKTSVRKLKLGRKWVFQHDNDPKHTSKVVAKWLKDNKVKVLEWPSQSPDLN PIENLWAELKKRVRARRPTNLTQLHQLCQEEWAKIHPNYCGKLVEGYPKRLTQVKQFKGNATKY
T94C	MGKSKEISQDLRKRIVDLYKSGSSLGAISKRLAVPRSSVQTIVRKYKHHGSTQPSYRSGRRRVLSPRDERT LVRKVQINPRTTAKDLVKMLEE**C**GTKVSISTVKRVLYRHNLKGHSARKKPLLQNRHKKARLRFATAHGD KDRTFWRNVLWSDETKIELFGHNDHRYVWRKKGEASKPKNTIPTVKHGGGSIMLWGCFAAGGTGALHK IDGIMDAVQYVDILKQHLKTSVRKLKLGRKWVFQHDNDPKHTSKVVAKWLKDNKVKVLEWPSQSPDLN PIENLWAELKKRVRARRPTNLTQLHQLCQEEWAKIHPNYCGKLVEGYPKRLTQVKQFKGNATKY
R126	MGKSKEISQDLRKRIVDLYKSGSSLGAISKRLAVPRSSVQTIVRKYKHHGSTQPSYRSGRRRVLSPRDERT LVRKVQINPRTTAKDLVKMLEETGTKVSISTVKRVLYRHNLKGHSARKKPLLQN**C**HKKARLRFATAHGD KDRTFWRNVLWSDETKIELFGHNDHRYVWRKKGEASKPKNTIPTVKHGGGSIMLWGCFAAGGTGALHK IDGIMDAVQYVDILKQHLKTSVRKLKLGRKWVFQHDNDPKHTSKVVAKWLKDNKVKVLEWPSQSPDLN PIENLWAELKKRVRARRPTNLTQLHQLCQEEWAKIHPNYCGKLVEGYPKRLTQVKQFKGNATKY
T295	MGKSKEISQDLRKRIVDLYKSGSSLGAISKRLAVPRSSVQTIVRKYKHHGSTQPSYRSGRRRVLSPRDERT LVRKVQINPRTTAKDLVKMLEETGTKVSISTVKRVLYRHNLKGHSARKKPLLQNRHKKARLRFATAHGD KDRTFWRNVLWSDETKIELFGHNDHRYVWRKKGEASKPKNTIPTVKHGGGSIMLWGCFAAGGTGALHK IDGIMDAVQYVDILKQHLKTSVRKLKLGRKWVFQHDNDPKHTSKVVAKWLKDNKVKVLEWPSQSPDLN PIENLWAELKKRVRARRPCNLTQLHQLCQEEWAKIHPNYCGKLVEGYPKRLTQVKQFKGNATKY
Lo DNA	CAGTTGAAGTCGGAAGTTTACATACACTTAAG

APPENDIX D: LIST OF DELIVERABLES

1. PUBLICATIONS

1.1. Peer-reviewed

1.1.1. Nesmelova, Irina V., Daria L. Melnikova, **Venkatesh Ranjan**, and Vladimir D. Skirda. "Translational diffusion of unfolded and intrinsically disordered proteins." Progress in Molecular Biology and Translational Science 166 (2019): 85-108.

1.1.2. Melnikova, Daria L., **Venkatesh V. Ranjan**, Yuri E. Nesmelov, Vladimir D. Skirda, and Irina V. Nesmelova. "Translational Diffusion and Self-Association of an Intrinsically Disordered Protein κ-Casein Using NMR with Ultra-High Pulsed-Field Gradient and Time-Resolved FRET." The Journal of Physical Chemistry B (2024).

1.2. Conference Paper

Venkatesh V. Ranjan, Yuri E. Nesmelov, and Irina V. Nesmelova. "DNA binding affinity and the assembly of a nucleoprotein complex by the sleeping beauty transposase." Biophysical Journal 123, no. 3 (2024): 503a.

1.3. Under review: **Venkatesh Ranjan**†, Gage O. Leighton†, Chenbo Yan, Maria Arango, Janna Lustig, Rosario I. Corona, Jun-Tao Guo, Zoltán Ivics, Irina V. Nesmelova. “DNA Binding of The Sleeping Beauty Transposase.” †=equal contribution. [Manuscript under review with Nucleic Acid Research]

1.4. Manuscript in preparation: “FRET-based structural model of the Sleeping Beauty paired-end complex.”

2. AWARDS, GRANTS AND FELLOWSHIPS

2.1. Travel Grant for Spring 2024 from the Graduate Professional and Student Government, UNC Charlotte.

2.2. Graduate School Summer Fellowship, Summer 2022.

2.3. STEM Communication Fellowship funded by the Burroughs-Wellcome fund for academic year 2020-2021

2.4. 2024 Graduate Teaching Award from the Department of Chemistry, UNC Charlotte.

3. POSTERS AND PRESENTATIONS

3.1. BPS 2024 Annual Meeting, Philadelphia, PA. “DNA binding affinity and the assembly of a nucleoprotein complex by the sleeping beauty transposase.”

3.2. Graduate Research Symposium, 2023, UNC Charlotte. “Structure of the SB Transpososome”