**University of South Bohemia in České Budějovice**
**Faculty of Science**

# The impact of DNA methylation on the stability and dynamics of i-motif in the ILPR promoter region

Bachelor thesis

**Eva Sýkorová**

Tutor: Mgr. Tomáš Fessl, Ph.D.

České Budějovice 2016

Sýkorová, E., 2016: The impact of DNA methylation on the stability and dynamics of i-motif in the ILPR promoter region. Bc. Thesis, in English. – 64 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

**Annotation:**

The stability of i-motif structure in the promoter of the insulin gene was investigated by circular dichroism and single-molecule fluorescence resonance energy transfer. The main factor studied was the cytosine methylation and its location in the DNA sequence. Also the molecular crowding effect was examined. In addition, i-motif folding kinetics were investigated by the single-molecule method.

**Key words:**

i-motif, cytosine methylation, ILPR, DNA folding, single-molecule FRET

**Affirmation:**

I hereby declare that I have worked on my bachelor thesis independently and used only the sources listed in the bibliography.

I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my bachelor thesis, in full in electronic form in publicly accessible part of the STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages.

Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defense in accordance with aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

České Budějovice ……………………..

April 21, 2016 Eva Sýkorová

**Acknowledgement:**

I would like to thank my tutor Mgr. Tomas Fessl, Ph.D. not only for his beneficial guidance and shared knowledge but also for introducing me the exciting side of science.

My thanks go also to Dr. Łukasz Bujak for inspiring discussions.

Further, I thank RNDr. Radek Litvín, Ph.D. and prof. RNDr. František Vácha, Ph.D. for financial support of my research.

Great thanks belong also to my family for their patience and support.

# 1 CONTENTS

# 2 INTRODUCTION

In this study, we wanted to elucidate the regulation process of insulin transcription using circular dichroism (CD) in combination with an established single-molecule method: fluorescence resonance energy transfer (FRET). Specifically, we focused on the still unclear conditions of-i-motif pH stability and folding dynamics in the insulin-linked polymorphic region (ILPR). The main factor studied was DNA methylation. In addition, the effect of molecular crowding was examined.

Because insulin is an essential hormone in the human body, dysfunctions in its production lead to serious illnesses [1]. Our contribution to the complete cognition of the insulin transcription regulatory process could therefore facilitate the development of treatment for insulin dependent diseases.

Fundamental biological and chemical aspects related to our investigation are provided further in this section.

## 2.1 Insulin

Insulin is a protein hormone produced by the pancreatic β-cells which are located in the Islets of Langerhans. The insulin molecule consists of two polypeptide chains, A and B, with two connecting disulfide bonds. It is synthesized into its final form from its precursor structure – proinsulin. In this precursor, the two peptide chains are joined continuously by a so-called C-peptide which is then cleaved to produce the insulin itself and the free C-peptide. This binding unit further functions as an indicator of the insulin production level during medical blood tests [2].

The importance of the insulin hormone lies in its irreplaceable role in glucose metabolism. The hormone enables cellular uptake and regulates the glucose blood level. It is also a regulator of carbohydrate, lipid and protein metabolisms [3].

Since glucose is the main source of energy in living organisms, dysfunctions in its metabolism can result in serious diseases. Both, the defects in production of insulin and resistance to it, lead to diabetes mellitus. The first defect causes diabetes mellitus type I (also known as insulin-dependent diabetes mellitus IDDM) while both of them inflict diabetes mellitus type II [4,5].

## 2.2 Insulin gene and insulin-linked polymorphic region

The insulin gene (INS) is located on the short arm of chromosome 11 (cytogenetic location: 11p15.5; molecular location: base pairs 2,159,779 to 2,161,209) [6].

An insulin linked polymorphic region (ILPR) was already shown to be directly related to insulin transcription and thus to IDDM. Its location on the INS is 363 base-pairs upstream from the transcription start site. The ILPR consists of a varying number of repetitive sequences: 5' - ACA GGGG TGT GGGG ACA GGGG TGT GGGG - 3' and complementary 3' - TGT CCCC ACA CCCC TGT CCCC ACA CCCC - 5' respectively [7].

The most common ILPR contains 14 variants of the slightly differing above mentioned sequence. This makes the region one of the most polymorphic guanine-rich or cytosine-rich areas in the human genome. Variants of C-rich sequences of this ILPR are shown in **Tab. 1**. The length polymorphism of the ILPR is considered to be one of the regulatory factors in insulin transcription and therefore is suspected of being responsible for the IDDM [6,7,8].

**Table 1**. Fourteen consecutive sequences occurring in the most common version of the ILPR with highlighted differences from the basic sequence of 5' – TGTCCCCAGACCCCTGTCCCCACACCCC – 3'; taken from [6].

| No. | Sequence | Number of nucleotides |
|---|---|---|
| 1 | TGTCCCCAGACCCCTGTCCCCAC-ACCCCTGT | 28 |
| 2 | TGTCCCCACACCCCTGTCCCCAC-ACCCCT**A**T | 28 |
| 3 | T**A**TCCCCACACCCCTGTCCCCA**GG**ACCCCTGT | 29 |
| 4 | TGTCCCCACACCCCTGTCCCCA**G-**ACCCCTGT | 28 |
| 5 | TGTCCCCACACCCCTGTCCCCAC-ACCCCTGT | 28 |
| 6 | TGTCCCCACACCCCTGTCCCC**CGG**ACCCCTGT | 29 |
| 7 | TGTCCCCA**G**ACCCCTGTCCCCAC-ACCCCTGT | 28 |
| 8 | TGTCCCCACACCCCTGTCCCCAC-ACCCCTGT | 28 |
| 9 | TGTCCCCACACCCCTGTCC**T**CA**G**-ACCCCTGT | 28 |
| 10 | TGTCCCCAGACCCCTGTCCCCA**GG**ACCCCTGT | 29 |
| 11 | TGTCCCCACACCCCTGTCCCCAC-ACCCCTGT | 28 |
| 12 | TGTCCCCACACCCCTGTCCCCA**GG**ACCCCTGT | 29 |
| 13 | TGTCCCCA**G**ACCCCTGTCCCCA**GG**ACCCCTGT | 29 |
| 14 | TGTCCCCACACCCCT-TCCCCA**GG**ACCCCTGT | 28 |

## 2.3 Secondary structures in ILPR

Nucleic acid secondary structures can have a regulatory function. In a gene promoter, certain structural elements are required for interaction with specific proteins that activate or stop gene expression [9].

G-quadruplexes are nucleic acid secondary structures generally found on G-rich strands, formed by guanine tetrads and stabilized by Hoogsteen hydrogen bonds [10]. Such structures can fold under physiological conditions on the ILPR G-rich sequences. Further, it was shown that the G-quadruplexes are the insulin regulating agents *in vitro* and the same function is highly presumed to occur in human cells [7,8].

Much attention has been paid to the ILPR G-rich strand in the past ten years while the complementary strand has been largely omitted in ILPR structure investigations. Here we focus on this C-rich strand, its secondary structure folding and stabilization in simulated intracellular conditions and thus its possible role in transcription.

As on the G-rich strands, the complementary C-rich strands can also form a tetraplex structure. This structure is called an i-motif and up to now there is far less known about its behavior and potential role *in vivo* or *in vitro* than its guanine analogy.

This secondary DNA structure was first described in 1993 [11]. Fundamental elements of the i-motif are hemi-protonated cytosine-cytosine$^+$ (C-C$^+$) base-pairs intercalated in a tetrameric structure. It is the only known DNA structure with intercalated base-pairs and the name i-motif was derived from this term.

The i-motif can be theoretically formed by one or more independent nucleic acid C-rich strands. However, the number of strands has no effect on the spatial arrangement of the cytosine tracks: two parallel strands form a duplex and two of these duplexes are stacked together in an anti-parallel direction [12].

In our case, 16 cytosines of one strand are involved in the i-motif structure, forming eight C-C$^+$ bonds in two vertical planes as can be seen in **Fig. 1**. The C-C$^+$ base-pair consists of three hydrogen bonds. The proton involved in the middle bond is shared by two nitrogen atoms from opposite cytosines, both at position 3 (N3). It was shown, that the N3…H$^+$…N3 bond is not symmetrical. The proton oscillates between the two nitrogen atoms with asymmetric double-well potential [13]. A scheme of the three hydrogen bonds in the C-C$^+$ base-pair together with the oscillating N3…H$^+$…N3 bond is shown in **Fig. 2**.

Two conformers have been described according to the way the cytosines interact. In the first structure, the outer hemi-protonated base pair involves the cytosine at the 3’ end of the DNA sequence; this conformer is then called a 3’E form. The second structure, with the outer base-pair opposite at the 5’ end, is called a 5’E form. It was shown that these two structures are able to convert to each other [14,15]. Both of the conformers are shown in **Fig. 1**.

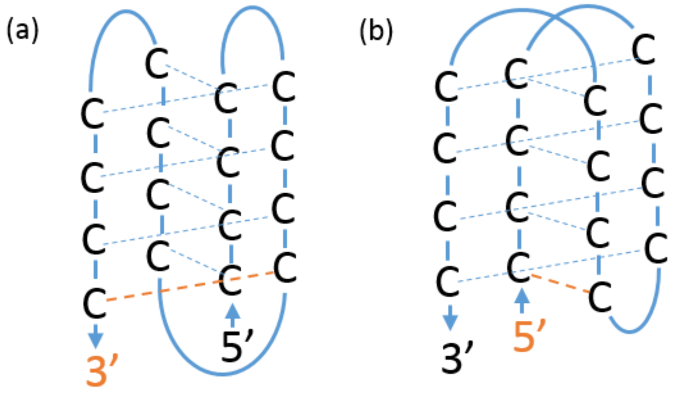


**Figure 1**. Two conformers of an i-motif structure with highlighted terminal C-C$^+$ base-pairs (orange scatter line); (a) 3’E form with the outer base-pair on the 3’ end and (b) 5’E form on the 5’ end of the DNA strand.

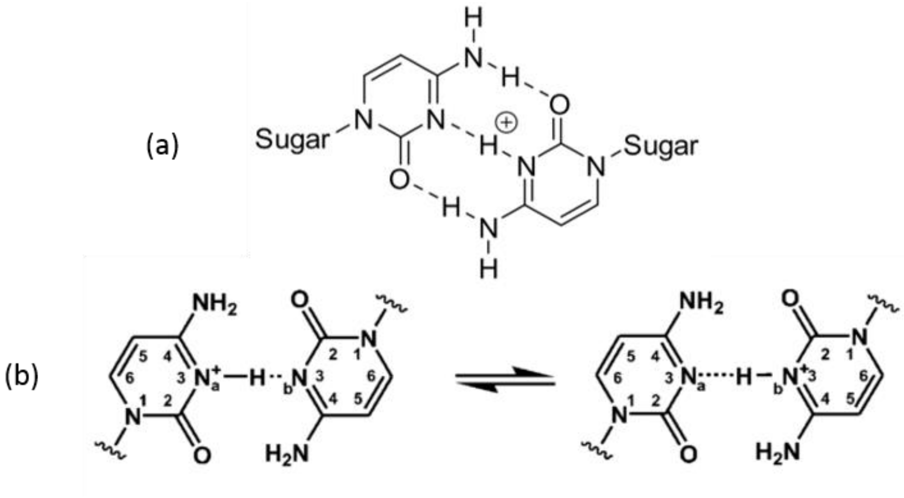


**Figure 2**. (a) Scheme of a cytosine-cytosine$^+$ base-pair with its three hydrogen bonds and indicated bonds to sugars in the nucleotides, taken from [14]; (b) scheme of an asymmetrical oscillating hydrogen bond between nitrogen atoms (both at position 3 in the pyrimidine circle) in the hemi protonated cytosine-cytosine$^+$ base-pair [13].

## 2.4 I-motif stability

Since the C-C$^+$ bond typical for the i-motif is protonated, pH has a great influence on its formation. The value of cytosine pKa varies with the solution, but it is generally slightly acidic (pH 4.6 in pure water at 25°C). As a consequence, acidic conditions stabilize while neutral or basic conditions destabilize the i-motif structure [12].

Apart from the fully folded i-motif structures, various intermediates such as triplexes exist at probably neutral or slightly acidic conditions (pH 6-7) [12]. These partially folded structures are currently being widely investigated in i-motif folding dynamics studies using single-molecule methods [16,17]. The partially unfolded triplex structures are suspected to be mechanically and thermodynamically stable enough to influence transcription. In this study, we looked in detail on the intermediate structures and their dynamics using the single-molecule method FRET.

Another i-motif stability factor is the number of C-C$^+$ base-pairs in the structure. Also the loops between the C-rich regions in the DNA strand, their length and composition, influence the stabilization of i-motif [12]. A classification has been made distinguishing two classes of i-motives: Class I with shorter loops (2-4 bases) and Class II with longer loops (up to eight bases). The Class I i-motives (including our structure) are described as less stable than Class II [18].

Ionic strength of the solution also has a noticeable influence on the i-motif stability. Nucleic acid structures generally are negatively charged at neutral pH induced by phosphate residues. Positively charged ions in solution can therefore compensate the anion character of the structure and stabilize it. On the other hand, the protonation specific in i-motives produces a positive charge which causes repulsion and destabilization in the presence of cations. Hence, ionic strength can cause conformational and stability changes in the i-motif structures [12].

In the past, due to its behavior, i-motif presence in a physiological environment with pH around 7.4 was thought to be improbable and its biological role was not considered. Recently, it was found that stabilizing effects increase the pH at which an i-motif can be folded, such as epigenetic modification of cytosines [19], the presence of certain cosolutes [20] or molecular crowding [19,21]. Here we focused on the effects of epigenetic modification and molecular crowding.

## 2.5 Cytosine methylation

Epigenetic processes have a regulatory role in human organism. They influence gene expressions and can result in heritable changes in the human genome. In general, various types of chemical modification are known to have an epigenetic function. In this study, we focused only on the effect of DNA methylation, specifically on the methylation of cytosines [22,23].

5-Methylcytosines, with the methyl group attached to the fifth atom of the base (shortly 5'mCs) (**Fig. 3**), are very common epigenetic tools spread over a wide range of eukaryotes including the human genome. It was first introduced as a gene activity controller in 1975 [24].

Two sites of cytosine methylation are distinguished. In the past, the 5'mCs occurring in a DNA sequence after a guanine nucleotide (such dinucleotides are called CpGs, meaning cytosine – phosphate – guanine) were studied more frequently and their function as transcription inhibitors is known [25]. CG islands (CGIs), short sequences containing many repetitions of CpGs and thus prone to cytosine methylation, are found mainly in gene promoters or close to the transcription start sites [26]. Later studies found non-CpG methylations where the 5'mC stands in a sequence after adenine (CpA), thymine (CpT) or another cytosine (CpC) nucleotide. In 1987, it was shown that the majority of 5'mCs in the human genome occurs in non-CpG dinucleotides [27]. Even though, the number of non-CpG studies rapidly increased until recently and the topology and occurrence of these methylations in the human genome still has to be properly mapped.

In this study, we assumed that non-CpG methylation can naturally occur in the ILPR. We further examined the potential influence of its presence and location on the i-motif folding and stability.

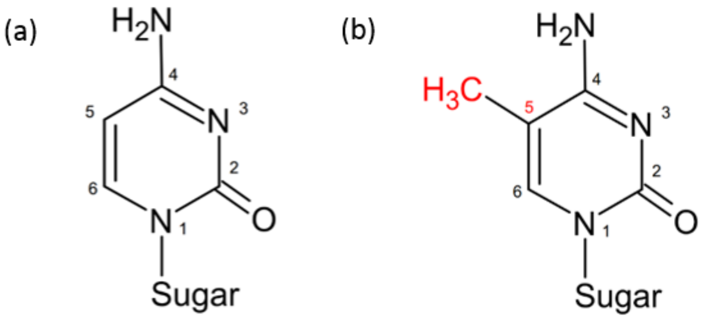


**Figure 3**. Non-modified and methylated cytosines with indicated bonds to sugar molecules and with numbered atoms of the main circle; (a) cytosine; (b) 5-methylcytosine; created in ACD/ChemSketch [28].

## 2.6 Molecular crowding

Molecular crowding is natural in cells. The intracellular environment contains 5 – 40 % mass of macromolecules, such as proteins, nucleic acids and complex sugars [29,30]. Thus, this factor has to be considered along with pH or ionic concentration when simulating physiological conditions.

The effect of molecular crowding has a great influence on the stability of proteins and nucleic acids [30]. The molecular crowding effect has been investigated on various DNA structures, such as small and large DNA duplexes [31], triplexes [32] or G-quadruplexes [33]. In [33], macromolecular crowding induced the formation of the G-quadruplex structure and stabilized it. I-motif stabilization by molecular crowding was observed on the c-MYC gene [34].

We used polyethylene glycol (PEG) with 8000 subunits to mimic intracellular conditions and thus to examine the impact of molecular crowding on the i-motif stability. PEG is the most common molecular crowding inducer in *in vitro* experiments. However, recent study warned that PEG does not properly mimic the intracellular environment and does not function as a crowding agent for human telomere G-quadruplex [35].

Nevertheless, it is out of the scope of this work to access direct contribution of PEG to structural change in the ILPR i-motif. For later use, we want to develop single-molecule experiments in more reliable physiological conditions, such as in lysate and ultimately *in cell*.

# 3 MATERIAL AND METHODS

## 3.1 DNA sequence

In all measurements we used a 31-base-long DNA sequence: 5`- TGT CCCC ACA CCCC TGT CCCC ACA CCCC TGT - 3`. For our purposes, this sequence was modified by cytosine methylation at different positions. In each of the five modified sequences, there were four 5'mCs located on formerly designed sites. One non-methylated sequence was used as a control. Terms **$C_0$**, **$C_{1\text{-}4}$**, **$C_{2,3}$**, **$C_{1,4}$**, **$C'_{1,4}$** and **$C'_1$** are further used for the samples. They were designed to mark the positions of methylations in the fourth C-tetrad in the sequence. The two samples described as **$C'_1$** and **$C'_{1,4}$** contain methylations also in the third C-tetrad. **$C_0$** is the control. All of the sample sequences with highlighted locations of the 5'mCs and with our designation are shown in **Tab. 2**. A 3D visualization of the 5mC's in each sequence folded into the i-motif is shown in **Supplementary mat. I**.

We chose the number of methylated cytosines (four in each modification) in the expectation of noticeable changes in the i-motif stability while, on the other hand, not exceeding its production availability. The exact positions of the methylations were selected to cover as many symmetrical variations as possible with only a few samples. All of the DNA samples were supplied by Integrated DNA Technologies, Inc.

**Table 2**. DNA sequences with cyanine dyes at the 3' and 5' ends and with highlighted (underlined) positions of the methylations; our designation is shown on the right (indexes 1 - 4 mark the positions of 5'mCs in the fourth C-tetrad (red text), sequences containing methylations in the third C-tetrad (green text) are described with **C'**).

| | |
|---|---|
| 5`-/Cy5/ TGT CCCC ACA CCCC TGT CCCC ACA CCCC TGT /Cy3/-3` | **$C_0$** |
| 5`-/Cy5/ TGT CCCC ACA CCCC TGT CCCC ACA CCCC TGT /Cy3/-3` | **$C_{1\text{-}4}$** |
| 5`-/Cy5/ TGT CCCC ACA CCCC TGT CCCC ACA CCCC TGT /Cy3/-3` | **$C_{1,4}$** |
| 5`-/Cy5/ TGT CCCC ACA CCCC TGT CCCC ACA CCCC TGT /Cy3/-3` | **$C_{2,3}$** |
| 5`-/Cy5/ TGT CCCC ACA CCCC TGT CCCC ACA CCCC TGT /Cy3/-3` | **$C'_{1,4}$** |
| 5`-/Cy5/ TGT CCCC ACA CCCC TGT CCCC ACA CCCC TGT /Cy3/-3` | **$C'_1$** |

## 3.2 Dyes

All of the samples were labeled for the single molecule experiments. We chose a well-established single-molecule FRET pair, utilizing cyanine dyes Cy3 (as a donor) and Cy5 (as an acceptor). They are widely used because of their convenient Förster distance (56 Å), low tendency for photo-induced blinking and quenching due to interaction with biomacromolecules, and for good water solubility and separation of the donor and acceptor emission spectra [36,37].

The dyes were located terminally: Cy5 on the 5' end and Cy3 on the 3' end of the sequence (**Tab.2**). Both were attached by a flexible linker, a single-bonded carbohydrate chain, to allow them a free motion in all directions. Cy3 and Cy5 properties, including their emission and absorption spectra and their formula, were provided by the supplier (Integrated DNA Technologies, Inc.) and are shown in **Tab. 3** and **Fig. 4**. The spectra are shown in **Supplementary mat. II**. A 3D model of the non-modified sample sequence folded into the i-motif with attached cyanine dyes is shown in **Supplementary mat. III**.

**Table 3**. Properties of Cy3 and Cy5, provided by Integrated DNA Technologies, Inc.

| **Dye** | **Absorbance max. [nm]** | **Emission max. [nm]** | **Extinction coefficient (at absorbance max.) [mol$^{-1}$cm$^{-1}$]** | **Quantum yield** |
|---|---|---|---|---|
| Cy3 | 550 | 564 | 136000 | 0.15 |
| Cy5 | 648 | 668 | 250000 | 0.30 |

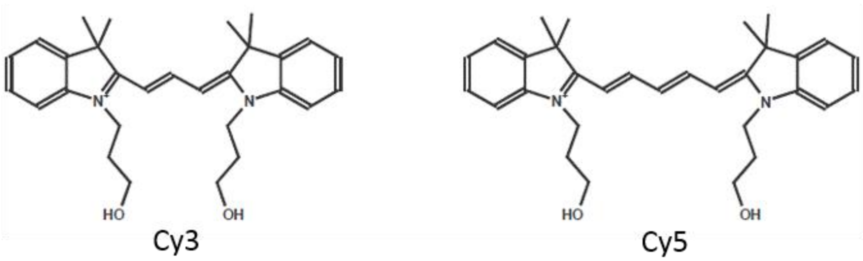


**Figure 4**. Formulas of Cy3 and Cy5, provided by Integrated DNA Technologies, Inc.

## 3.3 Buffer

All samples were dissolved in BPES buffer (6 mM $Na_2HPO_4.2H_2O$, 4 mM $NaH_2PO_4.H_2O$) with 140 mM concentration of $K^+$ ions (KCl added to the buffer) to simulate physiological conditions. The 12 mM concentration of $Na^+$ ions naturally occurring in human cells was already provided by the buffer composition itself. The buffer pH was adjusted by adding a negligible amount of 10 M NaOH or 37 wt.% HCl to generate a pH scale from 6.5 to 7.4, with 0.1 step. The effect of molecular crowding was simulated by addition of up to 50 wt. % polyethylene glycol (average mol. wt. 8000).

## 3.4 Circular dichroism

### 3.4.1 Principle

The principle of the CD method has been described in detail elsewhere [38,39]. Briefly, the measured signal is the difference in the absorption of left and right handed circularly polarized light by an optically active molecule (**Eq. 1**) [38].

$$CD = A_l - A_r \qquad \textbf{(1)}$$

$A_r$ represents absorbance (the amount of absorbed light) of right handed and $A_l$ the absorbance of left handed circularly polarized light.

A beam of light is polarized circularly when its electric and magnetic field vectors change their direction by rotating to the left or to the right in the way of propagation waves while having a constant magnitude [38].

A chiral, also called optically active, molecule is an asymmetrical object lacking a plane or a center of symmetry that cannot be identified with its mirror image. Two such non-superimposable mirror-image molecules are referred to as enantiomers.

The dimensionless absorbance is defined in **Eq. 2**, based on the Beer-Lambert law [39].

$$A = log_{10}(I_0/I) \qquad \textbf{(2)}$$

$I_0$ is the intensity of entering and $I$ as the intensity of outgoing light.

### 3.4.2 CD in nucleic acids

Nucleotides (monomer units of nucleic acids consisting of a base, a five-carbon sugar of either ribose or deoxyribose, and a phosphate group) yield CD spectra even though the bases, which are the chromophores of DNA and RNA (these absorb light and are responsible for the color of the molecule), are not chiral. This phenomenon occurs via the presence of sugar molecules in nucleotides inducing an asymmetry to the attached bases and therefore the ability of yielding a CD spectrum [39].

Further, the base-base stacking interactions frequently occurring in DNA and RNA secondary structures provide intense CD spectra which makes this technique extremely sensitive for nucleic acid investigations. However, CD measurements do not provide information on the atomic level and are only used for secondary structure studies [39,40].

The CD application in our investigation was to identify which secondary structure is adopted by the sample sequence under different pH and molecular crowding conditions. Such information is easily accessible since every DNA secondary structure has its own CD signature. This is caused by differently oriented transition dipole moments in the investigated molecules [39].

An electric dipole transition moment is described as a combination of charge distribution in the ground and excited states. It has a vector character and its direction determines the direction of light polarization at which the absorption is maximal. CD or absorption spectra of nucleic acids fall within the UV range, from approximately 300 nm to 180 nm, where the absorption of buffer and UV-induced radicals begins. The transition dipole moments responsible for the spectra are found in the bases while the rest of the nucleotides (sugars and phosphates) absorb mainly under 190 nm. Two types of transition vectors are allowed in the bases. Bonding π electrons have transitions in the plane of the base ($\pi$-$\pi^*$). For our measurements more important $n$-$\pi^*$ transitions are perpendicular with the plane and are represented by excitation of the oxygen or nitrogen non-bonding electrons. These are important when there are base-base interactions in alternative structural conformations, such as an i-motif. Therefore, the CD spectra yielded by an i-motif or by any other secondary structure element are unique [9]. See **Fig. 5** for the spectra of an i-motif and an unfolded DNA.

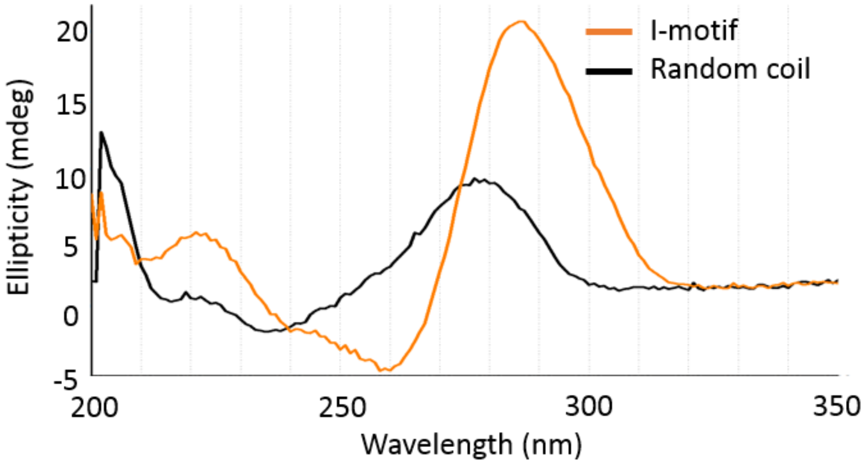


**Figure 5**. Typical CD spectra of a randomly coiled unfolded DNA strand (black line) and an i-motif (orange line); both spectra were measured on the non-modified sample at pH 6.5 (random coil) and 7.4 (i-motif).

### 3.4.3 Practical instrumentation

We measured CD on the Jasco J-715 spectropolarimeter (JASCO Corporation, 1995) with the parameter settings shown in **Supplementary mat. IV**. Each of the sample modifications was measured twice over the whole pH scale from 6.5 to 7.4, at 0.1 increments, first without any molecular crowding agent, and secondly in 50 wt. % PEG. The concentrations of all samples in the CD measurements ranged from 1.2 to 1.5 μM.

### 3.4.4 Data analysis

We analyzed the CD data to obtain a pH value at which 50 % of the present DNA molecules were folded into an i-motif, noted as $pH_{50}$. Two analytical methods were used to calculate this value. The first one, a two-state model analysis, takes a reaction path as a linear combination of starting and resulting states as is expressed in **Eq. 3** [41].

$$I(pH,\lambda) = c^0(pH) \cdot I(6{,}5;\lambda) + c^1(pH) \cdot I(7{,}4;\lambda) \qquad \textbf{(3)}$$

Matrix $I(pH,\lambda)$ represents already reconstructed CD data depending on pH and wavelenght, $I(6{,}5;\lambda)$ and $I(7{,}4;\lambda)$ are the CD spectra measured at pH 6.5 and 7.4 while $c^0(pH)$ and $c^1(pH)$ are relative concentrations of the initial and final states at exact pH. **Eq. 3** in matrix notation is expressed in **Eq. 4** [41].

$$I = C \cdot I^{0\infty} \qquad \textbf{(4)}$$

The relative concentrations of each state are calculated by using a linear least-square (LSQR) procedure (**Eq. 5**) [42].

$$C = I \cdot I^{0\infty T} \cdot (I^{0\infty} \cdot I^{0\infty T})^{-1} \qquad \mathbf{(5)}$$

A transposed matrix is denoted by $^{T}$.

The $pH_{50}$ values were obtained by fitting these relative concentrations to a dose-response function.

The second method of analysis is to derive the $pH_{50}$ values from the position of maxima. This puts the wavelengths of the maxima of each CD spectrum of a given sample modification measurement into one graph with the $pH_{50}$ taken as the inflex point of the sigmoid function fitted to it [43].

All data analyses were done either in home-built Python software [44] or in Origin, an established graphing and data-analysis software [45].

## 3.5 Fluorescence resonance energy transfer

### 3.5.1 Principle

FRET is a widely used technique in single molecule studies. The FRET process was first theoretically described in 1948 [46]. Briefly, a non-radiative transfer of energy from a donor molecule in an excited state to an acceptor molecule in the grounded state can occur over distances of 10 – 100 Å whenever their emission and absorption spectra overlap. No photons participate in the process; it proceeds via long-range dipole–dipole interactions of the two molecules involved [47,48].

**Eq. 6** shows the dependence of the energy transfer efficiency ($E_{FRET}$) on the donor and acceptor distance $r$ [49].

$$E_{FRET}(r) = \frac{1}{1+(\frac{r}{R_0})^6} \qquad \mathbf{(6)}$$

$R_0$ is a Förster distance at which the acceptor absorbs 50 % of the donor excitation energy.

The $E_{FRET}$ dependence on distance allows us to measure the lengths between donors and acceptors and therefore between two labeled locations on a macromolecule, which provides a structural information [48].

However, **Eq. 6** is a simplification, because the Förster distance is a function of the medium's refractive index $n$, the donor and acceptor spectral overlap integral $J$, orientation of the two molecules described by factor $\kappa^2$ and the donor's non-FRET quantum efficiency $Q_D$. This dependence is expressed in **Eq. 7** [47].

$$R_0^6 = \frac{9000(\ln 10)\kappa^2 Q_D J}{128\pi^5 n^4 N_A} \qquad \textbf{(7)}$$

$N_A$ is Avogadro's number.

The value of $Q_D$ is a known property of the donor molecule, in our case $Q_{Cy3} = 0{,}15$ (provided by the supplier). The spectral overlap integral $J$ is calculated from the known spectra of the donor and acceptor (shown in **Supplementary mat. II**). The orientation factor $\kappa^2$ describes the relative orientation of a donor and an acceptor. We designed the donor (Cy3) and acceptor (Cy5) dyes to be attached by flexible single bonded hydrocarbon chains at the ends of our sample DNA sequences so they could rotate freely in all directions. In such cases the orientation factor is generally taken as $\kappa^2$= 2/3 [48].

Because we were able to calculate the Förster distance (as described above), we could obtain the donor and acceptor distance by measuring only the E<sub>FRET</sub> (**Eq. 6**). Several ways exist for how to obtain the energy transfer efficiency. The one used in this study measures the intensities of the donor $I_D$ and acceptor $I_A$ emissions. In this method the observed efficiency is called a relative proximity ratio $E_{PR}$ (**Eq. 8**) [49].

$$E_{PR} = \frac{I_A}{I_A + I_D} \qquad \textbf{(8)}$$

The relative proximity ratio is consistent only when the photophysical and instrumental conditions are also consistent or when they are not required, e.g. when the purpose of given experiment is to extract transition or dwell times. In our experiments, we used different corrections to obtain the real energy transfer efficiency. **Eq. 9** describes the $E_{\mathrm{FRET}}$ corrected for factors $\beta$ and $\gamma$ [49].

$$E_{FRET} = \frac{I_A - \beta I_D}{((I_A - \beta I_D) + \gamma I_D)} \qquad \textbf{(9)}$$

The factor $\gamma$ combines the differences in the donor and acceptor quantum yields $\varphi_D, \varphi_A$ and in their detection efficiency $\eta_D, \eta_A$ (**Eq. 10**) [49].

$$\gamma = (\frac{\eta_A}{\eta_D}) \times (\frac{\varphi_A}{\varphi_D}) \qquad \textbf{(10)}$$

In FRET techniques on diffusing molecules (including our investigations), the parameters in **Eq. 10** are obtained empirically [48]. The detection efficiency of our experiments was measured on two control samples with known E$_{\mathrm{FRET}}$. We estimated $\gamma$ as 1.1.

We also considered two cross-talk factors. The first one, the $\beta$ factor (**Eq. 9**), represents the leakage of a donor emission into an acceptor detection channel. The second cross-talk factor, which is often not considered, corrects the acceptor emission caused by its direct excitation which appears to be due to donor and acceptor spectral overlap [50]. Both factors were experimentally measured.

A simplified scheme of structural information accessible from FRET measurements is shown in **Fig. 6**.

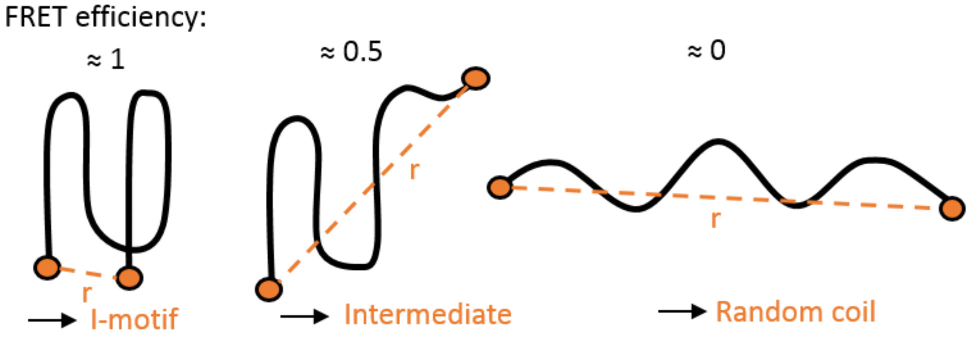


**Figure 6**. Scheme of a structural information accessible from FRET measurements; r is the obtained distance between two labeled sites on the molecule, according to which we were able to distinguish different structural conformers as for example an i-motif, a random coil or their intermediate.

### 3.5.2 Practical instrumentation

In our FRET measurements, labeled DNA molecules diffused in solution and optionally entered a confocal volume. It usually ranges in the scale of femtoliters and is designed for entrance of only single molecules [51].

In this volume, the donor molecule was excited with a green laser (532 nm) and its energy was to certain extent absorbed by the acceptor via non-radiative energy transfer. Then the fluorescence of both the donor and acceptor was detected in the form of photon bursts separated into two channels. The $E_{FRET}$ was obtained via comparison of the donor and acceptor emission intensities (**Eq. 9**). This was used to compute the distance between the fluorophores (**Eq. 6**) [52].

We built our own instrument for this method, partially inspired by similar investigation [52] and theoretical publications [48]. Its simplified scheme is described in **Supplementary mat. V**.

With this method, we measured all sample modifications only at the physiological pH 7.4. The effect of molecular crowding was omitted due to the properties of the solutions with 50 wt. % PEG, which are inappropriate for these types of measurements. The most suitable sample concentration for the single molecule investigation was around 0.5 nM.

### 3.5.3 Data analysis

Data obtained from the FRET measurements were interpreted in two ways. The first, further referred to as steady-state FRET, does not consider a molecule returning back into the confocal volume. Each detected photon burst is assigned to a different molecule. The values of $E_{FRET}$ are calculated individually for all detected photon bursts through the entire measurement time. Afterwards, they are put into a FRET efficiency histogram. This steady-state histogram can be further fitted with a probability distribution function (PDF, Maximum Likelihood Estimation). The recognized Gaussian peaks determine the $E_{FRET}$ and the proportions of the present components (structural conformers). However, this interpretation is useful only for studying the overall conformational distributions and cannot determine their time dependences or identify components with very short life times (short-living states) [52].

The second interpretation is based on the probability of one molecule diffusing more than once through the confocal volume. This method is called a recurrence analysis of single particles (RASP). In comparison with the steady-state FRET, more information is accessible using the RASP interpretation. For example more conformational subpopulations can be distinguished in the FRET efficiency histograms (the short-living states) as well as the dynamics of their distribution [52].

The reason why one molecule may return to the observed volume before a new molecule arrives is the low concentration (about 0.5 nM). We calculated the probability of two photon bursts being emitted by the same molecule $p_{same}(\tau)$ via a burst time correlation analysis (**Eq. 11**) [52].

$$p_{same}(\tau) = 1 - \frac{1}{g(\tau)} \qquad \textbf{(11)}$$

$\tau$ is the difference between the second and first burst times and $g(\tau)$ is an all bursts autocorrelation function.

An example of the autocorrelation function for just two sets of photon bursts, $B_1$ and $B_2$, each with a specific $E_{FRET}$ range, is shown in **Eq. 12** [52].

$$g_{B_1B_2}(\tau) = \frac{p(\{b_1, t\}, \{b_2, t + \tau\})}{p(\{b_1, t\}), p(\{b_2, t\})} \qquad \textbf{(12)}$$

$p(\{b_1, t\}, \{b_2, t + \tau\})$ is the probability that the bursts $b_1$ and $b_2$ are detected at both times $t$ and $t + \tau$ and $p(\{b_1, t\})$ and $p(\{b_2, t\})$ are the probabilities of their detection at time $t$. We assume that $b_1 \in B_1$ and $b_2 \in B_2$ and that $t$ is in the middle of the burst.

In the RASP FRET we constructed recurrence transfer efficiency histograms and fitted them with mixed Gaussian distributions. The recurrence histograms were made by investigation of a certain $E_{FRET}$ interval chosen from a steady-state FRET histogram and its redistribution in a short time interval. This recurrence interval had a known recurrence probability. Thus, we reliably observed only a single molecule in an appropriate recurrence time interval [52].

Further, we were able to derive kinetic information from the RASP interpretation. By putting histograms of very short recurrence time intervals (1-10 ms) that are continuously covering a longer time scale (10- 100 ms) into one graph, we could observe the time development (redistribution) of the distinguished components. Rate constants for each of the recognized Gaussian peaks were then obtained with an appropriate exponential fitting [52].

All data analyzes were done either in home-built Python software [44] or in Origin [45].

# 4 RESULTS

## 4.1 CD

We considered only two possible structural conformers in our first hypothesis, a random coil and an i-motif. The occurrence of intermediates was taken as unnoticeable at best. Thus, the baseline-corrected CD data (**Fig. 7a**) were first reconstructed via a two-state model. This model takes a reaction path as a linear combination of the starting (a random-coil) and resulting (an i-motif) states.

Then, $pH_{50}$ values were obtained from a dose-response function fitted to the calculated relative concentrations (**Fig. 7b**). However, the comparison of the reconstructed and raw CD data revealed positive residuals in a peak around 280 nm and negative residuals around 250 nm (**Fig. 7d**). These trends implied the presence of intermediates.

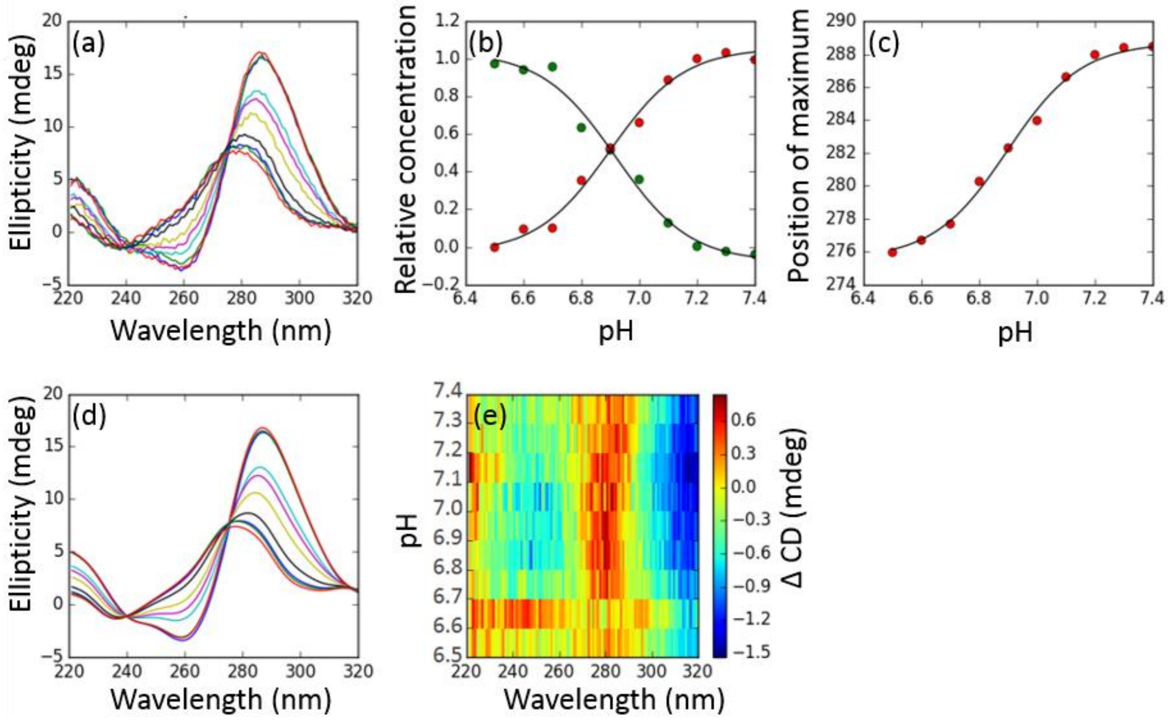


**Figure 7**. (a) Baseline-corrected CD data of the **$C_{1-4}$** sample measured at pHs varying from 6.5 to 7.4 at 0.1 step intervals; amplitude of the positive peak increases and moves to a longer wavelength with decreasing pH; (b) relative concentrations of initial and final states (random coil and i-motif) as a function of pH (black lines show sigmoid fit); (c) wavelengths of maxima as a function of pH (black line shows sigmoid fit); (d) reconstructed CD data of the same sample derived from a two-state model; (e) heat map of residuals between the two-state reconstruction and experimental data showing positive peak around 280 nm and negative around 250 nm.

Due to these results, we rejected the two-state model and performed further analyses with $pH_{50}$ values derived from the position of maxima. This method is more robust since the intermediate structures can affect the shape of the CD curves, but they cannot influence the positions of maxima. An example of the position of a maximum graph with a fitted sigmoid curve is shown in **Fig. 7c**.

We obtained two types of $pH_{50}$ values for each sample modification: one from the measurements in buffer without any PEG, further referred to as $pH_{50(noPEG)}$, and another in 50 wt. % PEG ($pH_{50(PEG)}$). All of the values with their standard deviations (SD) are shown in **Tab. 4** and **Fig. 8**.

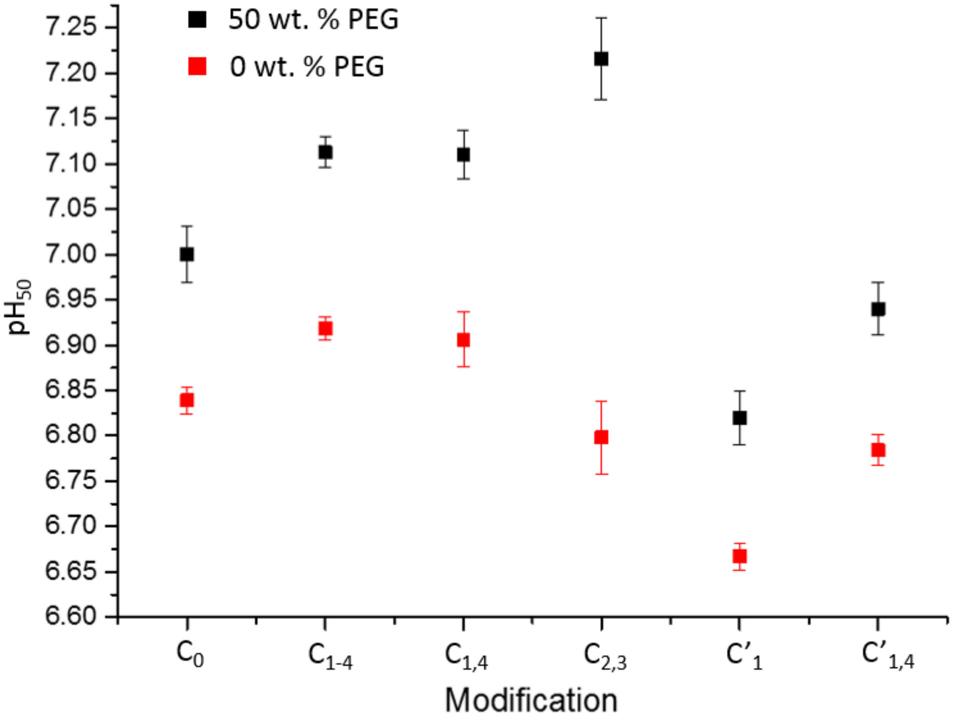


**Figure 8**. $pH_{50}$ values of each modified and control samples derived from the position of maxima; red color visualizing values of samples in 0 wt. % PEG and black color of samples in 50 wt. % PEG; error bars show standard deviations.

**Table 4**. $\mathrm{pH}_{50}$ values of each modified and control samples derived from the position of maxima; the difference in the $\mathrm{pH}_{50}$ values of samples in 0 wt. % and in 50 wt. % PEG and the difference in the $\mathrm{pH}_{50}$ values between modified samples and a control $\mathbf{C_0}$ sample in 0 wt. % and in 50 wt. % PEG are calculated.

| **Sample modification** | **$\mathrm{pH}_{50(\mathrm{noPEG})}$** | **Error (SD)** | **$\mathrm{pH}_{50(\mathrm{PEG})}$** | **Error (SD)** | **$\mathrm{pH}_{50(\mathrm{PEG})}$ - $\mathrm{pH}_{50(\mathrm{noPEG})}$** |
|---|---|---|---|---|---|
| | **0 wt. % PEG** | | **50 wt. % PEG** | | |
| $\mathbf{C_0}$ | 6.84 | 0.02 | 7.00 | 0.03 | 0.16 |
| $\mathbf{C_{1-4}}$ | 6.92 | 0.01 | 7.11 | 0.02 | 0.19 |
| $\mathbf{C_{1,4}}$ | 6.91 | 0.03 | 7.11 | 0.03 | 0.20 |
| $\mathbf{C_{2,3}}$ | 6.80 | 0.04 | 7.22 | 0.05 | 0.42 |
| $\mathbf{C'_1}$ | 6.67 | 0.02 | 6.82 | 0.03 | 0.15 |
| $\mathbf{C'_{1,4}}$ | 6.78 | 0.02 | 6.94 | 0.03 | 0.16 |

Generally, the $\mathrm{pH}_{50}$ values varied through all modifications. This fact reveals that the different methylation sites play an evident role in the i-motif pH stability. Focusing on the measurements without PEG, only two values were higher than the $\mathrm{pH}_{50(\mathrm{noPEG})}$ of the non-modified $\mathbf{C_0}$ sample (6.84), namely the $\mathbf{C_{1-4}}$ (6.92) and $\mathbf{C_{1,4}}$ (6.91) samples, meaning that these modifications showed a stabilization of the i-motif structure. On the other hand, the three remaining samples, $\mathbf{C_{2,3}}$ ($\mathrm{pH}_{50(\mathrm{noPEG})}$ = 6.80), $\mathbf{C'_1}$ (6.67) and $\mathbf{C'_{1,4}}$ (6.78), caused destabilization to various extents. A similar trend was observed in the molecular crowding measurements with the exception of the now stabilizing $\mathbf{C_{2,3}}$ modification. Its $\mathrm{pH}_{50(\mathrm{PEG})}$ was higher than the $\mathrm{pH}_{50(\mathrm{noPEG})}$ by 0.42. The molecular crowding effect increased the $\mathrm{pH}_{50}$ values in each sample. The two values ($\mathrm{pH}_{50(\mathrm{PEG})}$ and $\mathrm{pH}_{50(\mathrm{noPEG})}$) in the samples (other than $\mathbf{C_{2,3}}$) differed on average by 0.17 (SD 0.02). The significant gap between the PEG and no-PEG $\mathrm{pH}_{50}$ in the $\mathbf{C_{2,3}}$ sample implies a higher amount of unspecified intermediates present at 0 wt. % PEG, indicating that their folding into the i-motif probably proceeded only under molecular crowding conditions.

## 4.2 FRET

The CD measurements indicated the presence of intermediates. However, they provided no other information about them. To investigate these structures more thoroughly we used steady-state FRET and RASP FRET. The experiments were performed at pH 7.4 without the effect of molecular crowding. PEG induced a change in the diffusion coefficient and thus extended the time that the fluorophores spent in the confocal volume. This caused an early photo bleaching of the Cy5 molecules. For this reason, we measured all single molecule experiments only in the buffer without PEG.

### 4.2.1 Steady-state FRET

The collected data from the FRET measurements were analyzed first by the steady-state method. All detected photon bursts were put into FRET efficiency histograms, one for each sample modification. These histograms described the distribution of subpopulations with different $E_{FRET}$ (different structural conformers). Maximum Likelihood Estimation was used to divide each dataset into its individual components. The histograms for each sample modification with their PDF and the derived values of means and proportions are shown in **Fig. 9**.

At least three different conformers can be seen in each sample modification: one with a peak undoubtedly around $E_{FRET} = 0$ (a mixture of unfolded structures and bleached Cy5 molecules) and a second around $E_{FRET} = 1$ (fully folded i-motives). $E_{FRET}$ values under 0 and above 1 occur due to the symmetry of the PDF. A third peak can be seen around $E_{FRET} = 0{,}5$. This middle peak probably contains more than one subpopulation and represents folding intermediates. However, an effect of incomplete FRET pairs (caused by photo bleaching and blinking of one of the dye molecules) biases the interpretation of low FRET values. Also detailed interpretations are not possible because of an incapability to observe short-time distributions and thus to reliably distinguish all intermediate components.

Even though the steady-state FRET method cannot provide any information about the short-living states or the subpopulation dynamics, it is a suitable tool for observing the overall distributions of the structural subpopulations. For other purposes, the RASP FRET method was used.

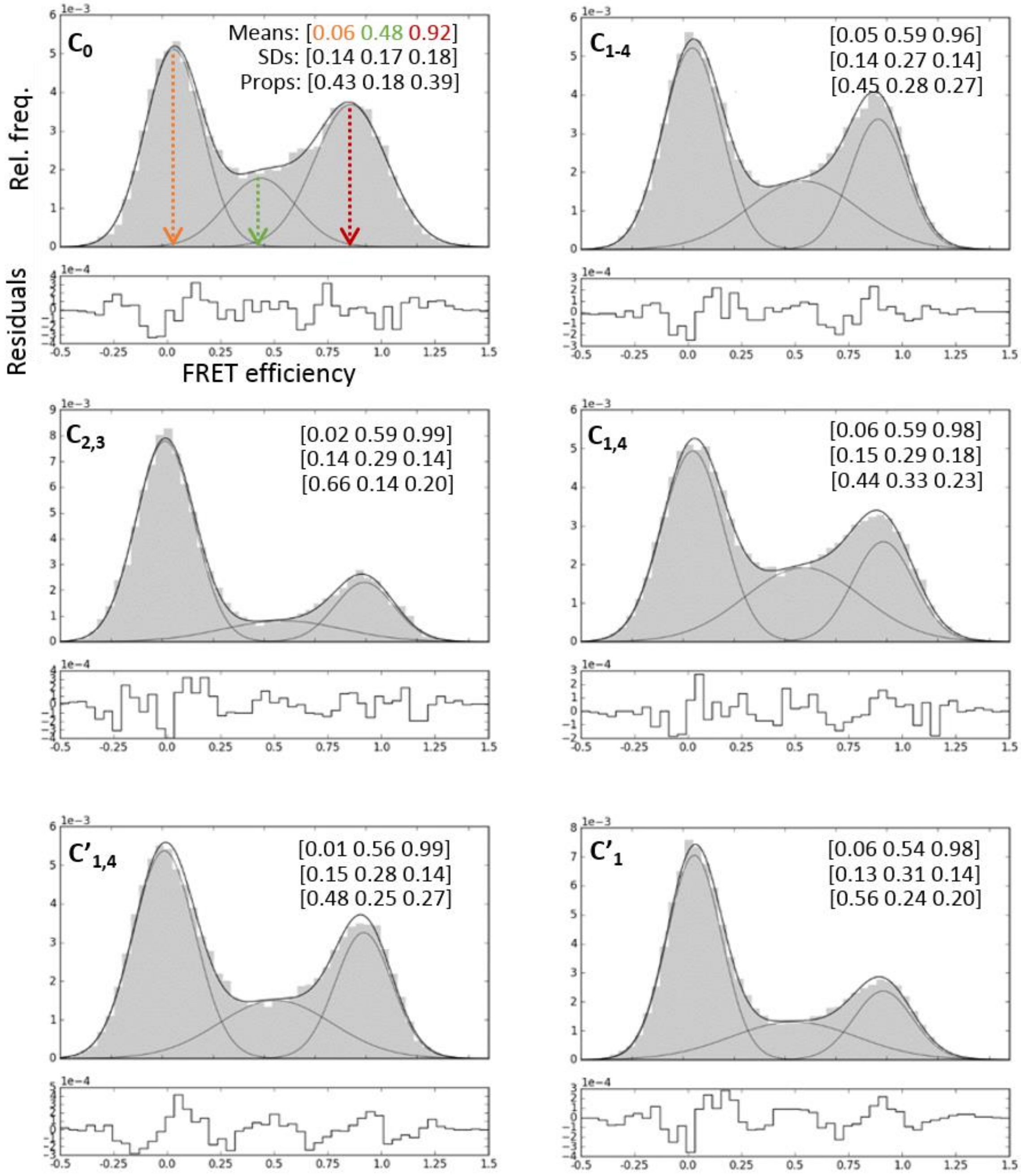


**Figure 9**. Steady-state FRET histograms with PDFs for all sample modifications showing relative frequency according to the $E_{FRET}$ of three different structural types: low peak for unfolded structures and bleached acceptor molecules, middle peak for unspecified intermediates and high peak for fully folded i-motives; values of peak midpoints (Means), their standard deviations (SDs) and peak proportions (Props.) are shown for each component (in order of the higher FRET eff.)

### 4.2.2 RASP FRET

To get an initial estimate of the positions of short-living subpopulations, we screened the different initial FRET efficiency ($E^0_{FRET}$) intervals in 0.1 steps, covering the range from -0.1 to 1.1. The resulting histograms (with the recurrence intervals 0 - 2 ms and 500 – 1500 ms) of all modifications are shown in **Supplementary mat. VI**. The short-time recurrence interval was chosen to detect potential short-living subpopulations. The long-time interval was chosen to make an almost steady-state histogram for comparison. Based on this screening we were able to distinguish five different conformational subpopulations (**Fig. 10)**. In addition to the two already observed steady peaks with means at $E_{FRET} = 0.00$ and $E_{FRET} = 1.00$ (**Fig. 9**), three new subpopulations appeared: their $E_{FRET}$ midpoints were at 0.23, 0.49 and 0.74. Because we were mostly interested in the time development of these subpopulations, we selected $E^0_{FRET}$ intervals for further RASP FRET analysis by iterative optimizing these five peaks widths. The first optimization step was to adjust each interval to cover only one subpopulation and thus not necessarily to correspond with the actual subpopulation peak width. In the second step, some of the intervals had to be extended due to the low amount of included photon bursts. The optimized values are shown in **Tab. 5**.

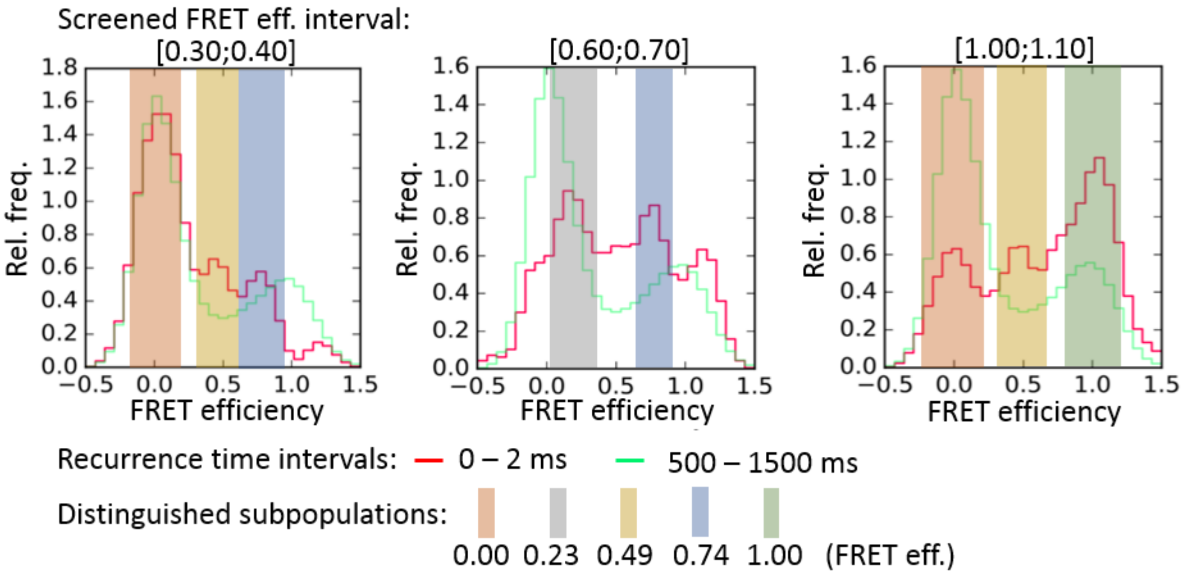


**Figure 10**. Example of three screening recurrence FRET efficiency histograms taken from the **$C'_1$** modification screening data set; all of the five distinguished subpopulations represented by separate peaks are clearly visible.

**Table 5**. Optimized initial FRET efficiency intervals used in RASP analysis.

| | $E^0_{FRET}$ intervals | | | | |
|---|---|---|---|---|---|
| $E^0_{FRET}$ (1) | -0.50 | 0.05 | 0.53 | 0.62 | 1.00 |
| $E^0_{FRET}$ (2) | 0.00 | 0.25 | 0.73 | 0.82 | 1.50 |

Further, we calculated the recurrence probability in time (see **Supplementary mat. VII**). The 50 % probability of two photon bursts being emitted by one molecule was observed at around 45 ms.

The recurrence transfer efficiency histograms were then plotted for each modification using the five optimized $E^0_{FRET}$ intervals. A noticeable time development was observed through the first 27 ms (see the histograms in **Supplementary mat. VIII**). We used the recurrence time intervals of 2 ms (to include a sufficient amount of photon bursts in each histogram). However, the intervals were overlaping (0 – 2, 1 – 3, …) to observe fluent time development. See **Fig. 11** for an example of the time development of recurrence histograms for $\mathbf{C_{2,3}}$ modification.

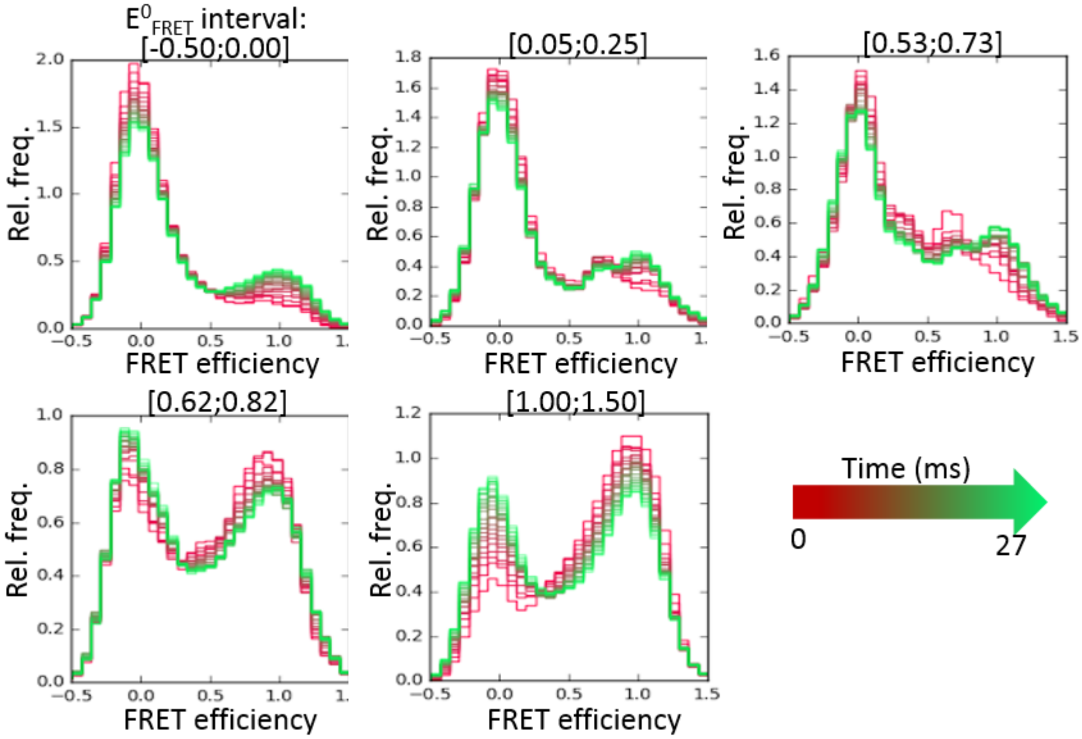


**Figure 11**. Example of recurrence FRET efficiency histograms plotted through the first 27 ms show the time development of five previously selected initial FRET efficiency intervals corresponding with recognized subpopulations, measured on $\mathbf{C_{2,3}}$ modification.

### 4.2.3 Kinetics

In the next step, we fitted the obtained recurrence transfer efficiency histograms to the PDFs. Five components describing the previously determined subpopulations were considered. For a better performance of the optimization process, we included data up to 50 ms, even though no significant changes occured after 27 ms.

Then we derived kinetic information from the time development of each of the five Gaussian components. These kinetics were exponentially fitted and for each function a rate constant was calculated. See **Supplementary mat. IX** for detailed information about the exponential fitting conditions and **Supplementary mat. X** for the kinetics of all modifications.

The RASP derived kinetics of each initial FRET efficiency interval through all sample modifications were evaluated together with the obtained rate constants and their standard errors (SE). We were able to determine the rate constants (with SE ≤ 10 % of a given value) for the overall i-motif folding path in each sample modification (**Fig. 12**). The folding constant was taken as the rate constant of an i-motif's exponential increase from the kinetics of the $E^0_{FRET}$ interval: (-0.50,0.00). Here, the random-coil population (around $E_{FRET} = 0$) changed to an i-motif population (around $E_{FRET} = 1$). The unfolding constant was taken as the rate constant of the random-coil's exponential increase from the kinetics of the $E^0_{FRET}$ interval: (1.00,1.50). Here, the populations' flow was reversible (from i-motif to random-coil).

Photo bleaching and blinking of dyes can induce errors into the $E^0_{FRET}$ interval (-0.50,0.00). However, the time scale of bleaching is in seconds and blinking in microseconds while our detection time range was from 2 to 50 ms. Thus the induced errors should not significantly influence the derived kinetics.

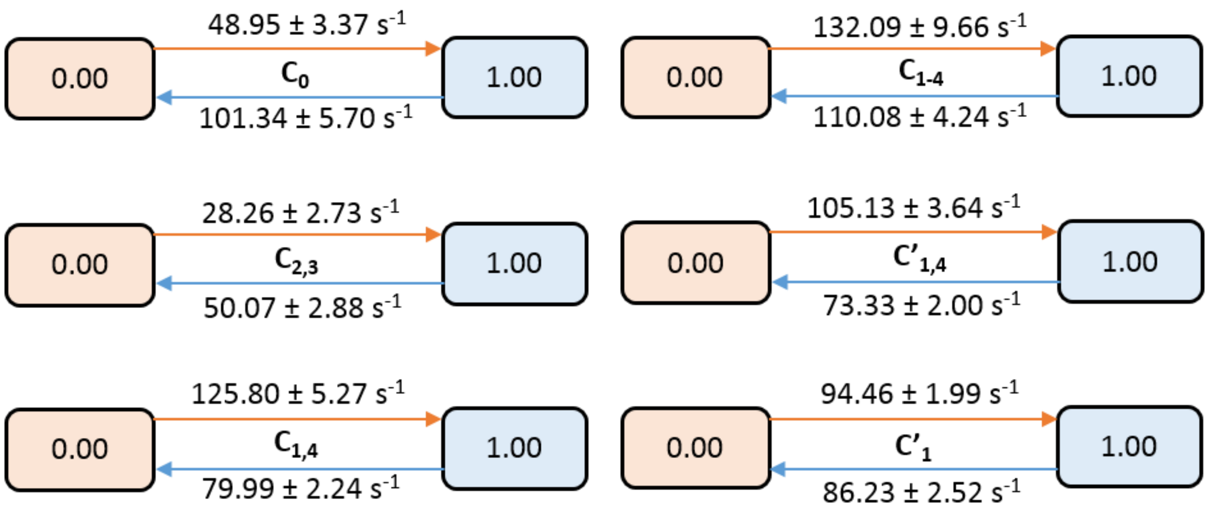


**Figure 12**. Scheme of overall i-motif folding paths for all samples with rate constants and their standard errors; the initial (a random coil) and the final (an i-motif) states are shown as their FRET efficiency means (0.00 for the random coil, 1.00 for the i-motif).

Again, as in the CD experiments, the observed values varied through all modifications and thus confirmed the methylation influence on i-motif stability. The majority of the observed kinetics provided constants in the range between 80 and 120 $s^{-1}$.

The values of the $\mathbf{C_0}$ sample were calculated as 48.95 ± 3.37 $s^{-1}$ for the folding and 101.34 ± 5.70 $s^{-1}$ for the unfolding rate constants. The $\mathbf{C_{2,3}}$ modification showed significantly lower values (28.26 ± 2.73 $s^{-1}$ for folding and 50.07 ± 2.88 $s^{-1}$ for unfolding path). According to the exponential fitting formula (**Supplementary mat. IX**), a lower rate constant means a higher reaction half time and thus a slower reaction. The low rate constants of the $\mathbf{C_{2,3}}$ modification indicate slower i-motif folding and unfolding compared to the other samples. The observed rate constants indicate that the unfolded molecule is more stable than the i-motif because the unfolding path is faster than the folding path in this modification. In addition to the anomalous behavior of this sample in the CD experiments, the $\mathbf{C_{2,3}}$ methylation is extremely interesting.

On the contrary, the $\mathbf{C_{1\text{-}4}}$ modification showed the highest rate constants for both the folding and unfolding paths (132.09 ± 9.66 $s^{-1}$ and 110.08 ± 4.24 $s^{-1}$). In this sample, the methylations enhanced the speed of the i-motif folding compared to the $\mathbf{C_0}$ sample.

In four cases (the **$C_{1,4}$**, **$C'_1$**, **$C'_{1,4}$** and **$C_{1-4}$** modifications), the folding constants were higher than the unfolding constants implying a higher stability of the i-motives compared to the random-coils. Only two samples showed the opposite behavior, the already mentioned **$C_{2,3}$** modification and the **$C_0$** sample. Here, the difference between the folding and unfolding constants (48.95 ± 3.37 $s^{-1}$ and 101.34 ± 5.70 $s^{-1}$) was the highest meaning that the i-motif folded on the non-modified sequence is rather unstable.

Further, we created detailed reaction path schemes by analyzing the kinetics of the three intermediate components (**Fig. 13**). We defined the maximum acceptable error (SE) as 20 % of the given rate constant and then chose such values that clearly corresponded with one sub-reaction. However, we also distinguished many paths with indeterminable or unreliable rate constants.

The three intermediates were observed in all sample modifications, but with different proportions, kinetics and detected reaction paths. In general, the "0.49" component was the most abundant of the intermediates in all investigated samples. Only in the **$C'_1$** modification we did not detect this component as an intermediate in the random-coil to i-motif folding.

The "0.23" component was most visible in the non-modified sample, but it was distinguished in smaller proportions in the other samples. The reversible path from a random-coil to this intermediate was observed in all modifications. However, the **$C_{2,3}$**, **$C'_{1,4}$** and **$C'_1$** samples showed no evidence of this intermediate folding into the i-motif structure.

The third "0.74" intermediate was registered in the smallest proportions, but it clearly appears in the reaction paths. Its transfer into all other components is clearly visible in all modifications from the (0.62,0.82) $E^0_{FRET}$ interval investigations. These paths are seen mainly in the shortest recurrence transfer efficiency histograms (0 - 2 ms). Thus, the "0.74" intermediate is probably a very short-living state with kinetics faster than 2 ms.

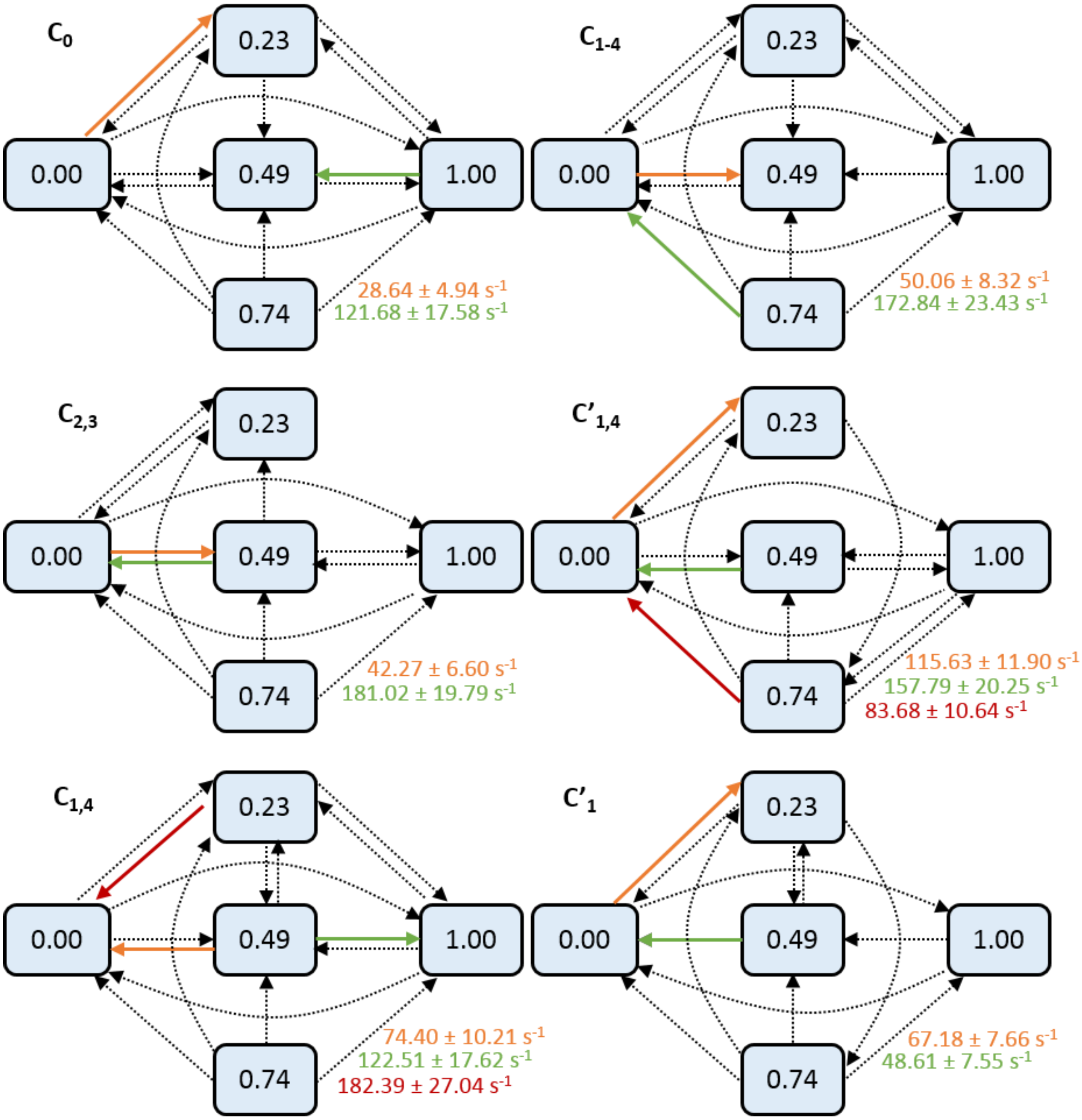


**Figure 13**. Schemes of i-motif folding paths with three intermediates through all sample modifications; design based on our single-molecule experiments with rate constants where the assignment was clear and the error was acceptable (SE ≤ 20 %); the components are described by their FRET efficiency means (0.00 for a random-coil, 1.00 for an i-motif, 0.23, 0.49 and 0.74 for the recognized intermediates).

# 5 DISCUSSION

In the CD experiments, we investigated five differently methylated ILPR sequences by analyzing the amount of folded i-motives as a function of pH and molecular crowding. The aims were to determine the potential i-motif occurrence under physiological conditions (pH 7.4, 140 mM $K^+$, 12 mM $Na^+$) and to evaluate the influence of the position of methylated cytosine on the i-motif stability.

I-motif folding, and therefore its potential regulating functions in living organisms, have long been uncertain due to its known stabilization in acidic conditions. For example, a study of double-stranded ILPR sequence found that, in similar ionic conditions (100 mmol $K^+$) at pH 7.4, only a G-quadruplex could fold while at pH 5.5 both structures, the G-quadruplex and i-motif, could occur even though they are mutually exclusive [53]. In a different study, an i-motif formation on the c-MYC gene was observed under physiological ionic conditions with molecular crowding at pH 6.7 [34].

We showed that, under a molecular crowding effect simulated by 50 wt. % PEG and with 5-methylated cytosines at specific positions (**$C_{2,3}$** modification), an i-motif structure in the ILPR can have a $pH_{50}$ of 7.22. Because 50 % of the investigated molecules were folded into an i-motif at the given $pH_{50}$, we assumed that there are fully folded i-motives present also at the physiological pH of 7.40. Thus, it is possible that i-motives can naturally occur in the ILPR.

The fact that the $pH_{50}$ values varied through all of the differently methylated samples indicates a strong influence of this epigenetic modification on i-motif stability. It was shown on the c-MYC gene that the presence of a single 5'mC in a C-rich sequence of its promotor stabilizes the i-motif structure by increasing its $pH_{50}$ by 0.2 (from 6.1 to 6.3) [19]. In our case, with four 5'mCs in each sequence and under the same 0 wt. % PEG conditions, there was a lower range of stabilization, with the shift in the i-motif $pH_{50}$ being only + 0.08 (from 6.84 to 6.92) for the most stabilizing (**$C_{1\text{-}4}$**) type of methylation. Even though the sequence used in the c-MYC experiments differed from ours, e.g. in the number of repetitive cytosines or in the loop lengths, the i-motif stabilizing influence of cytosine methylation was showed in both.

Until recently, the effect of molecular crowding was frequently not considered in simulating physiological conditions during nucleic acids secondary structures experiments. We investigated the influence of PEG induced crowding on the ILPR i-motif.

We observed a significant shift in the i-motif $pH_{50}$ values obtained from the molecular crowding measurements in 50 wt. % PEG. They were on average 0.17 (SD 0.02) higher than the $pH_{50}$ values of 0 wt. % PEG samples with the exception of the $\mathbf{C_{2,3}}$ modification, in which the $pH_{50}$ increased more significantly (shift by 0.42 from 6.80 to 7.22). The $pH_{50}$ of the non-modified sample increased from 6.84 to 7.00, which is consistent with other published studies [19,21]. The unusual shift observed in the i-motif $pKa_{50\%}$ value of the $\mathbf{C_{2,3}}$ sample, which changed from a destabilizing to a stabilizing modification, indicated a high amount of intermediates present in the $\mathbf{C_{2,3}}$ 0 wt. % PEG sample. The existence of partially folded i-motif structures in the ILPR sequence was already shown in [17]. We therefore focused in detail on the intermediates in the single-molecule experiments.

In summation, the CD experiments determined the influence of the chosen cytosine methylations on i-motif pH stability. In the molecular crowding condition natural to the cell environment, three of the modified sequences, $\mathbf{C_{1\text{-}4}}$, $\mathbf{C_{1,4}}$ and $\mathbf{C_{2,3}}$, showed higher $pH_{50}$ than the non-modified sample. On the other hand, the i-motives in the $\mathbf{C'_{1}}$ and $\mathbf{C'_{1,4}}$ samples were destabilized. Using this information, we analyzed the exact 5'mCs locations in each sample and tried to explain the observed shifts in the i-motives' pH stability. We numbered the binding cytosenes in our DNA sequence: 5' Cy5 – TGT $C_1C_2C_3C_4$ ACA $C_5C_6C_7C_8$ TGT $C_9C_{10}C_{11}C_{12}$ ACA $C_{13}C_{14}C_{15}C_{16}$ TGT – Cy3 3'. See **Fig. 14** for a visualization of the numbered sequence in the i-motif structure.

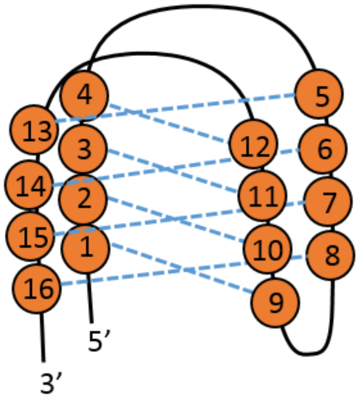


**Figure 14**. Schematic visualization of a numbered DNA sequence folded into an i-motif.

Probably the most important methylation enhancing i-motif folding was located on the thirteenth of the binding cytosines in our sequence ($C_{13}$). In a potential triplex intermediate, such a methylation could help to attach the last unfolded tetrad to form the i-motif (**Fig. 15**). We presume similar behavior in all potential symmetrically methylated sequences (for example with the methylation on $C_4$ instead of $C_{13}$).

The methylation of $C_{13}$ occurred in the most stabilizing $\mathbf{C_{1\text{-}4}}$ and $\mathbf{C_{1,4}}$ modifications. In the $\mathbf{C_{1,4}}$ sample, the bonding $C_5$ was also methylated although this double methylation was not necessarily required for i-motif stabilization, as was seen in the $\mathbf{C_{1\text{-}4}}$ sample. On the other hand, the presence of the methylated $C_{13}$ in combination with the methylated $C_{12}$ caused a destabilization in the $\mathbf{C'_1}$ and $\mathbf{C'_{1,4}}$ samples.

Based on these observations, the modification was generally taken as stabilizing when one or both of the bonded cytosines were methylated. Methylations located on the neighboring, but not interacting, cytosines led to destabilization (for example methylated $C_{13}$ and $C_{12}$). Here the methyl groups probably caused a sterical repulsion. The most destabilizing $\mathbf{C'_1}$ modification contained two such repulsing methylated pairs. Here, the repulsing pair of $C_9$ and $C_8$ could inhibit the folding of the potential triplex structure.

The $\mathbf{C_{2,3}}$ modification was specific because it had no methylated cytosines on the positions described above. Instead, the positions $C_{14}$ and $C_{15}$ were modified. This enhanced the triplex to i-motif folding only under molecular crowding conditions. The PEG molecules mechanically enforced the 5’mCs to interact.

The role of the methylation sites in i-motif folding ways was not analyzed other than in the triplex due to the lack of structural information about other potential intermediates.

A scheme of the suggested i-motif stabilizing and destabilizing sites of methylations can be seen in **Fig. 15**.

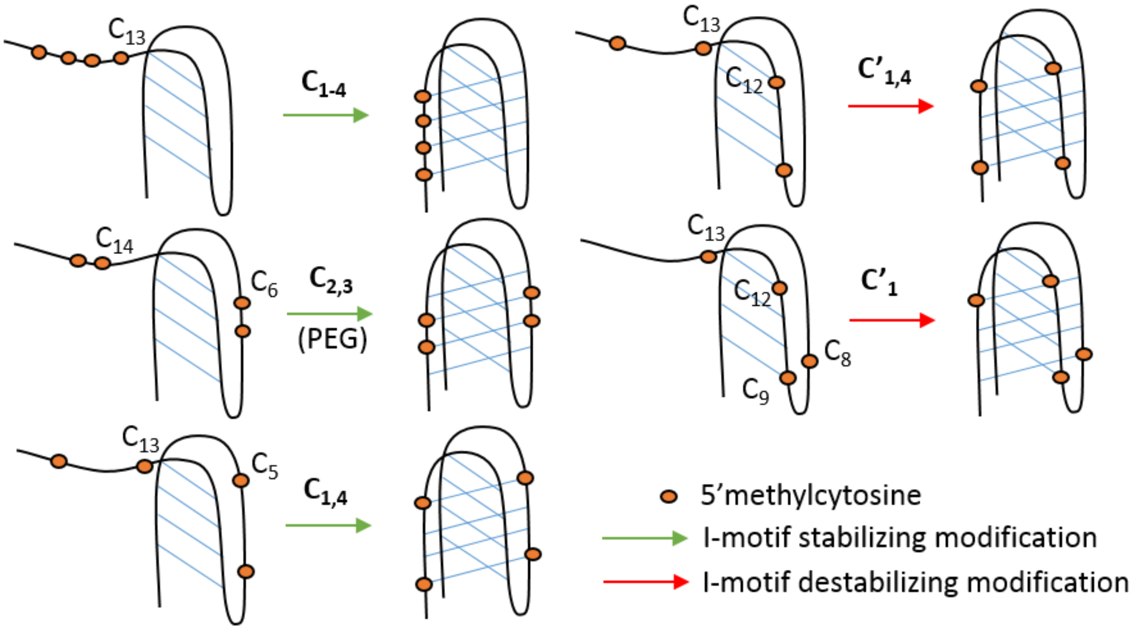


**Figure 15**. Simplified scheme of partially folded triplex structures folding into i-motives; methylation sites are indicated and the most stabilizing or destabilizing ones are numbered.

With the single-molecule FRET we investigated the folding path of the ILPR i-motif and how it differs through all of the sample modifications. With the RASP analysis we recognized three intermediate components and further examined their participation in i-motif folding.

I-motif intramolecular conformational changes were investigated in another single-molecule study [16]. FRET experiments confirmed the coexistence of partially folded structures with unfolded single strand coils. Moreover, three structures were observed at pH 5.8 with peaks at FRET efficiencies of 0.32, 0.59 and 0.86, while only two structures were recognized at pH 7.2 (peaks at 0.32 and 0.59). Considering the slight differences in DNA sequences and experimental conditions, the three peak values are comparable to our recognized intermediates (0.23, 0.49 and 0.74). However, there were some significant differences. In [16], the "0.86" peak was assigned to the i-motif structure. In our FRET experiments, an i-motif structure was clearly observed as a peak at $E_{FRET}$ 1.00. We also proved the existence of all three intermediates under physiological pH 7.4. In addition, we were able to incorporate the three components into the i-motif folding and unfolding path.

The "0.74" component was observed in all samples, even though with lower proportions than the other two intermediates. Its transition into all other components was seen in all modifications. Most times, we were not able to extract the transition rates, because of the low abundance of this state. In addition, the component observed in the $E^0_{FRET}$ interval (0.62,0.82) was often prominent only in the first recurrence interval of 0-2 ms. This indicates that the "0.74" intermediate undergoes conformational changes faster than we are able to detect by RASP.

Further, we suggest a triplex structure of this intermediate. This structure is known to exist on the C-rich strands [54]. The $E_{FRET}$ of 0.74 is closer to that of the i-motif (1.00) than to that of the random-coil (0.00) indicating only a small conformational difference from the fully folded structure. The kinetics show its ability to transform easily into the i-motif and random-coil alike. The suggested triplex structure with its transformations into the i-motif and random-coil are shown in **Fig. 16**.

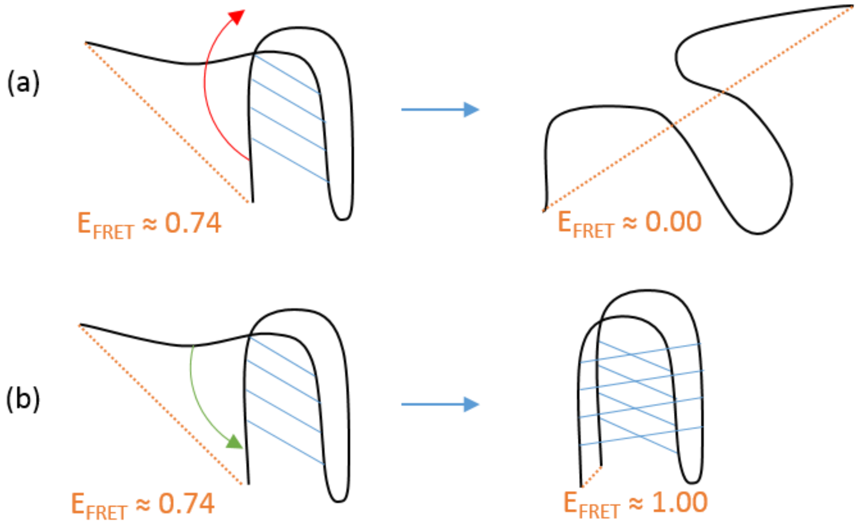


**Figure 16**. Simplified triplex structure: (a) unfolding into a random coil; (b) folding into an i-motif; suggested FRET efficiencies of each structure are shown.

We do not have enough information to assign specific structures to the other two intermediates with peaks at $E_{FRET}$ 0.23 and 0.49. However, we expect hairpin formation in the ILPR. This structure is known to occur with i-motif for example in human telomere [55] or in BCL2 promoter [56]. Therefore, the two subpopulations could be some of the hairpin conformers.

Besides distinguishing the intermediates, we derived kinetic information by observing the time developments of all the components. The random-coil to i-motif folding rate constants ranged from 28 $s^{-1}$ to 132 $s^{-1}$ and the unfolding rates from 50 $s^{-1}$ to 110 $s^{-1}$.

Rate constants of hairpin to i-motif folding and unfolding paths were measured under various $Cu^{2+}$ and EDTA concentrations [55]. The observed rate values ranged from 0.010 to 0.025 $s^{-1}$. Our results were up to four orders of magnitude higher. We interpret this disparity as the result of different experimental methods. The CD measurements used in [55] were incapable of detecting the initial short-time development with the dead time of 8 s. On the contrary, we observed the kinetics at 50 ms intervals in the single-molecule experiments. Thus, we were able to examine the detailed reaction mechanism and not only the overall i-motif stabilization in the given conditions. However, an intermediate existence was suggested in [55], which is in accordance with our results, although without any detailed specification.

In another kinetic study, the i-motif folding and unfolding rates were obtained from stopped-flow CD measurements [57]. The transformations were induced by pH changes. The observed i-motives folded and unfolded in the time scale of 100 ms. This value corresponds to the rate constant of 10 $s^{-1}$ (the kinetic exponential fitting was the same as in our study and can be seen in **Supplementary mat. IX**). Although this reaction rate is relatively close to some of our results (**$C_{2,3}$** folding rate of 28 $s^{-1}$), the majority of our rate constants were about ten times higher. Again, this discrepancy is most likely the result of different methods, even though the stopped-flow technique is more advantageous for this type of measurements than the classical CD.

We compared the obtained folding kinetics also to our CD experiments in 0 wt. % PEG. The highest i-motif $pH_{50}$ values were measured in the **$C_{1\text{-}4}$** (6.92) and **$C_{1,4}$** (6.91) modifications. Accordingly, these two samples showed the highest folding rate constants at pH 7.4 (**$C_{1\text{-}4}$** 132 $s^{-1}$, **$C_{1,4}$** 126 $s^{-1}$). Their unfolding rates (**$C_{1\text{-}4}$** 110 $s^{-1}$, **$C_{1,4}$** 80 $s^{-1}$) were lower than the folding and thus made the i-motives relatively stable. The difference in the folding and unfolding constants (+45 $s^{-1}$) in the **$C_{1,4}$** sample was the highest among those resulting in stability.

The destabilizing negative differences were observed only in the **$C_0$** (-52 $s^{-1}$) and **$C_{2,3}$** (-22 $s^{-1}$) samples. Also the folding rates of these two modifications were the lowest (**$C_0$** 49 $s^{-1}$, **$C_{2,3}$** 28 $s^{-1}$). In accordance with the no-PEG CD, the $pH_{50}$ of **$C_0$** (6.84) and **$C_{2,3}$** (6.80) were lower than that of the **$C_{1\text{-}4}$** and **$C_{1,4}$** samples.

In summation, the $\mathbf{C_0}$, $\mathbf{C_{1\text{-}4}}$, $\mathbf{C_{2,3}}$ and $\mathbf{C_{1,4}}$ modifications showed similar trends in both the no-PEG CD and FRET experiments. Based on this, we presume that the $\mathbf{C_{2,3}}$ sample would show a huge shift in the folding rates in 50 wt. % PEG, as was seen with the $pH_{50}$ in the PEG CD. However, the PEG FRET experiments could not be performed.

On the contrary, the $\mathbf{C'_1}$ and $\mathbf{C'_{1,4}}$ samples showed no reasonable similarities in the CD and FRET experiments. For example, the $\mathbf{C'_1}$ had the lowest $pH_{50}$ (6.67) in 0 wt. % PEG, but its i-motif folding rate was not the lowest (94 $s^{-1}$). Moreover, the folding rate was higher than the unfolding rate (86 $s^{-1}$) and showed a rather stabilizing effect. This contradiction may be caused by complications in the $\mathbf{C'_{1,4}}$ and $\mathbf{C'_1}$ folding paths brought about by the repulsing 5'mC interactions noted above (**Fig. 15**).

# 6 CONCLUSION

We investigated the human ILPR C-rich sequence using two complementary methods: CD and single-molecule FRET. With CD, we examined the stability of a secondary structure called an i-motif and how it is influenced by pH, cytosine methylations and molecular crowding. The effect of the methylations both stabilized and destabilized the i-motif structure at higher pHs. The molecular crowding induced by 50 wt. % PEG increased i-motif stability in all differently methylated samples. In one case, the $pH_{50}$ (value at which 50 % of molecules are folded into i-motives) reached 7.22 with a shift of 0.42 from a destabilizing to stabilizing modification. This observation suggested the possibility of the i-motif folding in living cells and indicated the presence of intermediates in the folding path.

Further, we focused on the participation of these intermediates in the random-coil to i-motif folding and unfolding paths and on the kinetics at the physiological pH 7.4 with 0 wt. % PEG using FRET with RASP analysis. We distinguished three different intermediate components in the RASP histograms with average FRET means at 0.23, 0.49 and 0.74. All of these structures were observed in the i-motif folding paths in all samples, albeit at different proportions. We proposed that the short-living “0.74” component has a triplex structure and is able to transform not only into the random-coil and i-motif, but also into the other two intermediates. The derived random-coil to i-motif folding rate constants ranged from 28 $s^{-1}$ to 132 $s^{-1}$, and the unfolding from 50 $s^{-1}$ to 110 $s^{-1}$.

The two highest i-motif folding rates were measured in the most stabilizing samples in accordance to our no-PEG CD experiments. Here, the folding rates were higher than the unfolding. On the other hand, the sample with the slowest kinetics showed a higher unfolding rate than the folding. This is in agreement with its destabilizing influence observed in the no-PEG CD experiments. However, the same sample showed the highest stabilization in the 50 wt. % PEG in CD measurements. Based on this, a big shift in the rate constants would also be expected in the 50 wt. % PEG FRET experiments. The molecular crowding FRET experiments had to be omitted due to extended diffusion time which increased photo bleaching.

The locations of the 5’mCs in each sample were analyzed. Methylation of the thirteenth binding cytosine in our sequence (first C in the fourth tetrad) was suggested as the most stabilizing site. The samples containing neighboring, but not interacting 5’mCs, each in a different C tetrad, were revealed to be destabilizing in the CD experiments, but provided relatively high i-motif folding rate constants in FRET. The methyl groups attached to the inner two cytosines of the second and fourth C tetrads stabilized the i-motif structure at the highest pH (7.22) but only under the molecular crowding effect.

In the future, FRET method capable of observing molecules under molecular crowding conditions would help to complete the investigation of i-motif occurrence in the natural cell environment. Also, FRET is recently often combined with the fluorescence correlation spectroscopy (FCS), another single-molecule method. FCS improves the time resolution meaning that i- motif folding could be studied in even shorter time scales. Moreover, our laboratory is currently developing a FRET *in cell* technique with high temporal resolution. With this, the three distinguished intermediates could be investigated in detail and the whole i-motif folding path could be clarified.

# 7 REFERENCES

1. Fu, Z., Gilbert, E. R. & Liu, D. Regulation of insulin synthesis and secretion and pancreatic Beta-cell dysfunction in diabetes. *Curr. Diabetes Rev.* **9,** 25–53 (2013).
2. Weiss, M., Steiner, D. F. & Philipson, L. H. Insulin Biosynthesis, Secretion, Structure, and Structure-Activity Relationships. (MDText.com, Inc., 2014).
3. Wilcox, G. Insulin and insulin resistance. *Clin. Biochem. Rev.* **26,** 19–39 (2005).
4. Owens, D. R., Grimley, J. & Kirkpatrick, P. Inhaled human insulin. *Nat. Rev. Drug Discov.* **5,** 371–2 (2006).
5. Bastaki, S. Diabetes mellitus and its treatment. *Int. J. Diabetes Metab.* **13,** 111–134 (2005).
6. Pruitt, K., Brown, G., Tatusova, T. & Maglott, D. The Reference Sequence (RefSeq) Database. (2012).
7. Catasti, P., Chen, X., Moyzis, R. K., Bradbury, E. M. & Gupta, G. Structure-function correlations of the insulin-linked polymorphic region. *J. Mol. Biol.* **264,** 534–45 (1996).
8. Schonhoft, J. D. *et al.* ILPR repeats adopt diverse G-quadruplex conformations that determine insulin binding. *Biopolymers* **93,** 21–31 (2010).
9. Bloomfield, V. A., Crothers, D. M. & Tinoco, I. *Nucleic Acids: Structures, Properties, and Functions*. (University Science Books, 2000).
10. Murat, P. & Balasubramanian, S. Existence and consequences of G-quadruplex structures in DNA. *Curr. Opin. Genet. Dev.* **25,** 22–29 (2014).
11. Gehring, K., Leroy, J.-L. & Guéron, M. A tetrameric DNA structure with protonated cytosine-cytosine base pairs. *Nature* **363,** 561–565 (1993).
12. Benabou, S., Aviñó, a., Eritja, R., González, C. & Gargallo, R. Fundamental aspects of the nucleic acid i-motif structures. *RSC Adv.* **4,** 26956 (2014).
13. Lieblein, A. L., Krämer, M., Dreuw, A., Fürtig, B. & Schwalbe, H. The nature of hydrogen bonds in cytidine···H+···cytidine DNA base pairs. *Angew. Chem. Int. Ed. Engl.* **51,** 4067-70 (2012).
14. Day, H. A., Pavlou, P. & Waller, Z. A. E. i-Motif DNA: structure, stability and targeting with ligands. *Bioorg. Med. Chem.* **22,** 4407–18 (2014).
15. Kanaori, K., Shibayama, N., Gohda, K., Tajima, K. & Makino, K. Multiple four-stranded conformations of human telomere sequence d(CCCTAA) in solution. *Nucleic Acids Res.* **29,** 831–40 (2001).

16. Choi, J., Kim, S., Tachikawa, T., Fujitsuka, M. & Majima, T. pH-induced intramolecular folding dynamics of i-motif DNA. *J. Am. Chem. Soc.* **133,** 16146–53 (2011).
17. Dhakal, S. *et al.* Coexistence of an ILPR i-Motif and a partially folded structure with comparable mechanical stability revealed at the single-molecule level. *J. Am. Chem. Soc.* **132,** 8991–8997 (2010).
18. Brooks, T. A., Kendrick, S. & Hurley, L. Making sense of G-quadruplex and i-motif functions in oncogene promoters. *FEBS J.* **277,** 3459–69 (2010).
19. Bhavsar-Jog, Y. P., Van Dornshuld, E., Brooks, T. A., Tschumper, G. S. & Wadkins, R. M. Epigenetic modification, dehydration, and molecular crowding effects on the thermodynamics of i-motif structure formation from C-rich DNA. *Biochemistry* **53,** 1586–94 (2014).
20. Day, H. A., Huguin, C. & Waller, Z. A. E. Silver cations fold i-motif at neutral pH. *Chem. Commun.* **49,** 7696 (2013).
21. Rajendran, A., Nakano, S. & Sugimoto, N. Molecular crowding of the cosolutes induces an intramolecular i-motif structure of triplet repeat DNA oligomers at neutral pH. *Chem. Commun. (Camb).* **46,** 1299–301 (2010).
22. Golbabapour, S., Abdulla, M. A. & Hajrezaei, M. A concise review on epigenetic regulation: insight into molecular mechanisms. *Int. J. Mol. Sci.* **12,** 8661–94 (2011).
23. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13,** 484–92 (2012).
24. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187,** 226–232 (1975).
25. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–22 (2011).
26. Nardo, L. *et al.* Effects of non-CpG site methylation on DNA thermal stability: a fluorescence study. *Nucleic Acids Res.* **43,** 10722–10733 (2015).
27. Woodcock, D. M., Crowther, P. J. & Diver, W. P. The majority of methylated deoxycytidines in human DNA are not in the CpG dinucleotide. *Biochem. Biophys. Res. Commun.* **145,** 888-894 (1987).
28. ACD/ChemSketch (Freeware), version 12.01, Advanced Chemistry Development, Inc., Toronto, On, Canada, www.acdlabs.com, 2016

29. Ellis, R. J. & Minton, P. A. Cell biology: join the crowd. *Nature* **425,** 27–28 (2003).

30. Miyoshi, D. & Sugimoto, N. Molecular crowding effects on structure and stability of DNA. *Biochimie* **90,** 1040–51 (2008).

31. Nakano, S., Karimata, H., Ohmichi, T., Kawakami, J. & Sugimoto, N. The effect of molecular crowding with nucleotide length and cosolute structure on DNA duplex stability. *J. Am. Chem. Soc.* **126,** 14330–1 (2004).

32. Goobes, R. & Minsky, A. Thermodynamic Aspects of Triplex DNA Formation in Crowded Environments. *J. Am. Chem. Soc.* **123,** 12692–12693 (2001).

33. Kan, Z.-Y. *et al.* Molecular crowding induces telomere G-quadruplex formation under salt-deficient conditions and enhances its competition with duplex formation. *Angew. Chem. Int. Ed. Engl.* **45,** 1629–32 (2006).

34. Cui, J., Waltman, P., Le, V. H. & Lewis, E. A. The effect of molecular crowding on the stability of human c-MYC promoter sequence I-motif at neutral pH. *Molecules* **18,** 12751–67 (2013).

35. Buscaglia, R. *et al.* Polyethylene glycol binding alters human telomere G-quadruplex structure by conformational selection. *Nucleic Acids Res.* **41,** 7934–7946 (2013).

36. Hohng, S., Joo, C. & Ha, T. Single-Molecule Three-Color FRET. *Biophys. J.* **87,** 1328–1337 (2004).

37. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5,** 507–516 (2008).

38. Rodger, A. & Nordén, B. *Circular Dichroism and Linear Dichroism*. (Oxford University Press, 1997).

39. Berova, N. & Nakanishi, K. *Circular Dichroism: Principles and Applications*. (John Wiley & Sons, 2000).

40. Sponer, J., Gabb, H. a, Leszczynski, J. & Hobza, P. Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study. *Biophys. J.* **73,** 76–87 (1997).

41. Tuma, R., Tsuruta, H., French, K. H. & Prevelige, P. E. Detection of Intermediates and Kinetic Control during Assembly of Bacteriophage P22 Procapsid (Supplementary material). *J. Mol. Biol.* **381,** 1395–1406 (2008),

42. Golub, G. H. & Loan, C. F. Van. *Matrix Computations*. (JHU Press, 1996).

43. Lannes, L., Halder, S., Krishnan, Y. & Schwalbe, H. Tuning the pH Response of i-Motif DNA Oligonucleotides. *ChemBioChem* **16,** 1647–1656 (2015).

44. Python (Python Software Foundation, at www.python.org) using Anaconda analytics platform (Anaconda Software Distribution, Continuum Analytics, version 2-2.4.0, at https://continuum.io)

45. Origin, version 8.5.1, OriginLab, Northampton, MA, at originlab.com, 2016

46. Förster, T. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.* **437,** 55–75 (1948).

47. Gell, C., Brockwell, D. & Smith, A. *Handbook of Single Molecule Fluorescence Spectroscopy*. (OUP Oxford, 2006).

48. Lakowicz, J. R. *Principles of Fluorescence Spectroscopy*. (Springer Science & Business Media, 2007).

49. McCann, J. J., Choi, U. B., Zheng, L., Weninger, K. & Bowen, M. E. Optimizing methods to recover absolute FRET efficiency from immobilized single molecules. *Biophys. J.* **99,** 961–970 (2010).

50. Lee, N. K. *et al.* Accurate FRET measurements within single diffusing biomolecules using alternating-laser excitation. *Biophys. J.* **88,** 2939–2953 (2005).

51. Rüttinger, S. *et al.* Comparison and accuracy of methods to determine the confocal volume for quantitative fluorescence correlation spectroscopy. *J. Microsc.* **232,** 343–352 (2008).

52. Hoffmann, A. *et al.* Quantifying heterogeneity and conformational dynamics from single molecule FRET of diffusing molecules: recurrence analysis of single particles (RASP). *Phys. Chem. Chem. Phys.* **13,** 1857–71 (2011).

53. Dhakal, S. *et al.* G-quadruplex and i-motif are mutually exclusive in ILPR double-stranded DNA. *Biophys. J.* **102,** 2575–84 (2012).

54. Lacroix, L. & Mergny, J. L. Chemical modification of pyrimidine TFOs: effect on i-motif and triple helix formation. *Arch. Biochem. Biophys.* **381,** 153–63 (2000).

55. Day, H. A., Wright, E. P., MacDonald, C. J., Gates, A. J. & Waller, Z. A. E. Reversible DNA i-motif to hairpin switching induced by copper(II) cations. *Chem. Commun.* **51,** 14099-14102 (2015).

56. Kendrick, S. *et al.* The dynamic character of the BCL2 promoter i-motif provides a mechanism for modulation of gene expression by compounds that bind selectively to the alternative DNA hairpin structure. *J. Am. Chem. Soc.* **136,** 4161–71 (2014).

57. Chen, C. *et al.* Study of pH-induced folding and unfolding kinetics of the DNA i-motif by stopped-flow circular dichroism. *Langmuir* **28,** 17743–8 (2012).

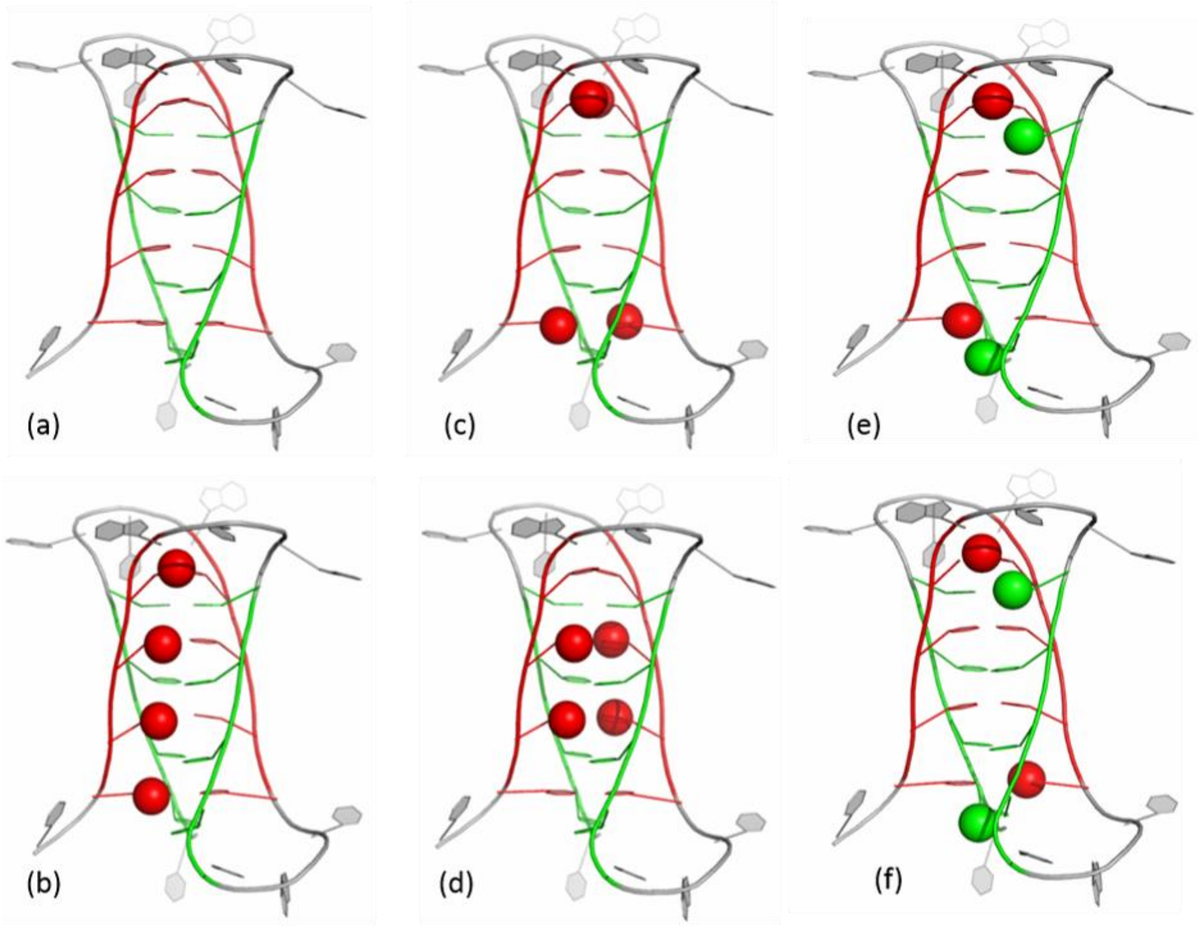
58. Flores, S. C., Sherman, M. A., Bruns, C. M., Eastman, P. & Altman, R. B. Fast flexible modeling of RNA structure using internal coordinates. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8,** 1247–57

59. The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC, at www.pymol.org, 2016

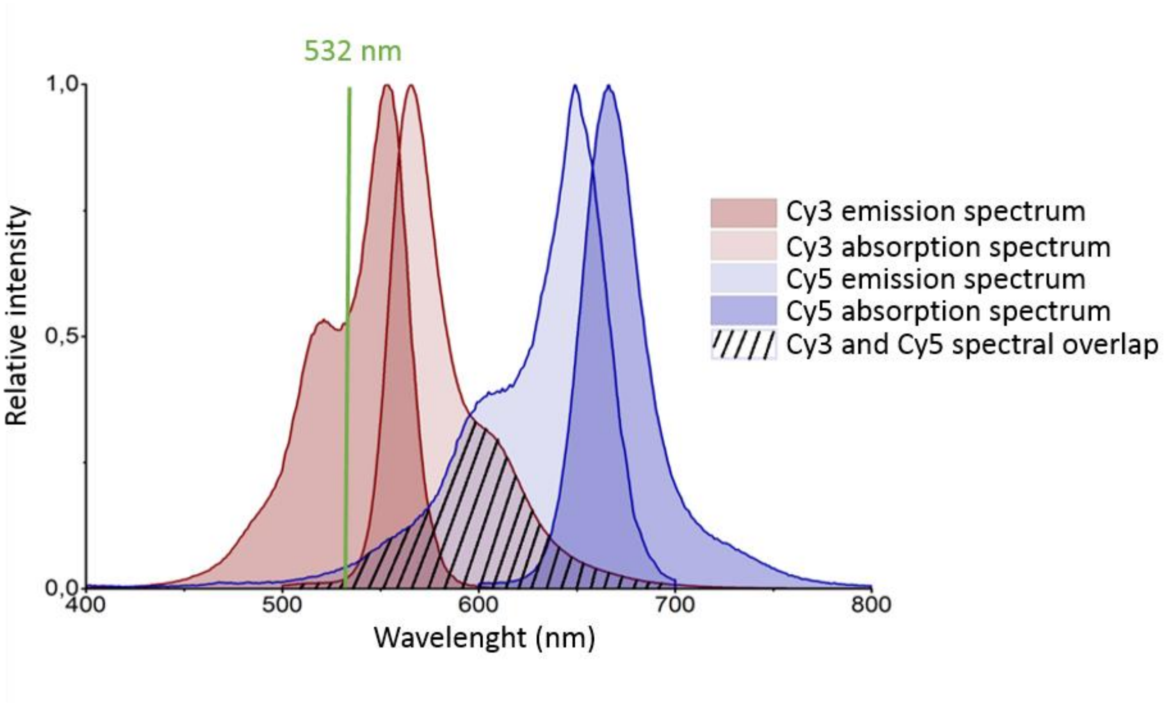
# 8 ABBREVIATIONS

| | |
|---|---|
| 3'E, 5'E | i-motif conformers |
| 5'mC | 5-methylcytosine |
| CD | circular dichroism |
| CGI | cytosine-guanine island |
| CpA | cytosine-adenine dinucleotide |
| CpC | cytosine-cytosine dinucleotide |
| CpG | cytosine-guanine dinucleotide |
| CpT | cytosine-thymine dinucleotide |
| Cy3, Cy5 | cyanine dyes |
| FRET | fluorescence resonance energy transfer |
| IDDM | insulin-dependent diabetes mellitus |
| ILPR | insulin-linked polymorphic region |
| INS | insulin gene |
| LSQR | an algorithm for sparse linear equations and sparse least squares |
| PDF | probability distribution function |
| PEG | polyethylene glycol |
| RASP | recurrence analysis of single particles |
| RET | resonance energy transfer |

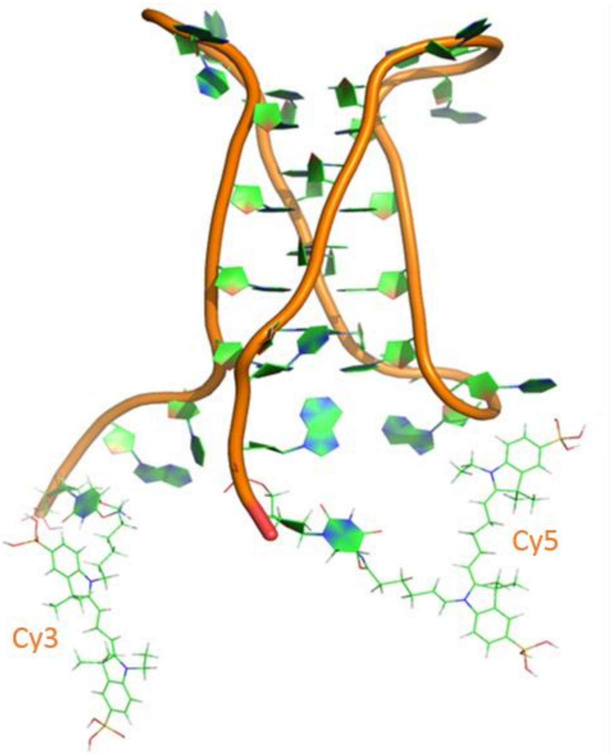
# 9 SUPPLEMENTARY MATERIAL



**I**. Schematic 3D visualization of the positions of 5-methylcytosines in a putative i-motif structure in (a) $\mathbf{C_0}$, (b) $\mathbf{C_{1\text{-}4}}$, (c) $\mathbf{C_{1,4}}$, (d) $\mathbf{C_{2,3}}$, (e) $\mathbf{C'_{1,4}}$ and (f) $\mathbf{C'_{1}}$ sample modifications; the core of the ILPR i-motif structure was taken from [6] and optimized for our needs by coarse-grained molecular dynamics [58].



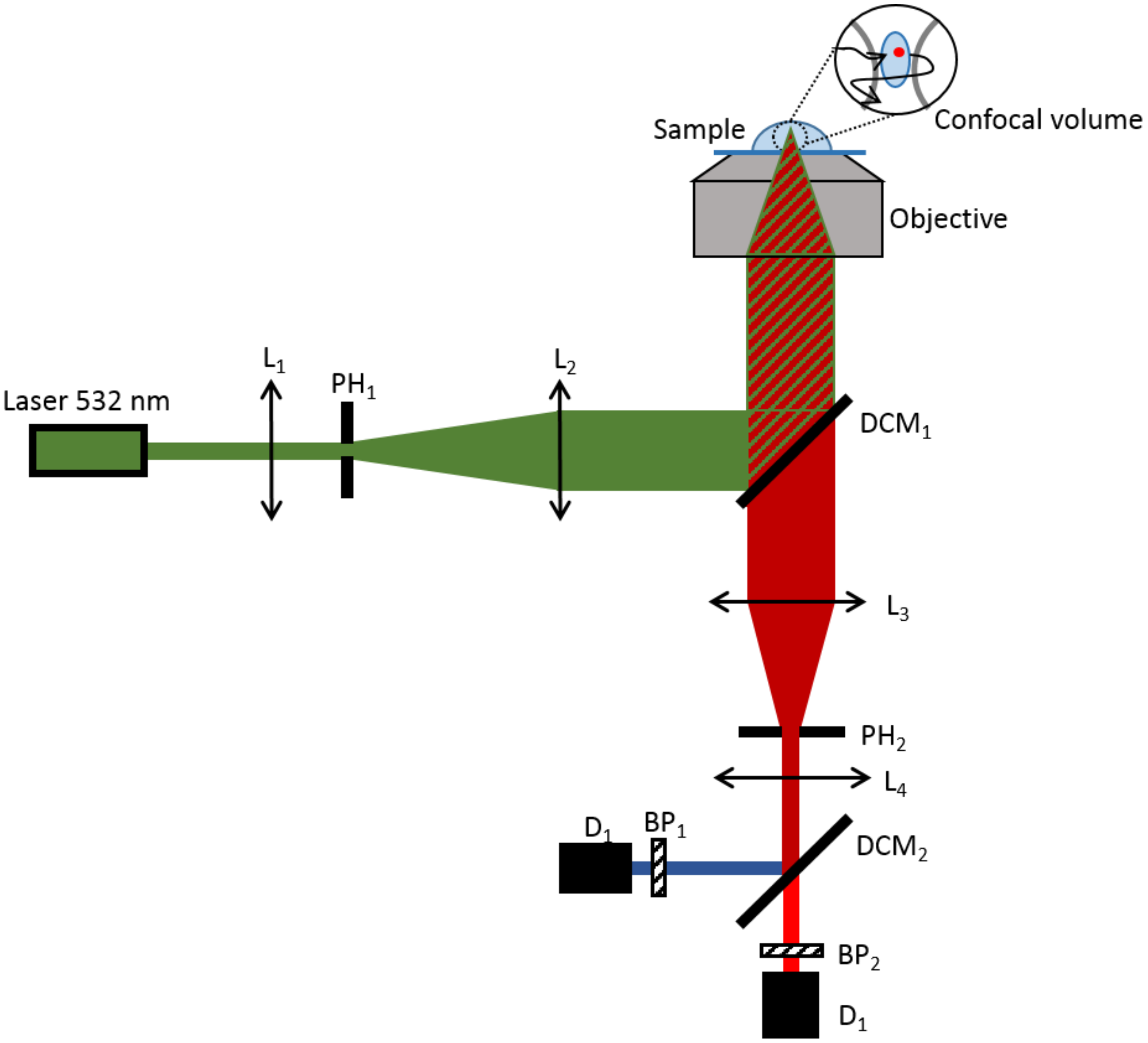
**II**. Absorption and emission spectra of Cy3 and Cy5 with indication of their spectral overlap and highlighted excitation wavelength of 532 nm; provided by Integrated DNA Technologies, Inc.



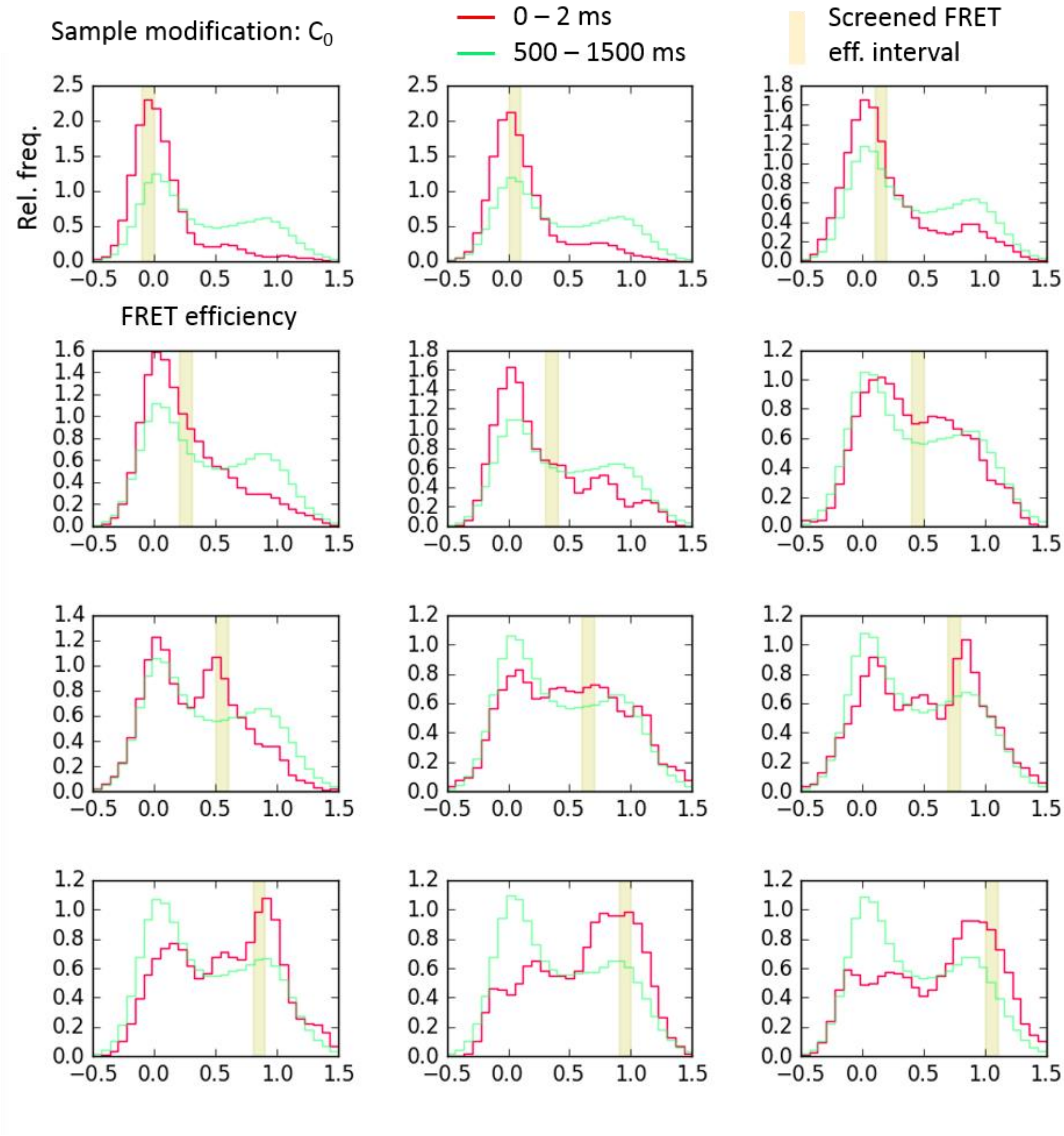
**III**. 3D visualization of the non-modified DNA sample sequence with attached cyanine dyes and folded into an i-motif; created in Pymol [59].

**IV**. Parameter setting for CD measurements.

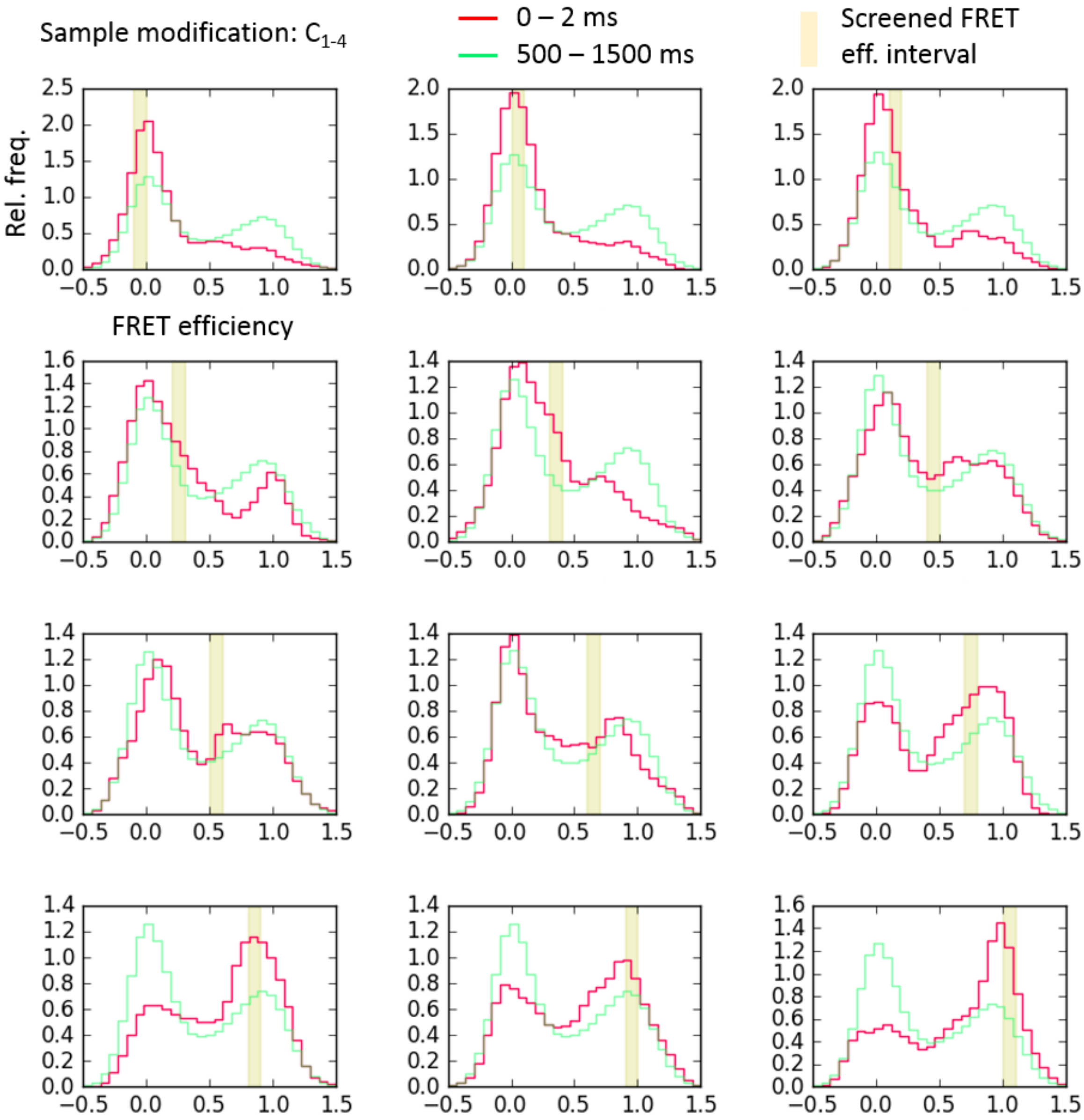
| Wavelength range (nm) | Step resolution (nm) | Band width (nm) | Sensitivity (mdeg) | Response (s) | Speed (nm/min) | Accumulations (units) |
|---|---|---|---|---|---|---|
| 350 - 200 | 1 | 2 | 20 | 0.25 | 200 | 5 |



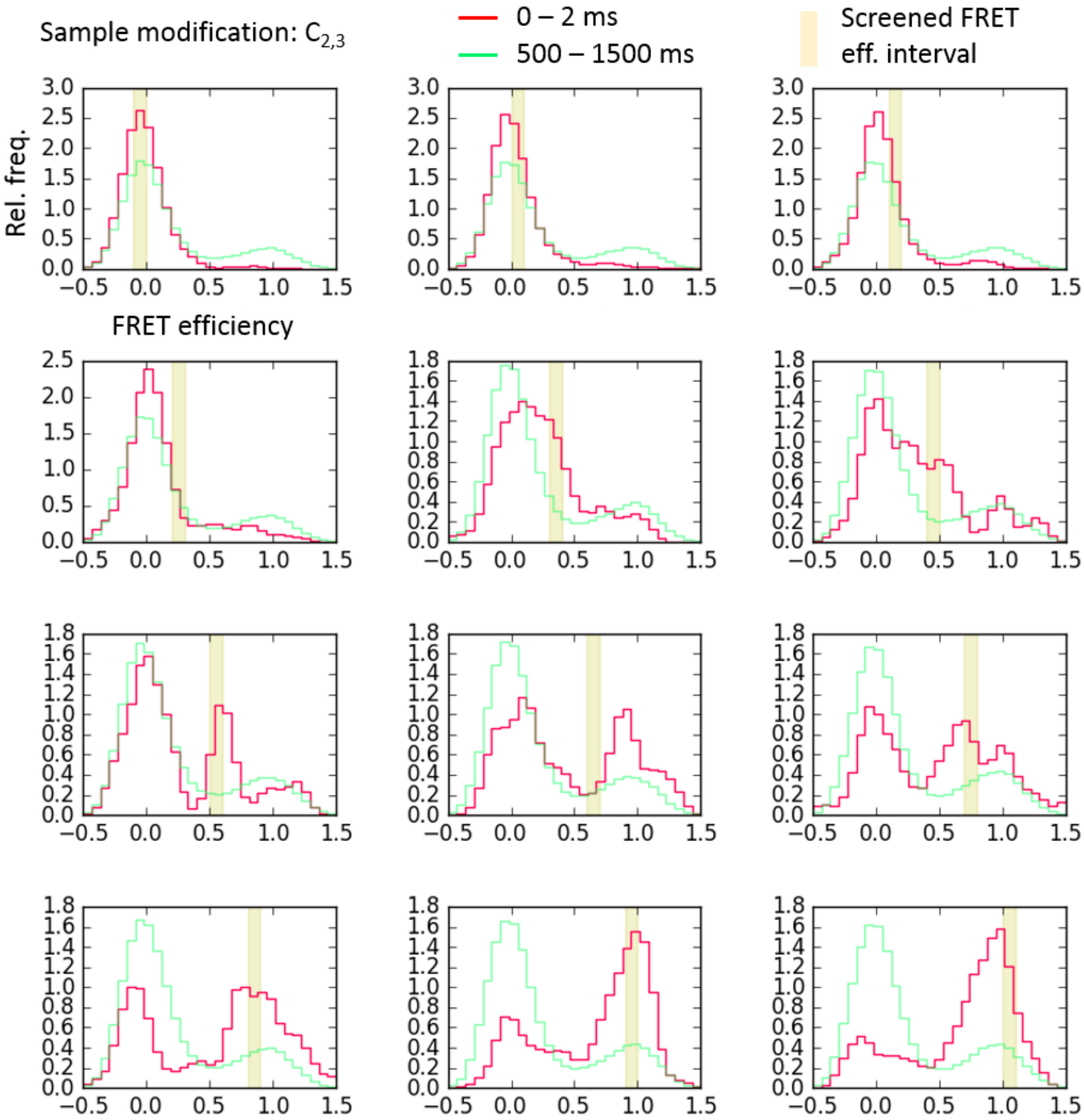
**V**. Home-built apparatus for FRET experiments: $L_1$ – $L_4$ , lens (Thorlabs); $PH_1$ (15 μ) and $PH_2$ (35 μ), pinholes (Thorlabs); $DCM_1$, long pass filter in microscope cube (XF3094 600ALP, Horiba); $DCM_2$, donor-acceptor emission separating dichroic mirror (zt640, Chroma); $BP_1$, acceptor emission band pass filter (FF02-685/40-25, Semrock); $BP_2$, donor emission band pass filter (Id01-640/8-12.5, Semrock); $D_1$ and $D_2$ , detectors (single photon avalanche diodes, SPCM QR-16, PerkinElmer); objective (Olympus), corrected to infinity, oil immersion, NA = 1.42.



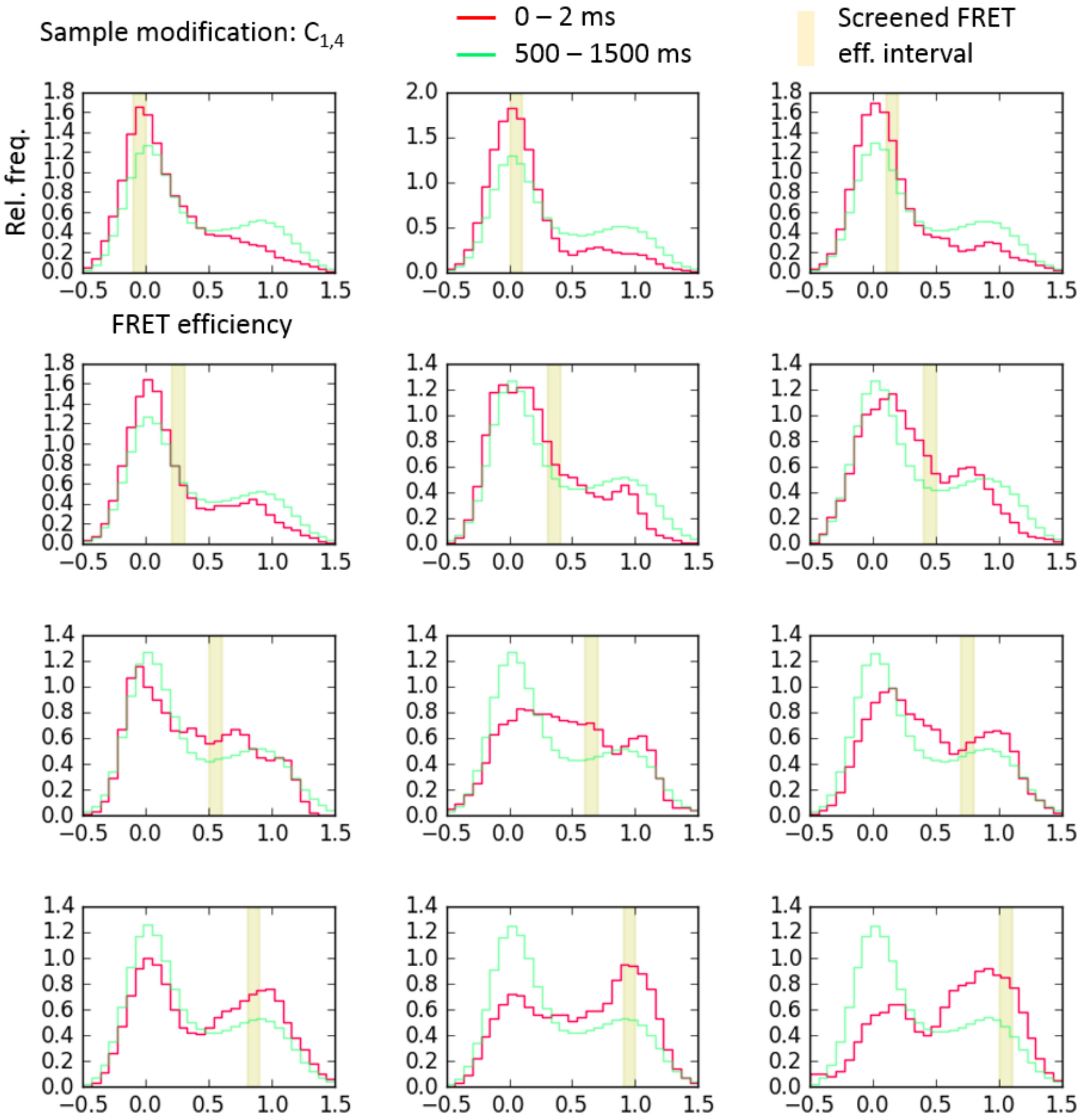
**VI.a**. Screening of initial FRET efficiency intervals; recurrence short-time histograms (red line), steady-state long-time histograms for comparison (green line).



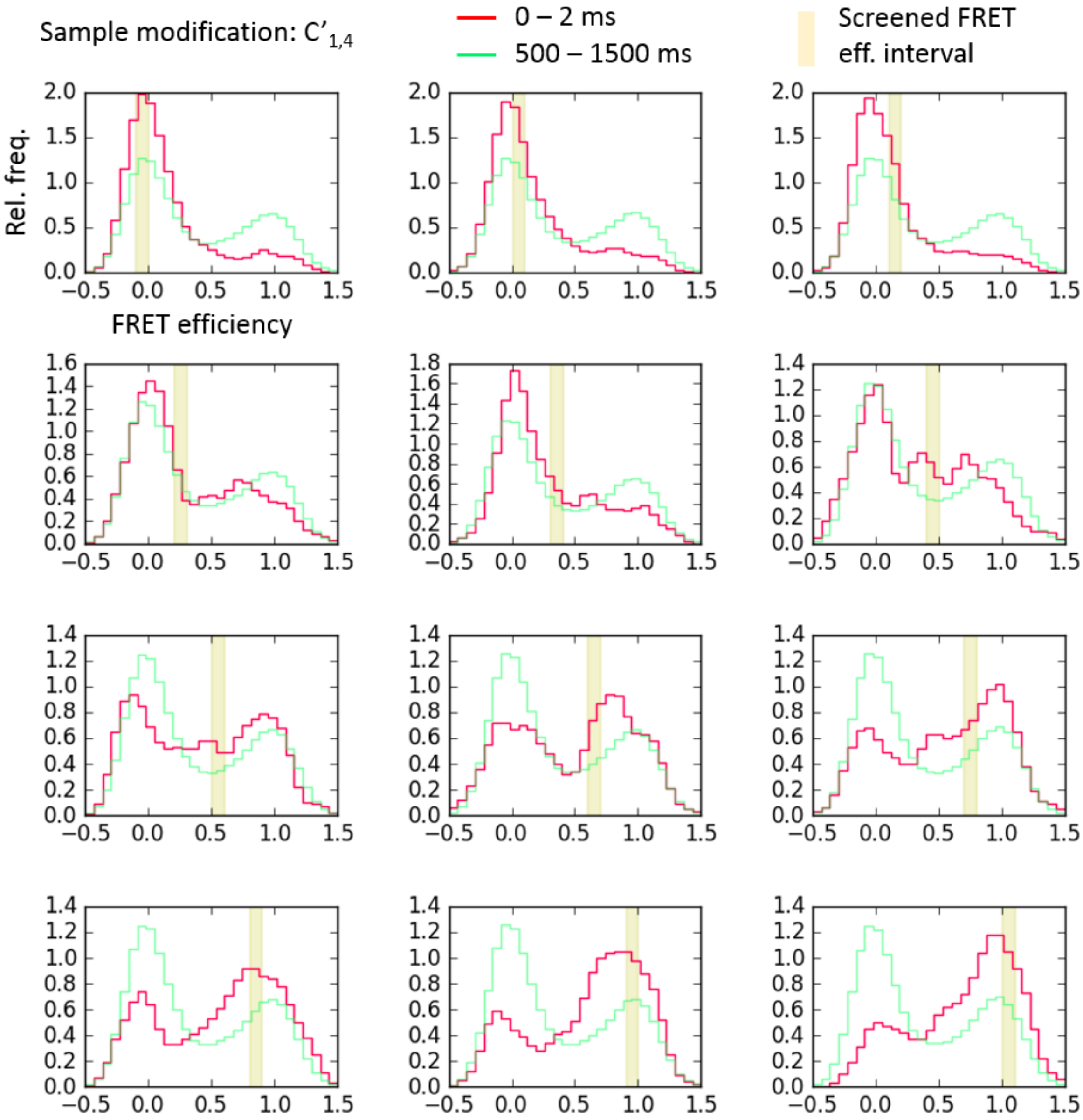
**VI.b**. Screening of initial FRET efficiency intervals; recurrence short-time histograms (red line), steady-state long-time histograms for comparison (green line).



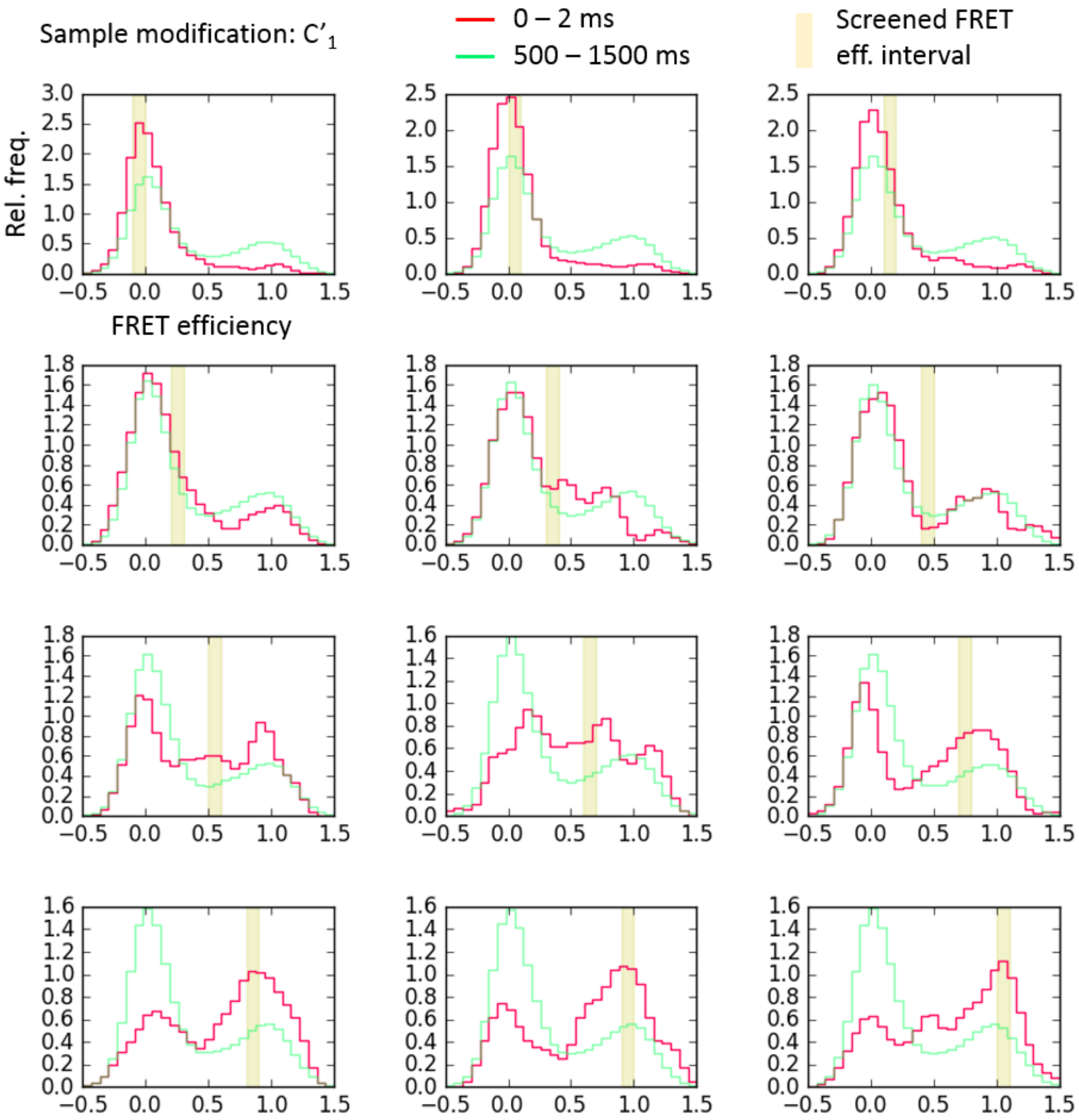
**VI.c**. Screening of initial FRET efficiency intervals; recurrence short-time histograms (red line), steady-state long-time histograms for comparison (green line).



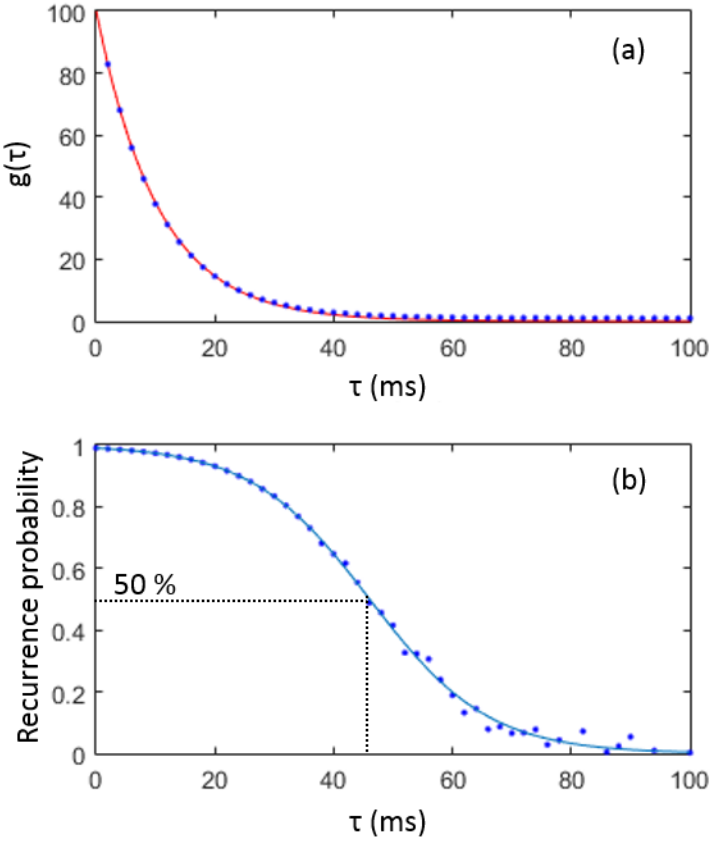
**VI.d**. Screening of initial FRET efficiency intervals; recurrence short-time histograms (red line), steady-state long-time histograms for comparison (green line).



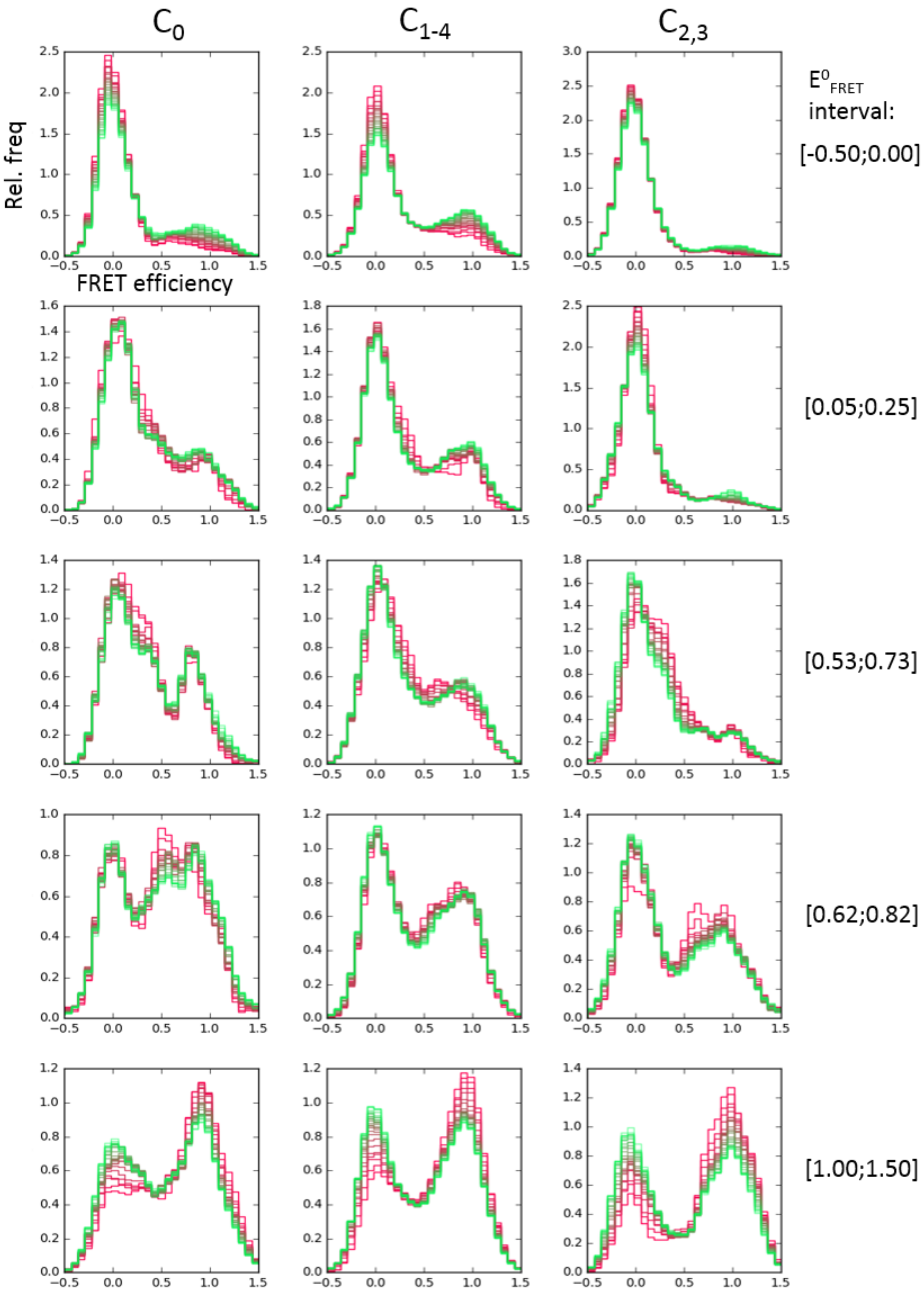
**VI.e**. Screening of initial FRET efficiency intervals; recurrence short-time histograms (red line), steady-state long-time histograms for comparison (green line).



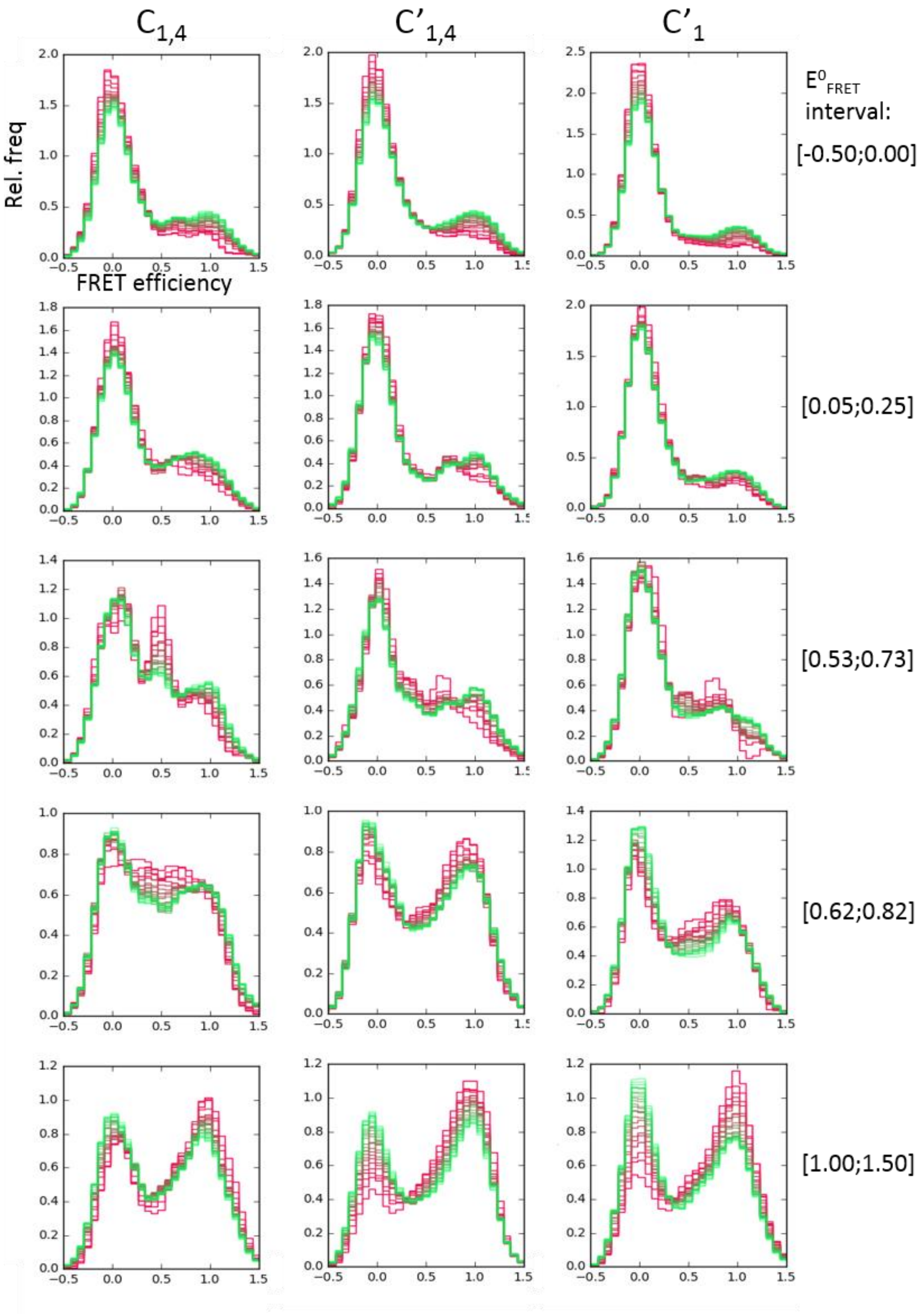
**VI.f**. Screening of initial FRET efficiency intervals; recurrence short-time histograms (red line), steady-state long-time histograms for comparison (green line).



**VII**. (a) Burst time autocorrelation function g(τ) depending on the time between two photon bursts τ; (b) the autocorrelation function converted into the same molecule burst probability with indicated time value at 50 % probability that two photon bursts originate from one molecule.



**VIII.a**. Recurrence transfer efficiency histograms for **$C_0$**, **$C_{1-4}$** and **$C_{2,3}$** modifications plotted through the first 27 ms showing the time development of five previously selected initial FRET efficiency intervals corresponding with recognized subpopulations; the time development goes from red to green.

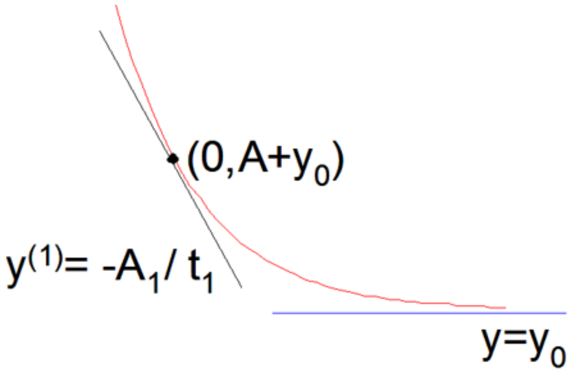


**VIII.b**. Recurrence transfer efficiency histograms for **$C_{1,4}$**, **$C'_{1,4}$** and **$C'_{1}$** modifications plotted through the first 27 ms showing the time development of five previously selected initial FRET efficiency intervals corresponding with recognized subpopulations; the time development goes from red to green.

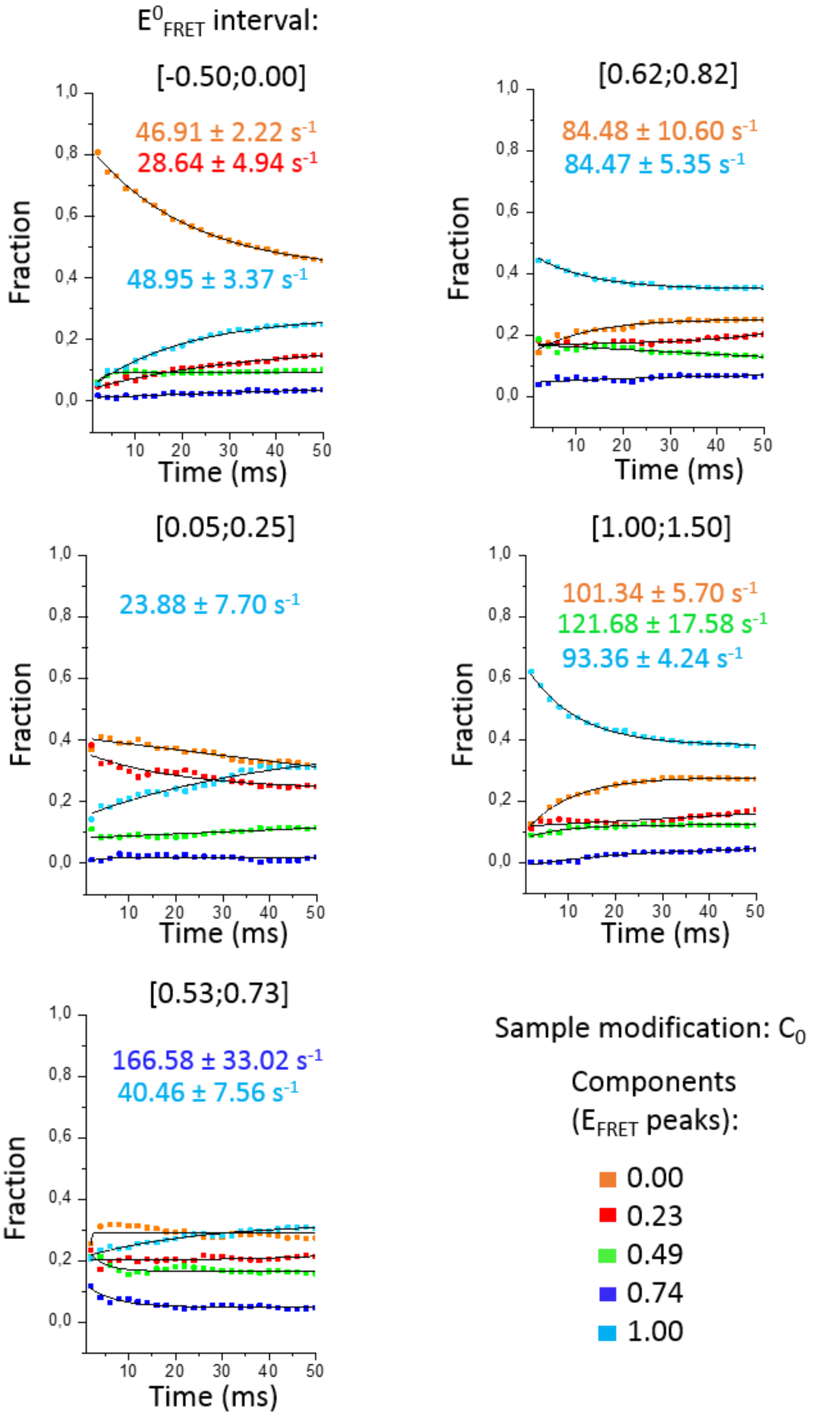
**Exponential fitting:**

Formula: $y = y_0 + A_1 e^{-\frac{x}{t_1}}$
$A_1$ ... amplitude
$y_0$ ... offset
$t_1$ ... decay constant [s]

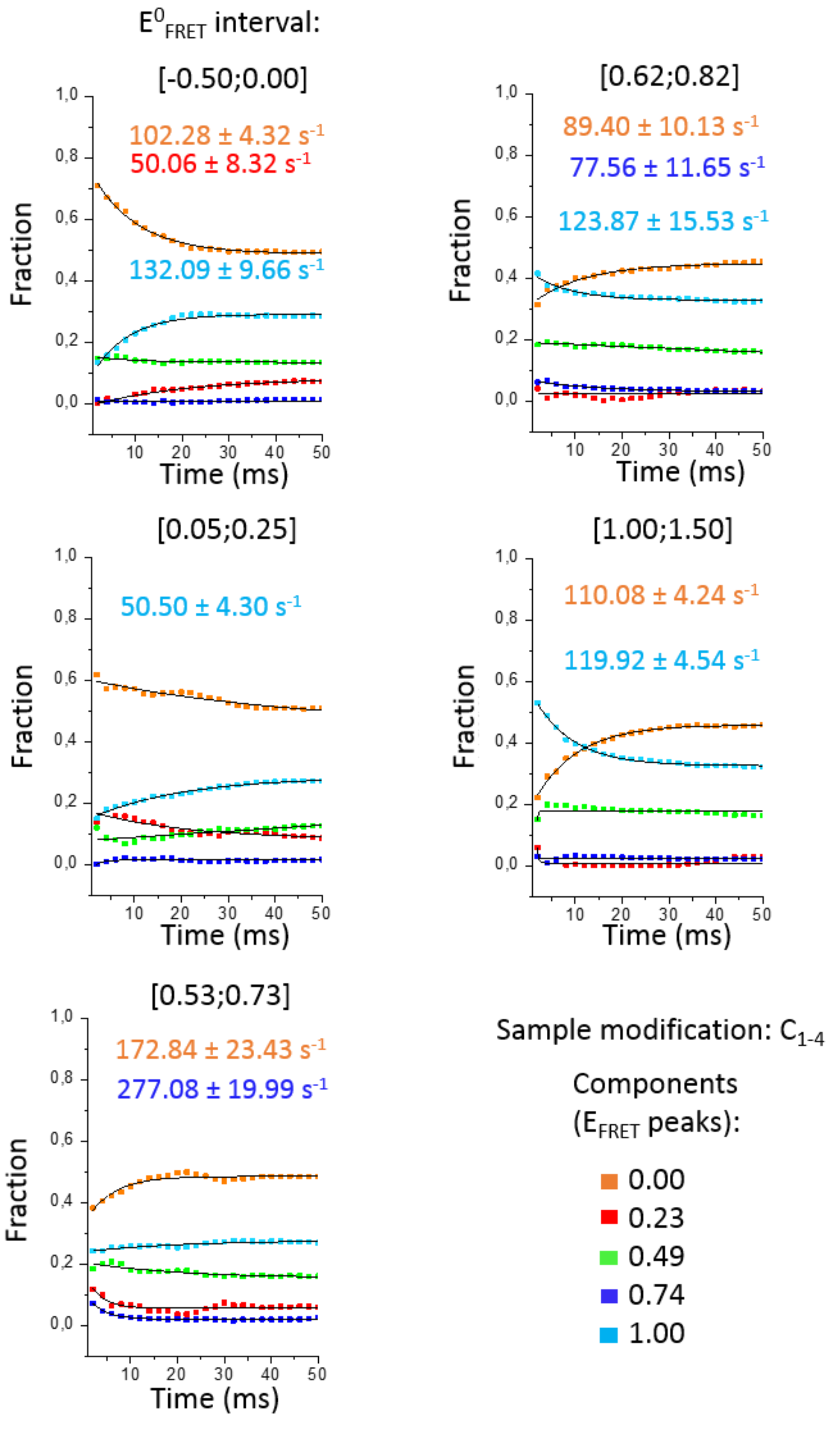
Derived parameters:
$k$ = 1/ $t_1$ ... rate constant [s$^{-1}$]
τ = $t_1$ ln(2) ... half life [s]



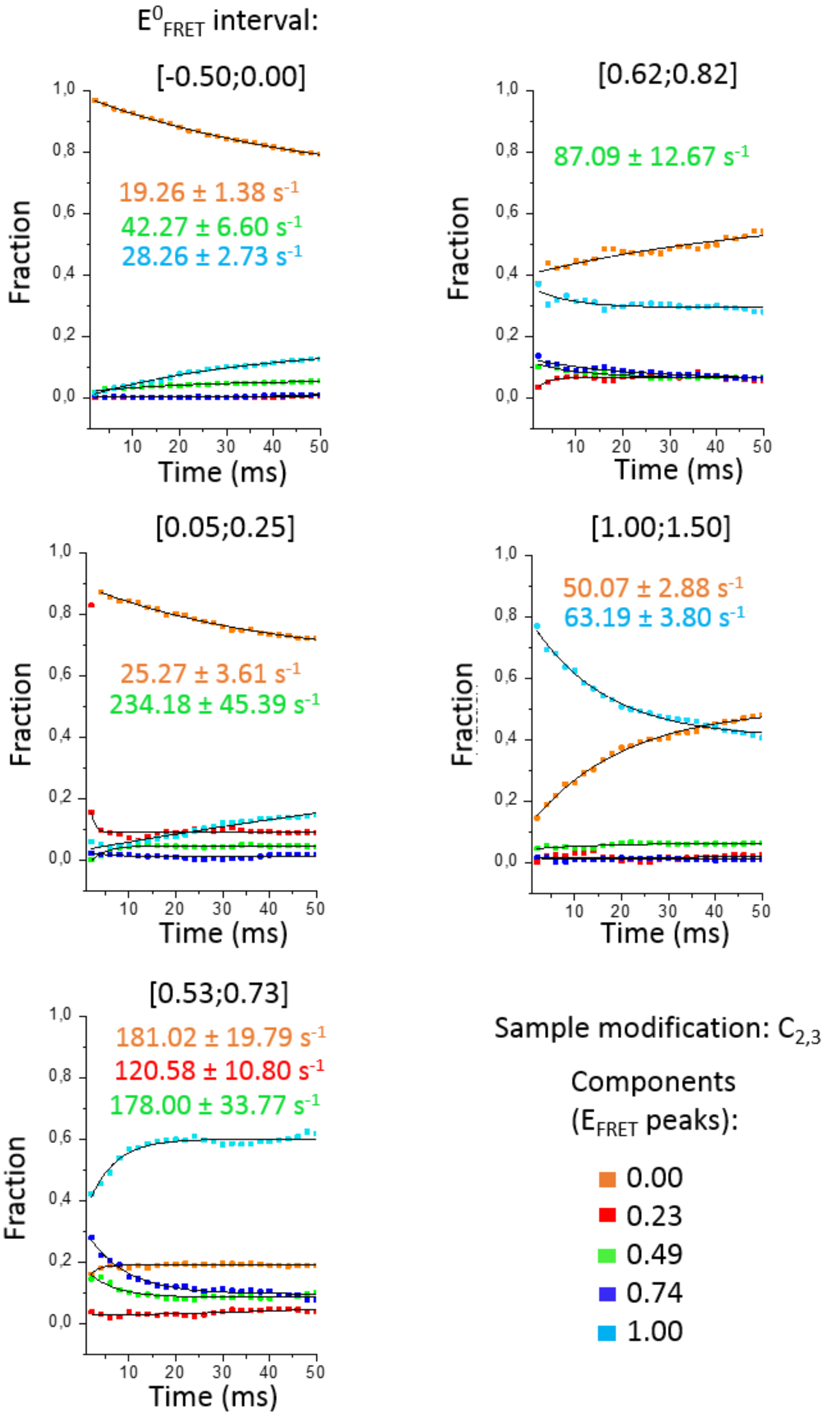
**IX**. Formula and parameters used in exponential fitting of a component development in time.



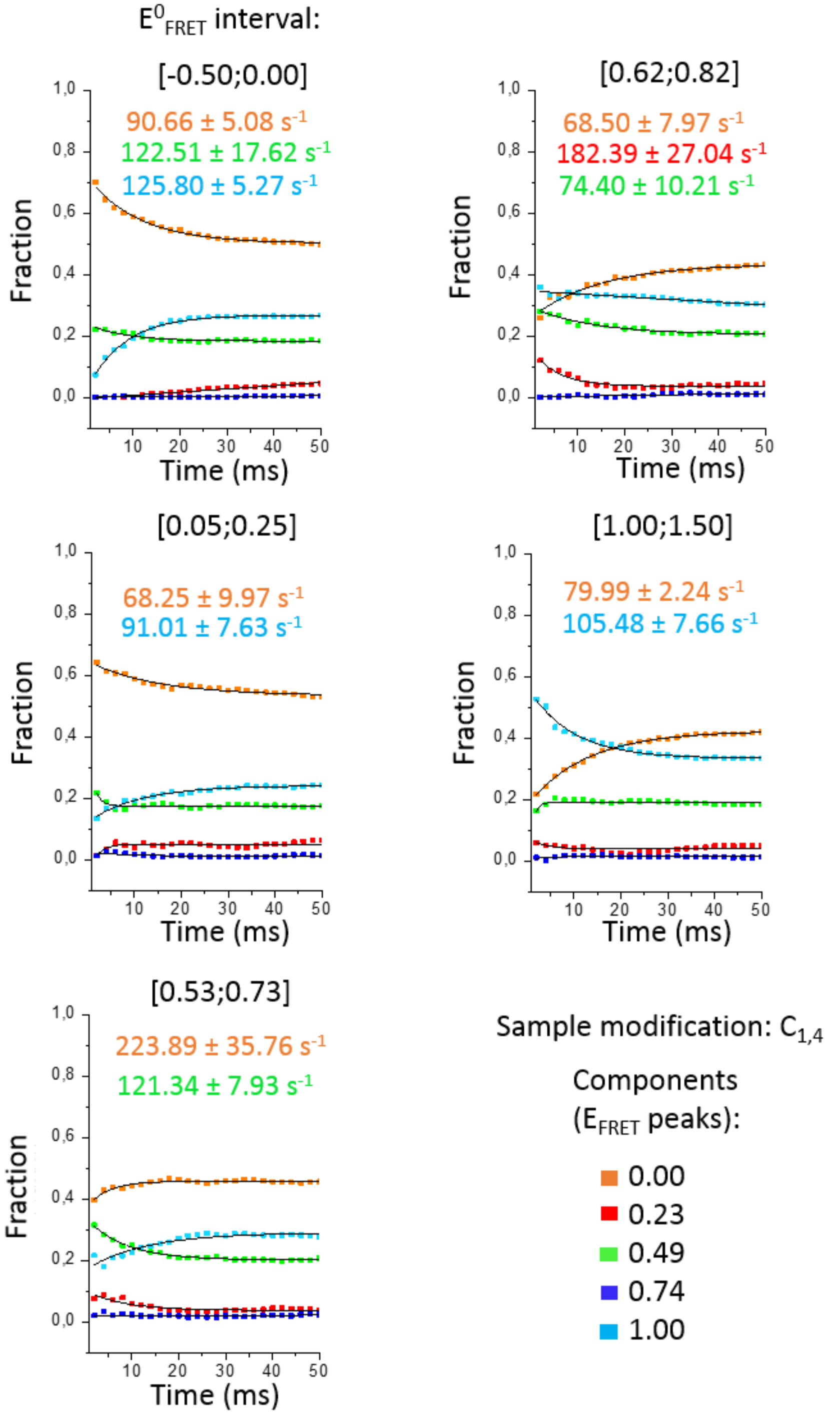
**X.a**. Time development of components obtained from PDFs of the RASP FRET histograms through 50 ms; exponentially fitted with calculated rate constants (SE ≤ 20 %).



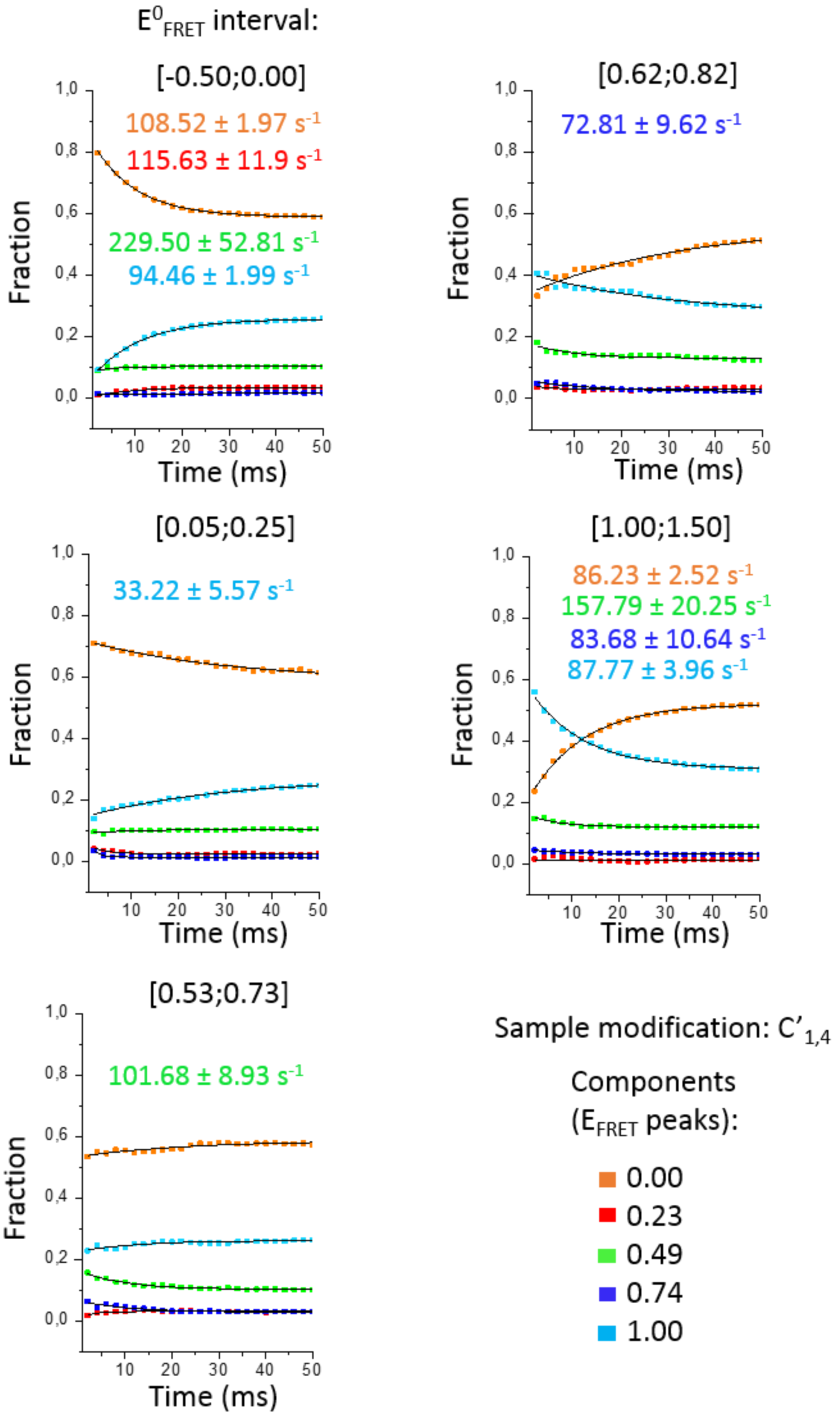
**X.**. Time development of components obtained from PDFs of the RASP FRET histograms through 50 ms; exponentially fitted with calculated rate constants (SE ≤ 20 %).



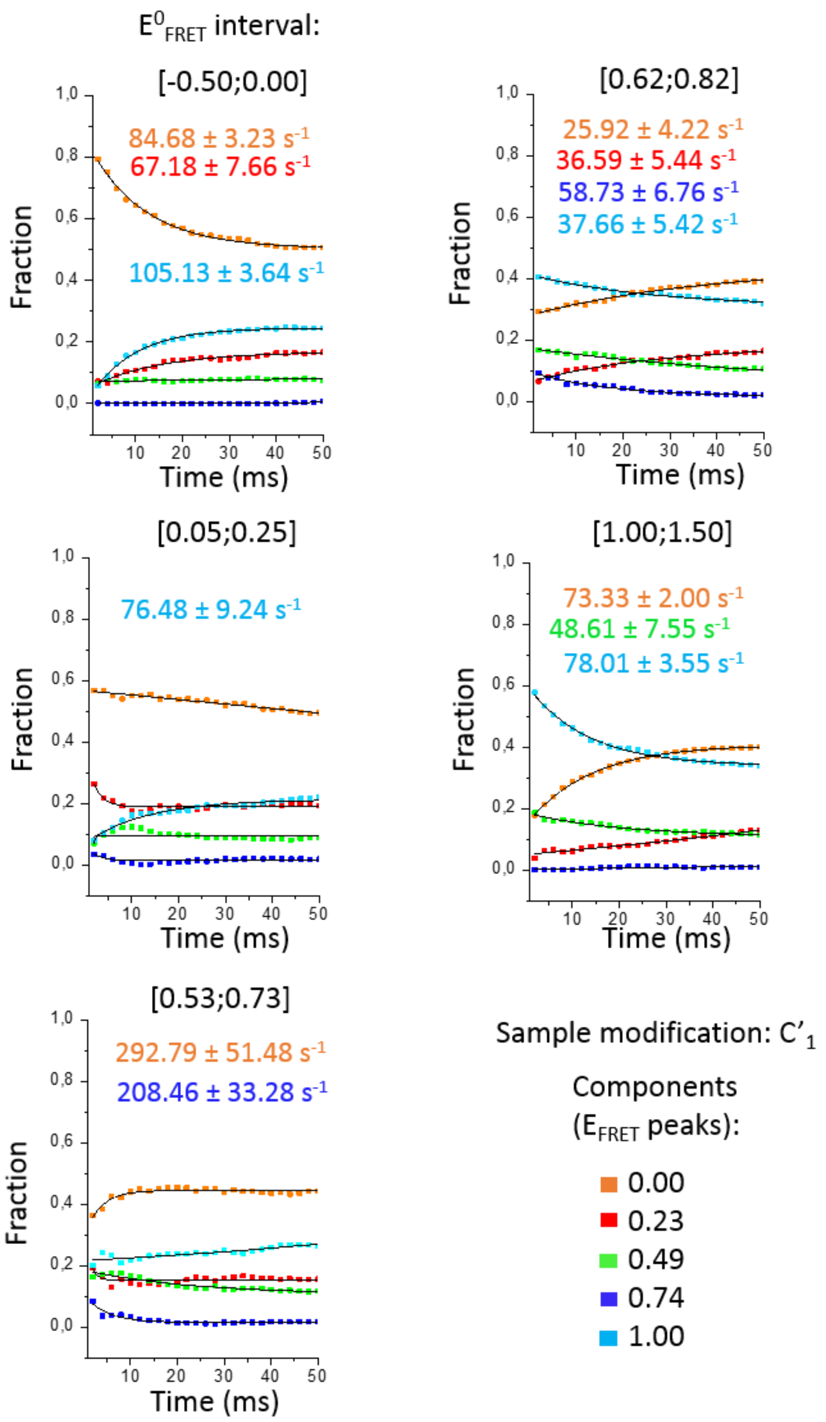
**X.c**. Time development of components obtained from PDFs of the RASP FRET histograms through 50 ms; exponentially fitted with calculated rate constants (SE ≤ 20 %).



**X.d**. Time development of components obtained from PDFs of the RASP FRET histograms through 50 ms; exponentially fitted with calculated rate constants (SE ≤ 20 %).



**X.e**. Time development of components obtained from PDFs of the RASP FRET histograms through 50 ms; exponentially fitted with calculated rate constants (SE ≤ 20 %).



**X.f**. Time development of components obtained from PDFs of the RASP FRET histograms through 50 ms; exponentially fitted with calculated rate constants (SE ≤ 20 %).