Fundamental aspects of the nucleic acid i-motif structures

5 AUTHORS, INCLUDING:

Anna Aviñó
Spanish National Research Council
84 PUBLICATIONS   **625** CITATIONS

SEE PROFILE

Ramon Eritja
Spanish National Research Council
359 PUBLICATIONS   **5,450** CITATIONS

SEE PROFILE

Carlos González
Spanish National Research Council
125 PUBLICATIONS   **2,046** CITATIONS

SEE PROFILE

Raimundo Gargallo
University of Barcelona
72 PUBLICATIONS   **1,148** CITATIONS

SEE PROFILE

Fundamental aspects of the nucleic acid i-motif structures

S. Benabou,[a] A. Aviñó,[b] R. Eritja,[b] C. González [c] and R. Gargallo [a*]

[a] Department of Analytical Chemistry, University of Barcelona, Martí i Franqués 1-11, E-08028 Barcelona, Spain.

[b] Institute for Advanced Chemistry of Catalonia (IQAC-CSIC), CIBER-BBN Networking Centre on Bioengineering, Biomaterials and Nanomedicine, Jordi Girona 18-26, E-08034 Barcelona, Spain.

[c] Institute of Physical Chemistry "Rocasolano", CSIC, Serrano 119, E-28006 Madrid, Spain.

ABSTRACT

The i-motif structure is formed in cytosine-rich sequences, being its building block the cytosine•cytosine+ base pair. This structure is particularly stable at pH values below the physiological one (~7.4) and, because of that, it has not attracted as much biological interest as other non-canonical structures such as the G-quadruplex. Nowadays, the proposal of potential roles *in vivo*, as well as nanotechnological applications, has produced an increasing interest in its study.

In this context, the present work provides an overall picture of the i-motif structure. Those aspects related with their formation and stability, such as chemical modifications or the interaction with ligands, are discussed. Special attention has been made to the i-motif structures that could have a hypothetical role *in vivo*, such as those present near the promoter region of several oncogenes.

INTRODUCTION

Recent genomic research has revealed that around 98% of biological DNA is comprised of non-coding regions, such as regulatory elements, introns, repeat sequences and telomeres, among others. Many types of non-coding DNA sequences have important biological functions, including the transcriptional and translational regulation of protein-coding sequences. Other non-coding sequences have likely, but not yet determined, functions. These non-coding regions often comprise repetitive sequences that are patterns of nucleic acid sequences (DNA or RNA) that occur in multiple copies throughout the genome. The most common DNA structure under physiological conditions is the B-DNA, which is a right-handed double helical structure with Watson-Crick base pairing. However, repetitive DNA sequences have the potential to fold into non-B DNA structures such as hairpin or left-handed Z-form under certain experimental conditions. Since the non-B DNA-forming sequences may induce genetic instability and, consequently, may cause human diseases[1], the molecular mechanism for their genetic instability has been investigated extensively.

G-rich sequences are able to fold into a non-B-DNA structure known as G-quadruplex. *In vitro* studies have determined the conditions at which the G-quadruplex may be formed and the spatial arrangement of bases in it. Because the fact that, at the physiological conditions of temperature, pH and ionic strength, G-quadruplex structures are stable, it must be expected the formation of these structures *in vivo*. In fact, evidence for the formation of G-quadruplex structures in the genome of mammalian cells by using appropriate ligands in a cellular context has been described recently[2–4].

C-rich regions have also the ability to form a folded structure known as i-motif (also named i-tetraplex or i-DNA) [5,6]. This structure is the only known DNA structure that consists of parallel-stranded duplexes held together through intercalated base pairs. Other structures, such as B-DNA or G-quadruplex are spatial arrangements of DNA strands held together by means of stacked base pairs. Unlike G-quadruplex, the formation of i-motif structures requires C base protonation to form the C·C$^+$ base pair. *In vitro* studies have shown that i-motif structures are stable at pH values lower than 7 at the physiological conditions of temperature and ionic strength. However, its stability is very low at neutral and basic pH values. Because of this, the hypothetical role of the i-motif structure in biological processes is uncertain and, consequently, much more effort has been done in the study of other non-canonical structures such as the G-quadruplex. In the last years, however, the sensibility of i-motif formation to narrow pH changes boosted the study of potential uses of this structure in the development of sensors and motors at nanoscale [7,8]. Finally, the isolation of proteins that specifically bind to C-rich sequences prompted again the study and discussion about its hypothetical biological role.

Few works have been devoted to review the advances in knowledge of the structure and solution properties of i-motif structures, including their potential role *in vivo*. Gilbert and Feigon included a section about the i-motif in their review about multistranded DNA structures [9]. The focus was mainly put on some striking features of a few structures previously determined by means of NMR or X-ray studies, like the parallel disposition of flanking base pairs on the i-motif core. The first review focused on the i-motif structure was published 14 years ago [10]. In that work, the latest discoveries about this structure in human telomeric and centromeric sequences, as well as the isolation of proteins that bind specifically to C-rich sequences were discussed.

The role of non-B-DNA structures, such as G-quadruplex or i-motif structures, as potential regulatory elements in transcription has been discussed recently [11]. It has been proposed that negative supercoiling may favour the formation of single strands in local unwounded regions of DNA. These single strands may form non-B-DNA structures which could play a role in the transcription regulation. This hypothesis is based on several evidences. First, the fact that most of the G-rich and C-rich sequences are in close proximity to the transcriptional start site suggest a role in transcription. In fact, ~43% of promoter regions in genes have the potential of form at least one G-quadruplex structure and hence, one i-motif structure. Second, the existence of proteins and ligands that recognize specifically the G-rich or C-rich regions suggest a role in transcription modulation. Third, the relatively low diversity observed among G-rich sequences in promoter regions suggests a conserved biological role.

Finally, Choi and Majima dedicated a short chapter to the i-motif structure in their extensive review about non-B DNA structures [12]. The existence of different species (i-motif-like and classical i-motif) is commented on the light of previous works, as well as the slow folding and unfolding kinetics associated to the formation of i-motif structures. In a later work, the most recent applications of fluorescence spectroscopy to the study of i-motif folding/unfolding processes have been reviewed by these two authors [13]. Special attention has been made to the application of these conformational processes in Nanotechnology, like the design of nanomachines, sensors or logic gates, among others.

In this framework, the present work reviews the recent literature about the fundamentals aspects of the i-motif structure, as well as the experimental methods used in its study. To our knowledge, recent studies dealing with the formation and stability of these structures have not been reviewed yet. Also, the latest results about the interaction with ligands and their potential presence *in vivo* are also reviewed. Advances in

nanotechnological applications of i-motif structures, such as the formation of metal nanoparticles, the design or nanomachines or molecular motors, and the development of sensing devices have not been included in this review.

THE I-MOTIF STRUCTURE

The building block of the i-motif structure is the base pair involving one neutral C and one protonated C at N3 (the C·C$^+$ base pair) bonded by three underline hydrogen bonds (Figure 1a). Based on NMR spectroscopy and theoretical calculations using the $(C_3TA_2)_3C_3$ sequence as model it has been proposed that N3···H$^+$···N3 bonds may be described as hydrogen bonds with asymmetric double-well potentials rather than a symmetric hydrogen bond with a single-well potential [14]. The formation of these hydrogen bonds produces a stronger base-pair interaction than the canonical G·C base pair, as denoted by *in silico* calculations [15]. Hence, the base-pairing energy (BPE) for the proton-bound dimer of cytosine (C·C) is 169.7 kJ/mol, whereas the BPEs of the canonical Watson–Crick G·C and neutral C·C base pairs are only 96.6 and 68.0 kJ/mol, respectively.

Figure 1. The i-motif structure. (a) the C·C$^+$ base pair, (b) 3D structure of an i-motif showing six C·C$^+$ base pairs in four strands (from PDB 1YBR). (c) Characteristic stacking pattern between adjacent C·C$^+$ base pairs in i-motif DNA. The top C·C$^+$ base pair from one parallel duplex is drawn in grey whereas the lower C·C$^+$ base pair from the second intercalated duplex is depicted in green. Hydrogen bonding interactions are drawn with broken yellow lines.

The name, i-motif, refers to the fact that is probably the only nucleic acid structure where base pairs are intercalated [5]. The i-motif structure may be formed from the spatial arrangement of C·C$^+$ base pairs involving C tracts present in one nucleic acid strand, producing a so-called intramolecular i-motif. On the other hand, it may be involve C tracts present in two or four independent nucleic acid strands, producing intermolecular i-motif structures. In any case, the C tracts are spatially arranged as a tetramer composed of two parallel-stranded duplexes that are interspersed in an anti-parallel way (Figure 1b). The depicted intramolecular i-motif contains several C·C$^+$ base pairs formed by the interaction of twelve Cs that are arranged in four tracts and into a single nucleic acid strand. In general, the hydrogen bonded to N3 in each C·C$^+$ base pair is not equidistant to the two N3 atoms, but adopts a position with the largest distance to the hydrogen of the next C·C$^+$ base pair [14].

The interaction of consecutive base pairs is mainly provided by the stacking of the exocyclic carbonyl and amino groups that are oriented with opposed dipoles, and is not observed between the aromatic heterocycles of the bases (Figure 1c). This apparent absence of stability provided by stacking interactions between the bases can be overcome by the formation of a systematic intermolecular C-H···O hydrogen bonding network between the deoxyribose sugar moieties of antiparallel backbones in the four-stranded molecule [16,17]. Overall, the base-pair distance is only 3.1 Å, a stacking much closer than in B-DNA (3.4 Å) and similar to A-DNA (2.9 Å). The helical twist between adjacent C·C$^+$ pairs (12-16°) is also smaller than in the case of B-DNA (36°) or A-DNA (32.7^o) [18]. This spatial arrangement produces the existence of two broad and flat major grooves and two extremely narrow minor grooves (Figure 1b). According to the spatial arrangement of the C·C$^+$ base pairs, i-motif structures have been classified in two classes: those with the terminal C·C$^+$ base pair at the 3' end (3'E intercalation topology), and those with the terminal base pair at the 5' end (5'E intercalation topology) (Figure 2) [19,20].

Figure 2. A C-rich sequence may fold into, at least, two different i-motif intercalation topologies, 3'E (up) and 5'E (down). For the $C_3TA_2)_3C_3$ sequence, the population ratio is 1:4 at 15°C. Reprinted with permission from reference [21].

In addition to the formation of three hydrogen bonds at each C·C⁺ base pair and the favourable interaction of consecutive pairs, the decrease of the <u>negative charge on the backbone</u> because of the protonation of C facilitates the association of the four strands. However, these considerations apply equally to duplex C·C⁺ structures and do not explain the formation of an intercalated structure. Part of the differential stability of the *i*-motif was initially explained in terms of close contacts between sugars in the narrow grooves of the structure[5]. These could give rise to favourable van der Waals energies (as observed from molecular dynamics simulations) and, as they exclude water, may represent hydrophobic contacts between deoxyribose sugars. As shown below, this effect would help to explain the lower stability of RNA i-motifs compared to their DNA counterparts [22].

Sequences such as $ACTC_3T_2CTC_2TCTCTA$ or $TCTCTC_2TG_2TC_2TC_2$ are examples of C-rich sequences that form duplex C·C⁺ structures lacking of the intercalated nature of i-motif [23]. In this case, it was observed the formation of parallel-stranded homoduplexes at pH 4–5.5, which were held by C·C⁺ base pairs. It was shown that the insertion of guanine or T into an oligomeric C sequence made the formation of the i-motif unfavourable.

INSTRUMENTAL APPROACHES TO STUDY THE I-MOTIF STRUCTURES

There are different approaches to study the formation and stability of i-motif structures. First studies demonstrated the formation of the i-motif structure by stretches of deoxycytidine at acid pH in base of <u>NMR</u> measurements. The study of tetrameric i-motif structures showed similar imino proton spectra, NOEs, and exchange properties that indicate that all the tetramers studied at that time had similar structures. It was also shown that the amino protons provide crucial information, particularly on the dynamical symmetry of the C·C⁺ pairs and on the lifetime of the open state, which is apparently at least 100 times longer than in Watson-Crick duplexes [24–26]. The imino protons involved in the C·C⁺ base pair show a characteristic NMR signal at ~15 ppm, whereas signals between ~12 and ~14 ppm are related to Watson-Crick G·C and A·T base pairing. The presence of signals between ~10 and ~12 ppm is usually related to additional non-canonical base pairs, such as T·T [27] or G·T mismatches [28,29]. These additional base pairs are usually located in the loop regions, but can be also found in the middle of C-tracts [30]. Tetrameric i-motifs formed by four short oligonucleotide strands normally render relatively simple NMR spectra and their structural characterization is, in general, straightforward. However, in the case of more complex (and presumably more relevant) sequences the assignment of the NMR spectra is much more challenging. Due to severe spectral overlap, many NMR studies could only be completed by introducing chemical modifications (usually methyl-cytosines [26] or site-specific isotopic labels [31].

Apart from NMR, one of the first techniques used to unravel the i-motif structure was <u>X-ray diffraction</u>. Hence, the crystal structure of C_4 solved at 2.3 Å resolution revealed the formation of a four-stranded molecule composed of two intercalated duplexes [32]. The information provided by X-ray crystallography may be significantly different from that obtained from solution-based techniques, such as NMR. Condensation of single molecules from solution into crystals represents a transition between distinct energetic states. In

solution, the atomic interactions within the molecule dominate. In the crystalline state, however, a set of additional interactions are formed between molecules in close contact in the lattice, the packing interactions. This fact was shown by Berger *and col.* in the study of the crystal structures of C_3T, TA_2C_3, C_3A_2T, and A_2C_4 [33]. These molecules showed intercalated C segments that were similar in their geometry, even though the sequences crystallized in different space groups.

UV molecular absorption spectroscopy may be a useful technique to observe the changes in protonation and stacking of the nitrogenated bases involved in the formation of i-motif. In this region, the protonation of Cs produces hyperchromicity at wavelengths from 275 to 300nm, as well a shift of the maximum wavelength from ~262 (for neutral C) to ~275nm (for protonated C). Hence, monitoring of absorbance changes in 275-295nm range is a nice way to study the stability of the i-motif as a function of temperature or pH changes. Due to the spectral characteristics of protonated and neutral Cs, the shape of the thermally-induced unfolding of i-motif structures monitored by molecular absorption spectroscopy is pH dependent. At pH values higher than the pK_a of C, the absorbance at 295nm decreases upon unfolding of the i-motif, whereas the opposite behavior is observed at pH values lower than the pKa of C. At pH values near the pK_a, no spectroscopic changes are observed upon unfolding [34]. In the case of melting experiments, it is possible to determine the T_m value and thermodynamic parameters such as ΔH and ΔS associated to the unfolding process. These parameters are calculated by using the van't Hoff equation, assuming a two-state process and values of ΔH and ΔS that are independent of temperature. It is possible to know the multimeric nature of the i-motif from determination of melting temperatures. If the melting of an i-motif-forming oligonucleotide show a concentration-independent profile, the i-motif will have intramolecular pairing, i.e., it is formed from the folding of only one, long strand. On the other hand, if the melting spectra show a concentration-dependent profile a bi- or multi-molecular association must be considered. When the i-motif is folded intermolecularly, T_m increases with the concentration.

Probably, the simplest way to detect the formation of the i-motif structure is recording the circular dichroism (CD) spectrum because it shows two characteristic negative and positive bands at ~265 and ~285 nm, respectively. Manzini *et al.* used CD to demonstrate the pH-induced formation of C-rich sequences in the human telomere and in the promoter region of the *k-ras* gene [35]. The exact position of these bands is also depending on the nature of bases located at the loops. The stability of the i-motif in front of temperature or pH changes is usually studied by monitoring ellipticity around 285nm. As example, the CD signal at this wavelength has been used to monitor the light-driven DNA conformational switch of the $(C_3TA_2)_3C_3$ sequence in presence of molecular malachite green carbinol base, a light-induced hydroxide ion emitter [36]. A related CD technique, the synchrotron radiation circular dichroism (SRCD), has been proposed for the study of i-motif structures [37]. The advantage of SRCD is the large available photon fluxes in the vacuum UV region where absorption is strong, which allows the exploration of excitation energy for the electronic coupling over a broad wavelength region. By using this technique, SRCD spectra of simple strands of C at different pH values were measured. The protonation state of bases determined the folding motif and as a result the extent of electronic coupling between bases. Different electronic couplings were found depending on the protonation of bases: single strands of all-protonated bases display no coupling, i-motifs have electronic coupling within a hemi-protonated base-pair and likely also nearest-neighbour coupling along the strand, and single strands of all-neutral bases have nearest-neighbour couplings along the strand.

Molecular fluorescence-based techniques have been also used to study the structure and kinetics of i-motif structures. Because canonical nitrogenated bases show little intrinsic fluorescence, C-rich sequences under study are usually labelled with appropriate fluorescent (or quencher) ligands at the 5' and/or 3' ends. Under

appropriate experimental conditions, the C-rich sequence folds allowing a closer contact between both labels giving rise to changes in the fluorescence of the bonded ligands. Lee *et al.* described the attachment of a pyrene moiety to the 5' end of a telomeric C-rich sequence. The high fluorescence of this molecule in the unfolded strand is strongly quenched upon stacking on the i-motif core [38]. The fluorescence almost disappears after addition of the complementary G-rich sequence and subsequent Watson-Crick duplex formation. <u>Fluorescence Resonance Energy Transfer</u> (FRET) has been shown to be a sensitive approach, provided that appropriate fluorophores are chosen [13,39,40]. In this technique, the excited state energy of a fluorescent donor chromophore is transferred to an unexcited acceptor chromophore. This yields quenched donor and increased acceptor fluorescence, being the efficiency dependent on distance and dynamics of the structure. Fluorophores such as FAM or TAMRA may be used with this purpose [41]. Finally, it has been shown that DNA sequences functionalized with pyrene moieties at both 5' and 3' ends show fluorescent properties that allow the study of the thermal stability of i-motif structures [23,42]. Another molecular fluorescence-based approach is the so called <u>molecular beacons</u>. Here, the fluorescence emitted by a label attached to one end of the DNA is quenched by a functional group attached to the other end. As example, BODIPY and DABCYL attached at the 5' and 3' ends of an i-motif-forming sequence have been used as fluorophore and quencher, respectively [43].

<u>Fluorescence Correlation Spectroscopy</u> (FCS) may provide information about the diffusion of the molecule as a function of pH, as well as information about the intrachain contact formation. The typical FCS setup consists of a laser line that is reflected into a microscope objective by a dichroic mirror. The laser beam is focused in the sample, which contains fluorescent molecules in such high dilution, that only a few are within the focal spot (usually 1–100 molecules in one fL). When the particles cross the focal volume, they fluoresce. This light is collected by the same objective and, because it is red-shifted with respect to the excitation light, it passes the dichroic mirror reaching a detector. Conclusions on physical phenomena have to be extracted from the measured signals with appropriate previously proposed models. The parameters of interest are found after fitting the autocorrelation curve to modelled functional forms [44].

<u>Time-resolved transient absorption and emission spectroscopies</u> have been also used to study the formation of i-motif structures [45]. These techniques take advantage of the fact that the excited states of cytidine and deoxycytidine multimers decay orders of magnitude more slowly than excited states in the corresponding monomers. Hence, variation in nucleic acid secondary structure profoundly affects the photophysical properties. It has been described that C-rich DNA sequences at physiological conditions of pH and ionic strength have similar excited state dynamics as the acid form of C-rich RNA sequences [45].

Figure 3. A typical force-extension curve obtained from the mechanical unfolding of the secondary structure of the TGTC$_4$ ACAC$_4$TGTC$_4$ACA sequence at pH 5.5. The unfolding event (~5 nm) is highlighted by a dashed green circle. Black curve is the fitting of the relaxing curve. Inset is the schematic of the laser tweezers experiment [46-48].

In a recent work, the study of complex transition kinetics in G-quadruplex and i-motif structures has been reviewed [49]. The single-molecule techniques used to investigate the kinetics of non-B DNA structures are shown with a special emphasis on the current progress in the mechanical manipulation and ligand binding of non-B DNA structures including i-motifs. The three major techniques used for the mechanical unfolding of biomolecular structures (AFM, laser tweezers, and magnetic tweezers) are explained in detail. The use of

laser-tweezers has been reported to investigate the structures formed in C-rich regions. Unlike traditional methods such as CD, NMR, UV/Vis… this method can reveal bimolecular structures in a highly dynamic fashion. This method is sensitive in identifying small populations, such as intermediates that can be formed during a folding process. Thanks to laser-tweezers an i-motif and a partially folded triple-helix-like structure have been shown to coexist in the C-rich human ILPR (insulin linked polymorphic region) oligonucleotide [46] (Figure 3).

Br_2 protection experiments are carried out to identify the specific C residues of a sequence that are involved in base pairings and intercalation to form the i-motif structure. Bromine is known to react selectively with double bond of the C within DNA, resulting in 5-bromodeoxycytidine. In particular, Br_2 reacts with the C residues in a single-stranded region 10-fold higher than those in the i-motif core. Hence, C residues in the loop regions are more reactive to Br_2 than other C residues involved in base pairing and intercalation, allowing the deduction of those specific C residues required for base pairing and intercalation in the i-motif structure [50].

Differential Scanning Calorimetry (DSC) is a technique that has been also used to determine mode-independent unfolding thermodynamics [51,52]. The main advantage over UV-monitored melting experiments carried out at just one wavelength is the unravelling of processes involving more than two steps, with the consequent determination of thermodynamic parameters such as the change in enthalpy. A disadvantage, however, is that the deconvolution of a DSC profile, i.e., the proposal of a given number of steps through which the folding process takes place, may be uncertain.

Separation techniques, such as non-denaturing polyacrylamide gel electrophoresis (PAGE), have been used extensively to demonstrate the mono or multimolecular nature of the i-motif formed [34,35,53]. Recently, PAGE has been used to resolve mixtures of dimeric, tetrameric and supramolecular structures formed from the tetramers [54]. Other separation techniques, such as Size-Exclusion Chromatography (SEC) have been used to determine the number of strands involved in the formation of multimeric i-motif structures, such as $T_2C_8T_2$, TC_5 or $(C_3TA_2)_3C_3$, a fragment of the human telomere [24,55]. As the elution time depends on the hydrodynamic volume which, in turn, is a function of the number of residues and of the spatial structure, the SEC system must be calibrated using an appropriate set of standards of known number of residues and structure before the unknown structures are analysed. Recently, SEC has been used to monitor the formation of dimers, tetramers and supramolecular structures from monomers along time [56].

Other techniques employed include Raman spectroscopy or the synchrotron small-angle X-ray scattering (SAXS). Raman spectroscopy has been used to detect the formation of i-motif structures as well as to quantify the protonation of C bases, both in solid and in solution phase [57]. Using SAXS, Jin *et al.* suggested that the conformation of i-motif DNA at mild acidic conditions is similar to that of the partially unfolded i-motif DNA rather than the fully folded i-motif [58]. Hence, the i-motif is structurally dynamic over a wide pH range, adopting multiple conformations ranging from the folded i-motif structure to a random coil conformation. In a later work, SAXS has been also used to study the effect of C60 fullerene attached to a C-rich sequence into the duplex – intramolecular structures equilibrium [59]. Representative tridimensional models may be postulated based on the analysis of the experimental SAXS data. Transient IR absorption (TRIR) study has provided for the first time the identification of signature IR bands for the long-lived species found upon UV-excitation of the i-motif in C-rich DNA [60]. The TRIR spectrum for UV-excited i-motif possess complex dynamics pointing to multiple decay processes, including possible charge transfer between packed hemi-protonated C bases. The slow decay is not due to simple protonation, but is rationalised in

terms of the specific structural features of the i-motif. The most likely origin is charge transfer between closely packed C bases.

The use of <u>Mass Spectrometry</u> (MS) to the study of i-motif-forming structures has also been described. In the earlier work, Rosu *et al.* revealed the existence of i-motif structures in the ions produced by electrospray MS [61] by using Infrared Multiphoton Dissociation (IRMPD) and Ion Mobility (IM). The i-motif structure was detected because its characteristic broadening of the IR band at 1650 cm^{-1} due to the contribution of a large blue-shift of the NH_2 scissoring and NH bending modes of the Cs. For the intramolecular human telomeric C-rich strand, ion mobility revealed that the lower charge states are more compact and give IR spectra characteristic of intramolecular hydrogen bond preservation in the gas phase. It has been described the formation of $C·C^+$ base pairs in gas phase from C derivatives [62]. The hemiprotonated dimers were observed in chloroform solution upon treatment with strong acid and in the gas phase by ESI-MS.

Other techniques include <u>sedimentation analysis</u> [63]. It has been described that the C_8 sequence adopts an A-like DNA structure at pH 7 and room temperature, whereas longer sequences containing 24 or 28 cytidine nucleotides assembled into an i-motif below ~15–20°C. The fourfold reduction in axial ratio is consistent with the formation of a four-stranded i-motif configuration. The results obtained in this work suggested that pure cytidine-containing oligonucleotides can form monomeric i-motif structures without heterologous nucleotides in the loops, such as thymidine or adenosine. The use of <u>analytical centrifugation</u> has been reported recently [64]. As pH increases, the hydrodynamic radius of individual DNA chains in aqueous solutions prepared by being heat-treated suddenly increases while the molar mass is constant, indicating that the conformation changes from an i-motif to a random coil. Finally, <u>competition dialysis</u> has been shown to be a useful technique to screen the selectivity of ligand-DNA interactions based on the differences in the secondary structure, such as i-motif [65].

MULTIVARIATE ANALYSIS

Modern instruments and computers allow the simultaneous recording of multivariate data, i.e., a complete set of spectra in function of an external variable (such as temperature or pH), and its appropriate storage. Multivariate analysis methods allow the recovery of information by unravelling these, *a priori*, complex data sets. In the case of melting experiments these methods not only allow the calculation of T_m or thermodynamic parameters, but also check the validity of the two-state process usually applied in the univariate analysis of these data. Multivariate analysis may be conducted in two different ways, depending on whether a physico-chemical model is initially proposed (hard-modelling approach) or not (soft-modelling approach).

For <u>hard-modelling</u> approaches, the proposed model depends on the nature of the process under study. Hence, for acid-base experiments the model will include a set of chemical equations describing the formation of the different acid-base species from the neutral species, together with approximate values for the stability. Kudrev *et al.* have studied the pH-induced formation of i-motif structures in several sequences based on a C-rich sequence found at the promoter region of the *c-kit* gene [66]. The spectra recorded along spectroscopically-monitored acid-base titrations were analysed by means of the hard-modelling-based <u>matrix method</u>. This enables the characterization of the protonation process in polymers in terms of the intrinsic protonation constant (K_{in}), and two additional parameters that model the cooperativity (ω_c) or anticooperativity (ω_a) of the process. Whereas the protonation and folding of mutated sequences containing only thymidines at the loops were successfully modelled, protonation and folding of the wild sequence, showing a greater base variability at the loops, could not be modelled. This fact was related to a larger

conformational variability in the wild sequence that in those mutated ones. Recently, a hard-modelling method has been applied to study the formation of i-motif structures in a sequence of the *n-myc* gene [67](Figure 4). In this case, CD and molecular absorption data were analysed simultaneously. For <u>melting experiments</u>, the physico-chemical model is related to the thermodynamics of DNA unfolding, according to the van't Hoff equation [68,69,70].

Figure 4. Resolution of the species present along the acid-base titration of a 34-mer sequence corresponding to the *n-myc* promoter [67]. (a) Selected set of CD experimental spectra. Inset: pH values at which the spectra were measured. (b) Selected set of molecular absorption spectra. (c) Calculated distribution diagram. (d) Calculated CD spectra. (e) Calculated molecular absorption spectra. Blue line: neutral form, probably a partially stacked single strand; red line: *i*-motif 1; green line: *i*-motif 2; black line: protonated form, probably a random coil. C_{nmyc01} = 1.3 µM, T = 25 °C. Reprinted with permission from reference [67].

Among the soft-modelling methods, <u>Singular Value Decomposition</u> has been traditionally used to determine the number of transitions during a process, such as a melting process [71]. Analysis of UV absorption spectra recorded along the thermally-induced unfolding of DNA structures by means of SVD allowed the detection of intermediates and a roughly estimation of the concentration dependence for each conformation with temperature. An evolution of SVD is the <u>Multivariate Curve Resolution – Alternating Least Squares</u> (MCR-ALS) method, which has been used to study the acid-base equilibria of a variety of DNAs, including C-rich sequences [72]. As SVD, MCR-ALS is a soft-modelling method that does not use any physico-chemical model in the calculation. The main advantages of MCR-ALS over previous methods are the inclusion of complementary information (such as non-negativity) as a mathematical constraint, and the possibility of carrying out a simultaneous analysis of data recorded along complementary experiments. Recently, this approach has been used to study the acid-base behaviour of oligonucleotides $(CAC)_3$, A_3C_6, and $C_3A_3C_3$ [73].

SIMULATION-BASED METHODS

There have been many applications of computational methods to study the formation and stability of i-motif structures. In particular, the use of molecular dynamics (MD) to simulate the solution behaviour of i-motifs has also been described. In an early work, Gallego *et al.* used the AMBER force field to simulate molecular dynamics trajectories of i-motif and duplex structures of centromeric DNA [74]. The electrostatic energy components were calculated using a Poisson-Boltzmann model, and the nonpolar energy components were computed with a van der Waals function and/or a term dependent on the solvent-accessible surface area of the molecules. It was found that the i-motif was mainly stabilized by favourable electrostatic interactions between hydrogen bonded protonated and neutral Cs, and by non-polar forces including the hydrophobic effect and enhanced van der Waals contacts, Later, Malliavin *et al.* studied the role of the phosphodiester backbones on the stability of tetrameric i-motifs formed by C_2 [75]. The 4 ns – long simulations were done for two different topologies (5'E and 3'E) and by using the program AMBER 6.0. Analysis of the trajectories showed that the main energetic difference between the two topologies is due to the van der Waals term. The comparison of the solvent-accessible area of each topology revealed that the sugar-sugar interactions accounted for the greater stability of the 3'E topology. This stresses the importance of the sugar-sugar contacts across the narrow groove which, enforcing the optimal backbone twisting, are essential to the base stacking and the i-motif stability. The hairpin structures formed by C-rich sequences at neutral pH values in explicit solvent have been also studied [76]. The GROMACS package and the modified AMBER force field were used. The results indicated that hairpin structures were the most stable configurations in the absence of protonated Cs. Recently, the structure and mechanical properties of ***i-motif*** nanowires based on a repeated TC_5 sequence were studied by means of MD simulations [77]. The i-motif nanowire seems to share similarities

with structural proteins, as far as its tensile stiffness, but is closer to nucleic acids and flexible proteins, as far as its bending rigidity is concerned. Furthermore, thanks to its very thin cross section, the apparent tensile toughness is close to that of a metal.

The interaction of porphyrin ligands with the i-motif structure formed near the *c-myc* NHE-III$_1$ (Nuclease Hypersensitivity Element III$_1$) has been studied *in silico* [78]. The NHE modelled consisted of single folded conformers of the polypurine intramolecular G-quadruplex and the polypyrimidine intramolecular i-motif structures, flanked by short duplex DNA sequences. The modelled i-motif structure used in that work was theoretical; being the central intercalated C core interactions based on NMR structural data obtained for a tetramolecular $(A_2C_4)_4$ i-motif. The loop structures were also *in silico* predictions of the *c-myc* i-motif loops. It was found that the cationic porphyrin compounds TMPyP2, TMPyP3, and TMPyP4 were all bound to both ends of the i-motif, one immediately adjacent to the thymidine triplet, near the interface between the G-quadruplex and i-motif, and the other at the opposite end of the i-motif. The analysis of the interaction energy in binding to both ends of the i-motif revealed that it was dominated by Van der Waal's energy, with very little electrostatic component. This was explained as due to the aromatic and hydrophobic nature of the porphyrin compounds. The planar, aromatic porphyrin ring structure tends to form stacking interactions with multiple pairs of bases of the DNA, and the pyridinium rings can interact favourably with other bases or to the deoxyribose rings of the DNA.

STABILITY OF THE I-MOTIF STRUCTURE

The relative stability of nucleic acid folded structures is often measured in terms of their melting temperature (T_m). T_m is defined as temperature of mid transition from the folded structure to the unfolded species often induced by heating of the nucleic acid sample. The transition is usually monitored using a spectroscopic technique, such as a molecular absorption or circular dichroism. As other nucleic acid structures, the stability of the i-motif depends on several important factors, such as the nucleotide sequence, ionic strength or temperature, among many others. Because the protonation of one of the C bases in the $C \cdot C^+$ base pairs is an absolute requirement for the formation of the i-motif structure, pH of the medium plays a crucial role. As the pK_a value of C is around 4.6 (in pure water at 25°C), it would be expected the formation of i-motif structures at pH values lower than 6.6, approximately [34,52]. Hence, at pH values from ~4 to ~7 and 25°C, the C bases are partially protonated and the DNA folds into the closed i-motif structure. In this pH range, the stability of the i-motif is a linear function of pH [34,79]. The highest stability of i-motif structures occurs at pH values close to the pK_a of C. At higher pH, the C bases deprotonate and the structure unfolds to a single-stranded form. On the other hand, if the pH value is too low (below 3, approximately) all the C bases are protonated and they cannot form the hydrogen bond pattern needed for the $C \cdot C^+$ base pair [79]. This is not the case of $G \cdot C$ and $A \cdot T$ Watson-Crick base pairs, whose stability is pH-independent in a broader pH range (from 2 to 10, approximately). At 37°C and 150mM NaCl, the i-motif is not formed at physiological pH (~7.4) but, depending on the sequence, may be formed at ischemic pH (~6.7) [80]. Nevertheless, some i-motifs have been observed at low temperature and at neutral [28] or even slightly basic pH values [81]. Recently, folding of i-motif structures in presence of silver cations at physiological pH has been reported [82].

The formation of the i-motif is strongly enthalpy driven at near neutral pH, being the value maximum at pH 6-7, and drops below pH 5.5 [34]. As example, using as a model a 31-mer sequence belonging to the *c-myc* gene, Mathur *et al.* studied the thermodynamic parameters associated to the formation of the i-motif [53]. At 20 °C, it was observed a ΔG of -10.4 ± 0.1 kcal·mol^{-1} with favourable enthalpy ($\Delta H = -76.0 \pm 1.0$ kcal·mol^{-1}) and

unfavourable entropy (TΔS = −65.6 ± 0.9 kcal·mol⁻¹) at pH 5.3 in 20 mM NaCl for i-motif folding. It is possible to determine the contribution of a C·C⁺ base pair to the thermodynamic parameters. Using a set of intramolecular i-motifs differing in the number and length of C tracts, Mergny *et al.* determined that each base pair decreases the ΔH by 10.5±0.5 kcal·mol⁻¹ and ΔS by 57±1 cal·mol⁻¹·K⁻¹ at pH 6 and 100mM NaCl [34]. Using sequences that form both intra- and intermolecular i-motif structures, values for ΔH and ΔS around 7 kcal·mol⁻¹ and 20 cal·mol⁻¹·K⁻¹, respectively, at pH 5.8 and 100 M Na⁺ were determined [52], with these values being larger than those corresponding for the disruption of base pairs in Watson-Crick duplex (~4.5 kcal·mol⁻¹ and ~11 cal·mol⁻¹·K⁻¹, respectively). From the NMR titration of TC₃, a free energy of -7.6 kJ·mol⁻¹ per cytidine base pair for the formation of the tetramer from single strands was calculated [24]. Overall, the values determined for the ΔH and ΔS have been shown to be independent on the nature (monomeric o tetrameric) of the i-motif structure. Finally, in the case of the *c-myc* i-motif, the formation of each C·C⁺ base pair leads to a decrease in enthalpy of 9-12 kcal·mol⁻¹ [53].

Apart from pH, the fundamental factor affecting the structure and stability of the i-motif structure is the <u>number of C·C⁺ base pairs</u> present in the folded structure. For a given pH value, the stability of two different i-motifs differing in the number of C bases in their sequences would be higher for the one showing a higher number of C·C⁺ base pairs [83]. It has been shown that the intra- versus intermolecular folding primarily may also depend on the length of the C-tracts [84]. Two stretches of six or fewer C residues favour the intermolecular folding of i-motifs, whereas longer C-tracts promote the formation of intramolecular i-motif structures with high thermal stability.

<u>Bases other than C</u> may be also accommodated into the i-motif core formed by C· C⁺ base pairs. Thymidines can form symmetrical T·T base pairs that are nearly isomorphic of C· C⁺ base pairs. The NMR-based study of the 9-mer sequence 5mC₂TCTCTC₂ showed the existence of two open/closed T3/T7 motifs, but the central thymidines form two long-lived T5·T5 pairs that are intercalated into the i-motif core [85]. Interestingly, whereas that sequence forms a tetrameric i-motif, the similar sequence 5mCCTCACTCC forms a dimer [27]. In this case, the dimeric i-motif is built by intercalation of two symmetrical hairpins held together by six symmetrical C·C⁺ pairs and an additional T·T pair. In summary, the formation of T·T base pairs in general stabilizes the i-motif structure because of extension of the i-motif core and decrease in flexibility [86]. Recently, the influence of methylation and halogenation of cytosine on the base-pairing energies (BPEs) in the *i*-motif has been studied [15]. To address this, proton-bound heterodimers of cytosine and 5-methylcytosine, 5-fluorocytosine, 5-bromocytosine, and 5-iodocytosine were studied. All modifications were found to lower the BPE and therefore would tend to destabilize DNA *i*-motif conformations. However, the BPEs in these proton-bound heterodimers still significantly exceed those of the Watson–Crick G·C and neutral C·C base pairs, suggesting that C·C⁺ mismatches are still energetically favoured such that *i*-motif conformations are preserved.

Apart from the number of C bases, another key factor is the <u>length and</u> <u>nature of loops</u>. First, the length of the loops has been shown to be critical in determining the multimeric nature of the i-motif structure. Hence, C-rich sequences with just one base at the loops tend to form mixtures of mono and bimolecular i-motifs [84], whereas longer loops favour the formation of intramolecular i-motifs [34]. It has been proposed that i-motif structures may be <u>classified into two groups</u> depending on the length of the loops [87]. In Class I, the loop sizes are 2:3/4:2 with four, five or six C·C⁺ base pairs. In Class II, the loop sizes are 6/8:2/5:6/7. In general, the midpoint of the pH-induced folding from the neutral strand for Class II-i-motif structures is slightly higher (~ 6.6) than that for i-motif structures identified as Class 1 (~5.8-6.4). due to the presence of stabilizing effects into longer loops. In this sense, it has been described an i-motif showing a 12-bases long loop that is

able to form a hairpin stabilized by Watson-Crick base pairs [67]. Interestingly, the potential to form a stem-loop structure in the long loop directs a particular and well stabilized DNA structure from diverse choices of high-ordered arrangements when bases at this loop were mutated to T.

The <u>nature of the bases at the loops</u> also influences strongly the structure and solution behavior of i-motif structures. Using NMR methods, it has been shown that protonation of A bases at the loops when the pH is lower than 4.6 produces a conformational change involving disruption of the i-motif core [88]. The studied sequence, a mutant of a fragment of the human centromeric satellite III, $C_2AT_2C_2AT_2C_2T_3C_2$, folds into a intramolecular i-motif structure. Protonation of adenine bases hinders the formation of an A·T base pair that extends the C·C$^+$ core at pH higher than 4.6. The application of a multivariate approach to the study of the acid-base equilibria of C_3T_3 and C_3TA_2 sequences allowed the quantitation of the relative contribution of all acid-base species in function of pH [69]. It was observed that the adenine-containing sequence formed two i-motif structures, probably differing in the protonation state of adenines at the loops. Moreover, it has been observed that adenine-containing i-motif structures unfold at lower temperatures than non-mutated sequences [34,69]. T·T base pairs can occur in the loops connecting the C-tracts [89]. Recently, using the telomeric i-motif sequence as model, it has been proposed that sterically demanding A bases at the double-loop side lead preferentially to the 5'E conformation [86]. In this sense, the systematic modification of the 4A loop region of the retinoblastoma gene with PyA fluorophore units allows discrimination of the fluorescence signals corresponding to structural dynamics from single-stranded to i-motif structures [90]. In terms of fluorescence signals, i-motif structures modified with fluorophores at the 1,2 and 1,4 positions of the 4A loop provide the most dramatic fluorescence changes at a single excitation wavelength upon conformational transitions from single-stranded to duplex to i-motif structures, respectively. Loop interactions are of particular relevance in the dimeric structure of d($TC_3GT_3C_2A$) [29] and d($TCGT_3CGT_2$) [28]. In both cases, the 5´-GT$_3$-3' loops form intermolecular GT base pairs that interact to each other through their minor groove side, forming a G:T:G:T tetrad.

The <u>bases at the 5' and 3' ends</u> of C-rich sequences able to form i-motif structures have also influence on the stability and molecularity of the i-motif structures. The study of the $C_3TA_2C_3$ and $C_3TA_2C_3TA_2$ sequences by means of PAGE and thermal melting experiments allowed the proposal that the inclusion of the additional –TAA segment at the 3'- end produces a change in the molecularity of the structure. Hence, the 9-mer folds in a bimolecular structure, whereas the 12-mer sequence exists in two (bimolecular and tetramolecular) forms [91]. The different arrangement of flanking bases may also influence the stability of the i-motif structure. Using NMR methods, it was shown that the A_2C_3 and C_3A_2 folded into different i-motif structures [92]. Whereas the first sequence adopts a unique structure showing A·A pairs stacked to C·C$^+$ base pairs, the second sequence, which adopts two distinct intercalation topologies (3'E and 5'E), fairly shows well stacked adenosine bases to the adjacent C·C$^+$ base pairs. Using X-ray crystallography, it was shown that the i-motif formed by the AC3T sequence was further stabilized at one end by a three-base hydrogen-bonding network, in which two As and a T form four hydrogen bonds via a reverse Hoogsteen and an asymmetric A–A base pairing [93].

As many other nucleic acid structures, the <u>ionic strength</u> of the solution has a certain influence on the stability of the i-motif structure. As described above, nitrogenous bases are neutral throughout a wide range of pH, from 5 to 9, approximately. The same occurs with the pentose sugar, which can only lose the proton at very alkaline pH values. However, at neutral pH values, DNA is a polyanion because of the negatively charged phosphate groups. The presence of appropriate counterions, such as sodium, potassium, and others balances the negative charge. Folded conformations may produce changes in the counterion atmosphere surrounding the phosphate backbone. Moreover, in the case of i-motif, protonation at N3 produces the appearance of

positive charge at the base pairs. Therefore, variations in ionic strength can cause conformational changes and variations in the relative stability of i-motif structures. In general, it has been observed than an increase in ionic strength from 0 to 100mM NaCl produces a destabilization of the structure [34]. Higher NaCl concentrations did not produce changes in T_m values. Using a C-rich sequence found near the promoter region of the *c-myc* gene, it was found that folding was induced by uptake of about two to three protons per mole of i-motif while a marginal (0.5–1 mol/mol) counterion uptake was observed [53]. Using CD spectroscopy, S1 nuclease digestion, and multivariate analysis it was shown later that two different i-motif conformations may be formed differing in the number of C·C+ base pairs[94]. The transition between both conformations is induced by a change in pH or ionic strength.

Apart from the canonical i-motif structure formed at acid pH values (~4-6), it has been proposed the formation of underlined partially folded structures at neutral and slightly acid pH values (~7-6). Using FRET and FCS techniques in the bulk phase and at the single-molecule level, it has been proposed that the partially folded species coexist with the single-stranded structure at neutral pH and room temperature [95].

INTERACTION WITH LIGANDS

The stability of folded structures of DNA may depend strongly on the interaction with inorganic and organic ligands. Hence, research is being done to find ligands which could modulate *in vivo* the stability of characteristic DNA structures, such as those G-quadruplex proposed near the promoter region of several oncogenes [96]. In contrast with G-quadruplex, few molecules have been reported to bind to i-motif structures (Figure 5). The reasons may be the low stability of i-motif at physiological conditions and the fact that the i-motif is a very compact structure, where planar ligands which usually stack on base pairs cannot be introduced easily. Therefore, non-specific electrostatic interactions throughout the phosphate groups or interactions with bases at the loops seem to be the most plausible interaction mechanisms.

Figure 5. Several ligands which have been reported to bind i-motif structures. (a) single-walled carbon nanotubes (SWNT), (b) TMPyP4, (c) Ru(bpy)$_2$(dppz)]$^{2+}$, (d) graphene quantum dots (GQDs) with peripheral carboxylic acid groups, (e) PMNT, (f) IMC-48.

The interaction of single-walled carbon nanotubes (SWNTs) with i-motif-forming sequences was firstly studied by molecular fluorescence and S1 nuclease digestion. The cleavage patterns showed that SWNTs bind to i-motif DNA at the end of the major groove [97]. It was proposed that SWNTs can selectively stabilize human telomeric C-rich DNA and induce i-motif DNA formation under physiological conditions or even at pH 8.0. The strong affinity of i-motif DNA for SWNTs has been also used for distinguish single- and multiwalled carbon nanotubes [98]. Later, the interaction of SWNTs with human telomeric i-motif DNA was shown to accelerate the S1 nuclease cleavage rate at the loop regions [99]. In a recent work, the use of SWNTs to inhibit telomerase activity through stabilization of i-motif structure has been reported [100]. The persistence of i-motif and the concomitant G-quadruplex eventually leads to telomere uncapping and displaces telomere-binding proteins from telomere.

TMPyP4 is a porphyrin that has been extensively used as a model ligand to study the binding characteristics of G-quadruplex structures. Hence, the study of its interaction with i-motif is the natural continuation of such studies. Moreover, the interaction of TMPyP4 with a C-rich sequence corresponding to the promoter region

of the *c-myc* gene at pH 7.4 has been shown to have an inhibitory effect on NM23-H2 DNA-binding activity [101]. From a biophysical point of view, Fedoroff *et al.* studied the interaction of TMPyP4 with a tetrameric i-motif formed by the A_2C_4 sequence [102]. From the results of NMR and docking studies, the ligand was suggested to bind peripherally near the ends by a non-intercalative mechanism that is independent on the multimeric nature of the i-motif structure. Using a intramolecular i-motif based on the telomeric sequence, it was shown that the binding of the porphyrin at pH 5 does not alter the structure, being the interaction mostly electrostatics in nature [69]. It was proposed that each i-motif binds up to two TMPyP4 molecules in an independent way, being the value of the binding constant for each binding site $\sim10^6$ M^{-1}.

The interaction of the <u>ruthenium complex</u> Ru(bpy)$_2$(dppz)]$^{2+}$ (where dppz stands for dipyrido[3,2-a:20,30-c] phenazine) with C-rich sequences has been studied at pH 7 and pH 5, where the i-motif is already formed [103]. Only a weak interaction was observed at pH 5, probably of electrostatic nature. Other complexes involving central metal atoms, such as the terbium-amino acid complex [Tb$_2$(DL-HVal)$_4$-(H$_2$O)$_8$]Cl$_6$ have been studied. In this case, the interaction at pH 5.5 with the telomeric i-motif structure produces a slight destabilization of the structure [104].

<u>Graphene quantum dot</u> (GQDs) have been described as stabilizing ligands for the human telomeric sequence $(C_3TA_2)_3C_3T$ [105]. GQDs are graphene sheets that possess novel chemical and physical properties. They can stabilize and induce the i-motif formation, as denoted by an increase in T$_m$ values upon addition of the ligand. It seems that GQDs not only stabilize the i-motif, but they can also induce the i-motif formation in alkaline or neutral pH. The mechanism of interaction has been explained in terms of an interaction with the internal TAA loop.

In a recent work, the interaction of <u>phenanthroline derivarives</u> with the i-motif structure formed by the $C_3(A_2TC_3)_3$ sequence has been studied [106]. These derivatives stabilize the structure of i-motif with an increase in melting temperature of ~8ºC in the presence of 10 times excess of these compounds. Their binding stoichiometric ratio and constant were 1:1 and 2·10^5 M^{-1}, respectively. Xu *et al.* studied the interaction of the hairpin structures formed by the sequences CGC$_4$TA$_2$C$_3$TA$_2$CTA$_2$C$_3$TGCG and A$_2$(C$_3$TA$_2$)$_2$C$_3$T$_2$C$_3$T$_4$ at pH 7 with the ligands doxorubicin and Hoechst 33258. These sequences may be used as a pH-driven C-rich DNA drug release device because, at neutral pH, may bind the ligands whereas no interaction was observed at pH 5.0, at which the i-motif structure is formed [107]. Finally, the interaction of a known dye, crystal violet, with the tetrameric i-motif structure 5'-(AC$_3$T)$_4$-3' has been studied recently [108]. The ligand was shown to bind externally to the terminus of the i-motif structure with a 1:1 stoichiometry and a 7·10^5 M^{-1} binding constant.

Apart from "canonical" aromatic and planar ligands, the interaction of i-motifs with other substances has been studied. Hence, Ren *et al.* reported the use of PMNT, a polythiophene derivative, to visualize by naked eyes the formation of i-motif structure upon pH changes [109]. In response to the conformational switch of i-motif, the conformation of PMNT that forms the complex changes correspondingly, resulting in a different colour. In a very recent work, Hurley *and col.* described the specific interaction of the IMC-48 ligand with an i-motif-forming sequence found at the *bcl-2* gene. At pH 6.6 the i-motif is in equilibrium with a hairpin showing Watson-Crick base pairs. By choosing the appropriate ligand, this equilibrium may be efficiently shifted, with dramatic effects on the gene expression [110].

The interaction of the linear carotenoid ligands crocin and crocetin, as well as the monoterpene aldehydes picrocrocin and safranal, with the i-motif structure formed by GC(TC$_2$)$_3$TC$_2$T(TC$_3$)$_3$ has been also studied [111]. Finally, the affinity of metal cations such as silver or palladium ions for i-motifs are being exploited for the preparation of DNA-mediated metal nanoclusters that may be used as fluorescent probes of gene sensors [112].

A review of studies dealing with the development and applications of these nanoclusters is beyond the scope of this work.

EFFECTS OF CHEMICAL MODIFICATIONS ON THE STABILITY

The stability of the i-motif structure may be modulated by introducing appropriate modifications in one or more of its building blocks. Hence, <u>modification of nucleosides</u> has been proposed as a way to increase the stability of i-motif structures. As example, the effect of the incorporation of 5-methyl-cytosine, deoxyuracil or 5-(1-propynyl)-deoxyuracil into C-rich DNA sequences was studied in detail in one of the first works dealing with this subject [113]. It was observed that oligonucleotides including 5-propynyl- deoxyuracil formed a stable i-motif which precluded triplex formation. Later, it was studied the influence of methylation or other modifications such as halogenation of cytosine on the base-pairing energies (BPEs) in the i-motif [15]. Proton-bound heterodimers of cytosine and 5-methylcytosine, 5-fluorocytosine, 5-bromocytosine, and 5-iodocytosine tend to destabilize DNA i-motif conformations. However, the BPEs in these proton-bound heterodimers still significantly exceed those of the Watson–Crick G·C and neutral C·C base pairs, suggesting that C·C+ mismatches are still energetically favoured such that i-motif conformations are preserved.

Inclusion of a single substitution of 2´-deoxy-2´-fluororibocytidine into the TC_5 sequence enhances the stability of the i-motif in sodium citrate buffer (50 mM, pH 4.2) [114]. Although multiple substitutions do not increase the stability further, they are relatively well tolerated. This is in contrast with the effect of inclusion of ribonucleotides in the same sequence, which is strongly destabilizing. On the other hand, inclusion of arabinonucleotides facilitates an increase in stability [115]. The different effect provoked by the relative orientation of the –OH group in riboses and arabinoses supports the importance of sugar–sugar contacts in stabilizing i-motifs. In the case of 2´-deoxy-2´-fluoroarabinocytidine, alteration of sugar-sugar contacts give rise to multiple effects, including the formation of alternative i-motifs.

Figure 6. Chemical modifications in the phosphate backbone. (a) phosphoramidate, (b) Sp and Rp phosphorothioate, (c) methylphosphonate, (d) phosphorothiolate.

The i-motif is also sensitive to alterations of the phosphate component of the <u>backbone</u> (Figure 6). Some modifications, such as phosphoramidate oligonucleotides have been shown to prevent the folding of C-rich sequences into i-motif structures at acidic pH values but not into triple helices [113]. Similarly, sequences showing methylphosphonate backbones do not form i-motif structures [113,116]. On the contrary, phosphorothioate oligodeoxynucleotides have been shown to form i-motif structures [113,116]. It was observed that the stability of phosphorothioate oligodeoxynucleotides (measured in terms of T_m values) was slightly lower than for phosphodiester backbone. Moreover, at neutral pH, the phosphorothioate associates and dissociates nine times faster than the phosphodiester oligodeoxynucleotide of identical sequence. Kanehara et al. also studied the stabilizing role of the R_p and S_p configurations of the phosphorothioate backbone in a tetrameric i-motif [117]. As for duplex DNA, it was observed that the S_p configuration produced a more stable structure than R_p. Later, it was suggested that the S_p configuration at phosphorus of the phosphorothioate linkage changes the sugar–phosphorothioate conformation and intermolecular interaction in the narrow groove, leading to the destabilization of the i-motif structure [118]. The comparison of the NOESY spectra showed that intraresidual H6−H3' and H2''−H4' NOE cross-peaks of the *all* S_p isomer are weaker than those

of the *all R*$_p$ isomer and PO-d(TC$_4$), indicating the change in the C3'-endo conformation and glycosidic bond angle. Another modification of the phosphate backbone involves the inclusion of 3'-S-phosphorothiolate linkages [119,120]. It has been shown that this incorporation stabilizes the i-motif with a minimum perturbation of the overall structure. This stabilization probably arises from the preference of the phosphorothiolate residues for the C3'-*endo* sugar pucker, which is universally observed in solution for the cytidine sugars in the i-motif. These results also strongly suggest that the destabilization of the i-motif by ribose sugars is due to the 2'-substituent and not the C3'-*endo* sugar pucker.

The formation of i-motif structures in DNA-modified backbones, such as Locked Nucleic Acids (LNA) (Figure 7) has also been studied. This is a conformationally constrained nucleic acid mimic with a 2'-O–4'-C methylene bridge that locks the nucleotide analogue in a C3'-endo sugar conformation. Generally, the introduction of LNA nucleotide monomers increases the thermal stability of DNA and RNA duplexes and triplexes. However, the *in vivo* potential use of LNA derivatives is still under study because they have been shown to produce hepatotoxicity in animals [121]. The i-motif formation with LNA-modified TC$_5$ oligonucleotides at low pH values was initially identified by means of CD, UV, and NMR spectroscopy [122]. Later, the effect of LNA nucleotides on the formation in modulating the formation and stability of the i-motif structure formed by a *c-myc*-based sequence was also studied [123]. Choosing the right mutations, it is possible to increase or decrease the stability, measured in terms of T$_m$ values, of a wild sequence.

On the other hand, Pasternak and Wengel studied the influence of Unlocked Nucleic Acid (UNA) monomers on thermodynamic stability of the 22 nucleotides human telomeric *i*-motif [124]. It was shown that UNA monomers modulate *i*-motif stability in a position-depending manner whereas no structural changes were observed in any case. In addition, 5-(pyren-1-yl)uracilyl UNA monomers were found to destabilize an i-motif structure at pH 5.2, both under molecular crowding and non-crowding conditions [125].

The formation of i-motif structures from Peptide Nucleic Acids (PNA) has also been described. One of the first works used alanyl-PNA, which consists of a regular peptide backbone with an alternating configuration of the alanyl amino acids. This backbone has nucleobases connected in the β position of the side chain [126]. It was found that the octamer -(AlaC-AlaC)$_4$-Lys-NH$_2$ did not form any i-motif structure at pH 4.5. On the contrary, the H-(Gly-AlaC)$_4$-Lys-NH$_2$ sequence formed the i-motif. It was concluded that the formation of i-motif structures in PNA sequences required the inclusion of a glycine unit as every second amino acid thus providing a spacer that facilitates the intercalation of linear double strands. Using the DNA and PNA analogues d(TC$_5$) and p(TC$_5$), it has been shown that the PNA i-motif is stable over a pH range (∼4.1 – 4.5) narrower than in the case of the DNA i-motif (∼4.7-5.7). The difference was explained in terms of a response to pH to the polyanionic nature of the DNA backbone which probably shields pH changes more effectively than the neutral PNA backbone [127]. This narrower pH range of existence was also confirmed for i-motif formed by p(TC$_8$) being its stability, however, slightly higher than that of the corresponding DNA i-motif [128]. The use of a peptide template has been described as a method for the design of i-motif structures that may be stable even at neutral pH values [129]. In this approach, a cyclic peptide scaffold is used as a topological template for directing the intramolecular assembly of anchored oligonucleotides into an i-motif topology.

Figure 7. Other backbone modifications: (a) locked nucleic acid (LNA); (b) unlocked nucleic acid (UNA); (c) peptide nucleic acid (PNA).

A mixed strategy involves the formation of <u>hybrid i-motif structures</u> from a binary mixture of C-rich DNA (TC_5) and PNA (H_2N-TC_5-Lys-COOH) sequences [130]. ESI-MS confirmed the formation of a tetrameric species, composed of PNA–DNA heteroduplexes whereas NMR spectroscopy confirmed that PNA and DNA form a unique complex comprising five C·C^+ base pairs per heteroduplex. H1′–H1′ NOEs show that both heteroduplexes are fully intercalated and that both DNA strands are disposed towards a narrow groove, invoking sugar–sugar interactions as seen in DNA i-motifs. As in the case of PNA i-motif, this hybrid i-motif showed enhanced thermal stability and intermediate pH dependence. This is likely to be because two of negatively charged backbones in the DNA i-motif are replaced by neutral polyamide PNA backbones thereby reducing the electrostatic repulsion associated with multi-stranded structures. Furthermore, the DNA i-motif is net negatively charged, the PNA i-motif is net positively charged and the hybrid is an electrically neutral complex.

<u>Circular oligonucleotides</u> possess many distinctive properties, when compared to their linear counterparts, such as higher DNA-binding affinity, greater sequence selectivity, enhanced resistance to degradation by exonuclease, and an ability to serve as efficient templates for DNA and RNA polymerase [131]. In this sense, it has been shown that the i-motif can direct the sequence-specific formation of a phosphodiester linkage and thus representing a new type of structural template for constructing circular oligonucleotides [131,132]. A particular example are the nanometer-sized circles range in length from 36 to 60 nucleotides based on the C-rich human telomere repeat, $(C_3TA_2)_n$ [133]. These cyclic DNAs may act as templates for synthesis of human telomere repeats *in vitro*. The circles were constructed successfully by the application of the A-protection strategy, which allows for cyclization/ligation with T4 DNA ligase. Thermal denaturation studies showed that at pH 5.0, all circles form folded structures with similar stability, while at pH 7.0 no melting transitions were seen. Recently, it has been reported a minimal i-motif structure formed by the cyclic sequence <$pTCGT_2CGT_2$>. This sequence form a dimeric structure stabilized by a unique C·C^+ base pair capped at both ends by G:T:G:T tetrads [28]. Finally, Zhou *et al.* described the synthesis of a fluorescein-labelled circular i-motif structure. The most striking feature of their work is that the dye is not bonded to the usual 5' or 3' ends, but to the central loop. This is accomplished by covalent bonding to one of the bases at the loops not involved in the i-motif core [134]. It was shown that this cyclic oligonucleotide resists the hydrolysis by exonucleases and the fluorescent moiety is indeed present in the ligation product.

Yang *et al.* studied the effect of tetra(ethylene glycol) (EG_4) substitution at the <u>loop</u> region of intramolecular i-motif formed by $(C_3T_2)_3C_3$. In general, single substitution at all loops preserved the stability of *i*-motif, while the triple substitution of three bases at one loop resulted in a significant decrease of stability, and the more EG_4 substitutions, the less stable of the system. The substitution at the narrow groove results in a faster migration in agarose gel comparing to those at other two wide grooves, suggesting modification at different positions has different extents of influence on the topology [135]. The influence of the insertion of a non-nucleotide pyrene moiety into the loop between two C-rich regions has also been studied [136]. The stability of the i-motif structures was measured at different pH values under non-crowding and crowding conditions (20% poly(ethylene glycol)). When (R)-3-((4-(1-pyrenylethynyl)benzyl)oxy) propane-1,2-diol (TINA) was inserted, the oligonucleotides still formed i-motif structures of similar stability (at pH 6.2 but not at pH 5.2) than those observed for the corresponding wild type oligonucleotide. Interestingly, incorporation of pyrrolic-modified porphyrines into C-rich sequences induces i-motif stabilization [137]. This is probably due to porphyrine-porphyrine interaction between the parallel oriented strands in the tetrameric i-motif.

Other modifications include the addition of C60 fullerene to <u>both 5' and 3' ends</u> of a telomeric C-rich sequence [59]. Upon addition of the G-rich complementary strand, it was found that fullerene shifted the pH-

induced conformational transition between the i-motif and the duplex structure, possibly due to the hydrophobic interactions between the terminal fullerenes and between the terminal fullerenes and an internal TAA loop in the DNA strand. Finally, Robidoux *et al.* described the synthesis of <u>branched parallel</u> cytidine-rich oligonucleotides that are joined at the ends with a riboadenosine linker. It was shown that this molecular architecture enhances the stability of the resulting i-motif [138]. In addition, these branched constructions were used to explore the effect of arabino- vs ribo- substitutions in the stability of i-motif, showing that arabino-cytidines do not provoke the dramatic destabilizing effect of rC. The formation of branched i-motif structures incorporating 2-deoxy-5-propynylcytidine residues was also confirmed by temperature-dependent CD- and UV-spectra as well as by ion-exchange chromatography. The low pK_a-value of this nucleoside ($pK_a \sim 3.3$) compared to cytidine ($pK_a \sim 4.5$) required strong acidic conditions for i-motif formation. The immobilization of oligonucleotides incorporating multiple residues of that nucleoside on 15 nm gold nanoparticles generated DNA–gold nanoparticle conjugates which are able to aggregate into i-motif structures at pH 5 [139].

KINETICS OF I-MOTIF FORMATION

Appropriate sequence and environmental conditions (such as pH, temperature or ionic strength, among others) may favor the formation of thermodynamically stable i-motif structures. However, kinetic aspects such as time needed to accomplish the folding of a given DNA sequence or the folding pathway must be taken into account. As the most-studied i-motif-forming sequence, the C-rich strand of telomere DNA was recently reported to transiently exist as a 5' single-stranded overhang at the chromosome ends in the S-phase of replicating human cells [140]. The rates of folding and unfolding are crucial in determining the opportunity or possibility of the formation of i-motif under *in vivo* situations within the narrow "time window", during which the DNA is transiently liberated [141].

As expected, measurements done <u>at different pH values</u> show that both the folding- and unfolding-rate constants (k_f and k_u) are strongly dependent on pH. Zhao *et al.* studied the variation of the rate constants with pH for a intramolecular i-motif structure by SPR [141]. They observed that the folding and unfolding of $(C_3TA_2)_4$ occur on the timescale of minutes at pH ~5. These values, which were larger than those obtained in solution [132], were explained as due to the immobilization of the DNA on the SPR chip. The results showed that promotion of i-motif formation by protons is achieved by a combination of increased k_f and decreased k_u, which result in a rapid increase in the folding equilibrium constant, K_F, with decreasing pH. The impact of pH on the kinetic aspects of i-motif folding is also reflected on the observation of irreversible melting/annealing curves. FRET studies revealed high reversibility of the pH-induced folding of i-motifs and multiphasic folding kinetics with folding and unfolding time constants in the order of minutes [142].

Intramolecular i-motifs form faster than those multimeric in nature. Hence, the <u>formation times</u> for intramolecular and tetrameric structures are in the order of seconds and hours, respectively. Liu and Balasubramanian observed that the folding and unfolding processes of $(C_3TA_2)_3C_3$ i-motif are both completed in about 5s in a proton-fueled DNA nanomachine using the fluorescence spectroscopy [143]. However, the formation of dimeric $(mC_2TCACTC_2)_2$ and tetrameric $(TC3)_4$ i-motifs takes place in hours [27,85].

Most studies have been done on the <u>formation pathway</u> of i-motif structures. As example, the formation and dissociation rates of three $[TC_n]_4$ tetramers (n=3, 4 and 5), their dissociation constants and the reaction orders for tetramer formation by NMR have been reported [144]. The experimental observations suggested that i-motif formation proceeds via sequential strand association into duplex and triplex intermediate

species and that triplex formation is rate limiting. The proposed model is different to the one involving association of two preformed duplexes. In the case of the $[TC_3]_4$ i-motif structure, the reaction order was 3. Studying the folding kinetics of a dimeric i-motif, Canalia *et al*. found that the association rate of this structure is smaller than in the case of Watson-Crick duplex. As oligonucleotide association in Watson–Crick duplexes is described by a model involving the formation of a correct nucleus of a few base pairs, followed by rapid 'zipping' into the fully paired duplex, the formation of an i-motif structure could involve additional steps [27]. These could be: (i) the formation a hemiprotonated nucleus; (ii) adequate intercalation during the nucleus elongation of a third strand with appropriate orientation; and (iii) association, in parallel orientation with this stand, of a fourth strand that finally locks the strand assembly into a long-lived structure.

Lieblein *et al*, by means of NMR spectroscopy, investigated the kinetics of folding of the i-motif initiated by a pH-jump from pH 9 to pH 6. Under these conditions, folding follows a kinetic partitioning mechanism, where two conformations form in the first step with a rate constant of the order of 2 min^{-1}. Subsequent refolding of the kinetically favoured conformation to the thermodynamically more stable conformation is slow, with rate constants of the order of 10^{-3} min^{-1}. They proposed that the two conformations differ in the intercalation topology of the $C \cdot C^+$ base pairs and in the formation of $T \cdot T$ base pairs. At equilibrium, the closing $C \cdot C^+$ base pair can either be formed at the end of the C-rich strand (5'E) in the major conformation or at the 3'-end (3'E) in the minor conformation [21].

Using the stopped-flow circular dichroism (SFCD) technique, the kinetics of the pH-induced folding and unfolding process of the i-motif formed by $(C_3TA_2)_3C_3$ have been also studied [145]. The results showed that the molecule can fold or unfold on a time scale of ~100 ms when the solution pH is changed from 8 to 5, or vice versa. On the assumption of irreversible folding or unfolding processes, theoretical models to decipher the respective kinetics were proposed, suggesting that the cooperativity of protons is crucial for both the folding and unfolding process. In the unfolding process, the cooperative neutralization of two protons (out of the total six protons in the i-motif core) is the only rate-limiting step. In the folding process, on the contrary, there exists a critical step in which three protons bind cooperatively to the DNA strand.

The influence of pH on the dynamics of intramolecular folding of the $(C_3A_2T)_3C_3$ sequence has been studied recently by means of FRET and FCS techniques [95]. The measured fluorescence decay profiles were explained as the result of a mixture of two different species (the unfolded strand and a partially folded strand) that evolve simultaneously to yield the i-motif structure at pH 4.8. Interestingly, it is proposed that the partially folded species as well as the single-stranded structure coexist at neutral pH, supporting that the partially folded species may exist substantially *in vivo.* Finally, Zhou *et al.* claimed the formation of i-motif structure by the telomeric sequence at neutral and even slightly basic pH values, at 4oC [81]. The kinetic data provided by CD and FRET were fitted to a single exponential describing first-order kinetics. Surprisingly, the determined time constants depended strongly on the spectroscopic data used for the calculation (214 s for CD data, and 493 s for fluorescence data). The difference was explained in terms of a delaying effect of the dyes attached covalently into the sequence in FRET measurements.

I-MOTIF VERSUS INTERMOLECULAR WATSON-CRICK DUPLEX COMPETITION

In vivo, C-rich regions are usually accompanied by the complementary G-rich regions. It is expected, therefore, a thermodynamic equilibrium involving the Watson-Crick duplex structure and the potential

tetraplex (G-quadruplex and i-motif) structures. The hypothetical role *in vivo* of both tetraplex structures must be closely related to the shift of the equilibrium from the Watson-Crick duplex. This competition has been extensively studied, mainly in the case of the human telomeric DNA, but also in the case of guanine- and C-rich sequences found in the promoter regions of several genes.

Phan and Mergny published one of the first works studying the competition between the tetraplex (G-quadruplex and i-motif) and the Watson-Crick duplex structures in the human telomere [146]. Using NMR and UV melting experiments, phase diagrams for the 1:1 mixtures of the $AG_3(T_2AG_3)_3$ and $(C_3TA_2)_3C_3T$ sequences were proposed in 100 mM NaCl or KCl. It was observed that the tetraplex structures are predominant at pH lower than 5, approximately, in presence of KCl. In a coetaneous work, Sugimoto *et al.* showed that for these two sequences, at pH 7, 100 mM NaCl and 0°C, the duplex is the predominant species, being the concentrations of the intramolecular structures residual [147]. The binding constant of the two DNA strands in the presence of 10 mM Mg^{2+} at pH 7.0 was shown to be $5.3 \cdot 10^7$ M^{-1} at 20 °C, about 400 times larger than that in the presence of 100 mM Na^+ at pH 5.5. Using multivariate data analysis methods, the relative concentrations of all species involved in the equilibrium between intra- and intermolecular species in the case of C and G-rich regions at the *bcl-2* [148], *c-kit* [149] and *n-myc* [67] genes have been determined. In all these three cases, the concentration contribution of the tetraplex structures at pH 7 and 25°C is below 10%.

The <u>thermodynamic parameters</u> behind the competition between the Watson-Crick duplex and the intramolecular structures have been also calculated. Lee *et al.* reported the measured (using ITC) and calculated (using Hess cycles) enthalpies for the reactions involving G-quadruplex structures and C-rich sequences [150]. The overall results showed that the favourable free energy terms for the interaction of DNA intramolecular complexes with its complementary strands are enthalpy driven. The main observation is that an intramolecular structure can be disrupted by a complementary strand, provided that its length and sequence is appropriate. However, the favourable free energy term of these targeting reactions may well be increased by improving the stability of the duplex products, by using longer single strands with complementary sequences and/or DNA intramolecular structures with loops containing a larger number of unpaired bases.

The <u>kinetics</u> of the duplex formation from the equimolar mixture of G-/C-rich complementary sequences has been also studied. Li *et al.* investigated this transition at both pH 7.0 and pH 5.5 [151]. Fitting to a single-exponential function gave an observed formation rate of $8 \cdot 10^{-3}$ s^{-1} at 20 °C in 10 mM Mg^{2+} buffer at pH 7.0, which was about 10 times the observed rate at pH 5.5 under the same conditions. Both of the observed rates increased as temperature rose, implying that the dissociation of the intramolecular structures was the rate-limiting step for the Watson-Crick duplex formation. In base of SPR measurements involving immobilized DNA, it has been proposed that pH affects to the association process involving the single strands, whereas it has little effect on the dissociation step of the Watson-Crick duplex [141]. Hence, whereas the value for the association rate is $\sim 10^6$ $M^{-1} \cdot s^{-1}$ in the pH range 4.8 – 7.0, the corresponding value for the dissociation rate is reduced from 10^{-3} s^{-1} (at pH 4.8) to 10^{-4} s^{-1} (at pH 7.0). Overall, these values correspond to association equilibrium constants around $\sim 10^8 – 10^9$ M^{-1}, values higher than those obtained in solution.

Recently, it has been demonstrated that the formation of G-quadruplex and i-motif conformations destabilize directly proximal duplex regions [152]. It has been shown that the large diameter of these folded conformations is mitigated with increased distance from the duplex region: a spacing of five base pairs or more is sufficient to maintain duplex stability proximal to predicted G-quadruplex/i-motif-forming regions. This difference in

stability reflects a stronger contribution from entropy than from enthalpy, although both kinetic and thermodynamic influences are present.

The effect of pH and cation on the structures and stabilities of the isolated sequences $G_4T_4G_4$ and $C_4A_4C_4$ and of their mixtures were studied at 5°C and 100 mM NaCl [153]. In addition to the expected effect of pH, it was found that Ca^{2+} ions induced a parallel G-quadruplex structure and then inhibiting the duplex formation at pH 6. Interestingly, however, the addition of Mg^{2+} to the equimolar mixture of $G_3(T_2AG_3)_3$ and 5-'$(C_3TA_2)_3C_3$ at pH 7 stabilizes the duplex and destabilizes the G-quadruplex [147].

To understand the structure of biomolecules *in vivo*, their properties studied *in vitro* are extrapolated to the *in vivo* condition, while the condition in a living cell is inherently molecularly <u>crowded</u> and a non-ideal solution contains various biomolecules. The effect of molecular crowding on the structure and stability of the telomeric G-rich and C-rich sequences has been studied. Cell-mimic crowding can increase i-motif stability at acid pH and cause dehydration [154]. These crowding conditions are achieved at pH 5.5 using concentrations of polyethylene glycol (PEG) 200 (average molecular weight of 200) ranging from 0 to 50 % (w/w). However, crowding cannot induce i-motif formation at physiological pH. On the other hand, it was shown that, in crowding conditions, the 1:1 (G-rich:C-rich) mixture folds into the parallel G-quadruplex and i-motif but does not form the Watson-Crick duplex, as observed in absence of crowding conditions [155]. The ITC measurements indicated that the thermodynamic stability ($\Delta G°_{20}$) of the duplex formation between the G-rich and C-rich DNAs in the non-crowding condition was −10.2 kcal mol^{-1}, while only a small energy transfer was observed in the ITC measurements in the molecular crowding condition. Overall, these results suggested that the molecular crowding condition may prevent the Watson-Crick duplex formation.

The competition between intramolecular tetraplex structures and intermolecular Watson-Crick has been used to construct a nanomachine that, depending on the pH of the medium, is able to bind or release the telomere-binding protein TRF 1, and to release small quadruplex-binding molecules to impede progress of the polymerase [41]. In a similar approach, a 36-mer circular oligonucleotide containing both a lateral 25-mer i-motif structure and a 11-mer A-rich region was synthesized [43]. The A-rich region was able to form Watson-Crick interactions with the corresponding T-rich complementary strand. At neutral pH, the i-motif structure is not formed and the duplex predominates. At low pH values, the formation of the i-motif structure generates a backbone bending on the non-C-rich segment that prevents the formation of the Watson-Crick duplex.

SUPRAMOLECULAR I-MOTIF STRUCTURES

C-rich sequences showing self-associative properties may form supramolecular (or higher order) structures (sms) that may be potentially interesting from a nanotechnological point of view. Among the pioneer works in this field, Ghodke *et al.* presented a strategy to build 1D scaffolds by using i-motif structures [156]. The formation of these structures, known as I-wires, was induced from a highly concentrated solution of the monomeric C_7 or C_9 sequence by annealing at pH 5.5 from 90°C to 25°C followed by incubation at 4°C. The growth propagates through non-Watson–Crick base-pairing and leads to nanowires more than 3 μm long. Recently, the structure and mechanical properties of DNA i-motif nanowires were studied by means of molecular dynamics simulations [77].

The formation of sms has been investigated in sequences containing two C stretches of unequal length (C_nXC_m) [54]. These sequences may associate into a tetrameric i-motif with a core of $C \cdot C^+$ base pairs and two dangling non-intercalated strands of the shorter C stretch at each end (Figure 8). These dangling ends allow the formation of the sms structures by interacting with other tetrameric i-motif structures. Interestingly, the formation of these sms structures competes with the formation of dimeric i-motif structures, and is dependent on pH, DNA concentration and temperature. Later, it was shown that C stretches as short as CC can link i-motif tetramers into sms [56]. By using NMR and SEC, it was observed that the sms grew in competition with i-motif tetramers or dimers and their formation rate was controlled by the availability of the building block involved in sms formation. The sms stability increased with the number of cytidines in the shorter C stretch. The comparison of the sms lifetimes with those of i-motif tetramers showed that incorporation of tetramers into a larger structure has a stabilizing effect. The presence of a single non-C residue in the oligocytidine sequences prevents the formation of structures with staggered strands, an effect certainly favorable to sms elongation into linear structures. Due to sugar backbone stretching induced by systematic intercalation, the i-motif is a stiff structure. Hence, association of i-motif tetramers into sms can potentially form unbendable rods with a particularly large persistence length.

Figure 8. Postulated association pathway of C_7GC_4 into i-motif sms. The monomer is in equilibrium with one or several hairpin dimer(s) and Te, the tetramer formed by full intercalation of the C_7 stretches. Association and mutual intercalation of the C4 stretches of two Te building blocks result in the formation of the Te2 species. The i-motif symmetry gives to the assembly of several building blocks a elongation capacity similar to that of the building blocks themselves and allows the association of preformed Te_n and Te_m species into structures including (n+m) building blocks. Reprinted with permission [54].

Apart from i-motif sms, a mixed strategy involves the formation of DNA pillars from i-motif stems and Watson-Crick duplex branches [157]. The central stem has some overhanging structures that can enable hybridization with complementary units by Watson–Crick pairing and, thus, multiple i-motifs can join to form the pillar.

Complex DNA molecules containing parallel and antiparallel duplex elements as well as i-motif structures have been designed and synthesized recently [158]. Oligonucleotide duplexes with parallel orientation containing reverse Watson–Crick A·T base pairs and short C_2 tails were shown to be stabilized under slightly acidic conditions by $C \cdot C^+$ base pairs. Corresponding molecules with antiparallel orientation containing Watson–Crick A·T base pairs did not show this phenomenon. This chimeric DNA with parallel duplex elements and long C_5 tails at one or at both ends assemble to tetrameric i-motif structures, even at pH 6.8. Moreover, molecules with two terminal C_5 tails form multimeric assemblies which have the potential to form nanoscopic scaffolds. On the contrary, antiparallel hybrid molecules were only able to aggregate into multimeric assemblies at pH 6.0.

I-MOTIF RNA?

In general, oligoribonucleotides form less stable i-motif structures than the corresponding oligodeoxynucleotides. The first work that studied the potential formation of i-motif structures by oligoribonucleotides described a difference in T_m of 29°C for the 18-mer DNA *vs.* the corresponding 18-mer RNA, indicating a lower stability of RNA i-motif structure. In fact, oligoribonucleotides were shown to adopt

preferably a triplex structure, instead of the i-motif [159]. As uracil-substituted oligodeoxynucleotides did not destabilize the i-motif, it was deduced that the methyl group of T does not play a role in the stabilization of the motif and, consequently, the difference between RNA and DNA must arise from other sources, such as the presence/absence of the 2'-OH group, the sugar conformation or other steric hindrances. Collin *et al.*, studying the relative thermal stabilities of DNA and mixed DNA/RNA tetrads [22], proposed an explanation for the inability of RNA to form a stable i-motif. As the glycosidic angles χ in DNA i-motif are generally high anti conformation and sugar puckers are mostly C3'-*endo*, which are also typical for RNA A-type duplexes, the steric hindrance between 2'-hydroxyls in the narrow groove should be most responsible for the absence of an RNA *i*-motif. This was shown by the positional dependence of *i*-motif stability in the studied dihydroxylated tetrads, by the absence of effect of a 2'-hydroxyl substitution in tetrads containing arabinose and by the complete intolerance of 2'-*O*-methyl modifications.

The spatial structure of an RNA i-motif was shown to be very similar to the corresponding DNA i-motif, being difficult to decide whether the differences between them are significant [160]. As example, the $r(UC_5)$ sequence forms two i-motif structures that differ by their intercalation topologies. The stacking topology of the main structure avoids one of the six 2'-OH/2'-OH repulsive contacts expected in a fully intercalated structure. The C3'-*endo* pucker of the RNA sugars and the orientation of the intercalated C·C$^+$ pairs result in a modest widening of the narrow grooves at the steps where the hydroxyl groups are in close contact. Finally, as observed previously, the free energy of the RNA i-motif, on average -4 kJ mol^{-1} per C·C$^+$ pair, is half of the value found in DNA i-motif structures.

Hybrid i-motifs may be a way to increase the poor stability of RNA i-motifs. It has been shown that a hybrid consisting of two DNA strands and two RNA strands is formed faster than the corresponding DNA i-motif [161]. However, the thermodynamically more stable structure corresponds to the DNA i-motif, as disproportion is observed after 5 days. A mixed strategy involves the formation of hybrid i-motif from a binary mixture of C-rich RNA and PNA sequences [162].

I-MOTIFS *IN VIVO*?

The Watson-Crick duplex is thermodynamically more stable than the intramolecular structures at the physiological conditions of pH and temperature. However, it is known that nuclear processes such as transcription, replication, recombination and repair produce negative supercoiling. A way to reduce the stress produced by this negative supercoiling is the local unwinding of the double helix [11]. In these conditions, the formation of intramolecular structures, such as G-quadruplex or i-motif, could be favoured over the Watson-Crick duplex. On the other hand, it seems that the requirement of an acidic pH value to maintain a stable i-motif structure is a barrier that cannot be overcame. Dysregulated pH is known to be an adaptive feature of most cancers, regardless of their tissue origin or genetic background. In normal differentiated adult cells, intracellular pH is generally lower (around 7.2) than the extracellular pH (around 7.4). However, cancer cells have a higher intracellular (around 7.4) and a lower extracellular pH (6.7–7.1) [163]. In these conditions, C-rich sequences may adopt i-motif structures and modulate the formation of the other nucleic acid structures. It is also worth mentioning that some biological process can provoke local acidification in the cell. For example, Poly(ADP-ribose) polymerases (PARPs) produce 1 mol of proton and nicotinamide for each mol of NAD consumed. This reaction could cause temporary acidification allowing the transient formation of i-motif structures.

The crowding nature of the *in vivo* environment must be taken into account. In general, the effect induced by the crowded intracellular environment is defined by the excluded volume effect and dehydration effect. The role of these effects on the formation and stability of i-motif structures has been studied extensively. The results, however, are not conclusive; whereas some studies suggest the formation of i-motif at neutral pH under crowding conditions [164], other studies did not observe such formation [13]. Rajendran *et al.* reported evidence that the molecular crowding induces the formation of the intramolecular i-motif structure formed by the $5'$-$CG_2(C_2T)_nCG_2$-$3'$ sequence, where n = 4, 6, 8 and 10, at pH 7.0 and 122mM Na^+. 20% (w/w) PEG 2000 and 8000 were used as cosolutes because they may mimic the intracellular environments [164]. Recently, the same cosolute was used to study *c-myc* promoter sequence i-motifs. It has been proposed that the *c-myc* i-motif can exist as a stable structure at pH values as high as 6.7 in 40% w/w polyethylene glycols having molecular weights up to 12,000 g/mol [165].

Despite the uncertain existence of i-motif structures *in vivo*, it has been found that the existence of C-rich sequences that could potentially form i-motif structures may be a source of errors in the Polymerase Chain Reaction (PCR) amplification of DNA fragments [166]. Undetected, this phenomenon may produce systematic errors in genetic analyses that may lead to misdiagnoses in clinical settings and, in consequence, the authors propose that PCR products should be checked for G-quadruplex and i-motifs to avoid the formation of allele dropout-causing secondary structures.

Telomeric DNA

Two of the most studied G-rich and C-rich sequences are those corresponding to the end of the human telomeres. This region of the genome, which plays an important role in cell replication, has been related to cancer diseases as well as aging. The telomere is a superstructure that protects the telomere from degradation. It is known that the length of the telomere is shortened after each cell replication cycle. After a series of cycles, the telomere is so short that it cannot protect the telomere from degradation and the cell dies. However, the telomeric specific polymerase, telomerase, prevents the shortening of the telomere. Hence, a relatively high activity of the telomerase enzyme is detected in 80-90% of tumour cells [167].

In this context, the C-rich strand of human telomere DNA, which could form an i-motif structure, has been reported to transiently exist as a 5' single-stranded overhang at the chromosome ends in the S-phase of replicating human cells [140]. This strand, which repetitive sequence is $5'$-$(C_3TA_2)_n$-$3'$ has been studied in depth since the early works on the i-motif structure [25,35,55] (Table 1). Depending on the experimental conditions and on the specific sequence studied, different structures may be observed. Hence, by using NMR methods, it has been observed the formation of additional T·T or A·A base pairs that help to stabilize the structure. Slow conversion of 3'E and 5'E conformations has also been observed for a tetrameric structure based on *Tetrahymena* sequence [92]. In the case of the human telomere, it seems that intramolecular folding of long sequences favours the formation of the 5'E conformation [168]. It has been proposed that long telomere sequences may show different behavior than truncated sequences [51,91]. Hence, while the 9-nt sequence adopts a bimolecular i-motif structure, the double repeat (12-nt) sequence exists in two (bimolecular and tetramolecular) forms.

The hypothetical telomeric i-motif might be an important and interesting matter of study in biomedical research. Firstly, it has been proposed as an anti-cancer target in gene regulation processes due to its potential ability to inhibit telomerase and to stop the multiplication of cancer cells. Hence, it has been reported recently that the i-motif may interfere to RNA transcription [100]. Secondly, the i-motif structures have been proposed to play an important role in chromosome recognition [169]. In eukaryotes cells the sister

chromatids are meet together until they separate in mitosis to produce two daughter cells. Thanks to these structures the two antiparallel ends of telomeres might associate.

CENTROMERIC SATELLITE DNA

The sequence $TC_2CGT_3C_2A$ is part of a region of the human centromeric alpha-satellite DNA called the CENP-B box, the binding site of centromere protein B (CENP-B). At acidic pH, this sequence and the full $TC_3GT_3C_2A_2CGA_2G$ CENP-B box strand all fold and dimerize in solution forming a stable bimolecular structure containing two GT_3 hairpin loops that interact through a G:T:G:T tetrad [29]. The stem region of the dimer is a four-stranded intercalated motif in which two hairpins associate in a head-to-head way. These minor groove tetrads have been observed recently in other i-motifs structures and may contribute to stabilize the $C·C^+$ base-pairs [28]. Later, the folding of this sequence was studied by means of molecular dynamics simulations (see above). Another sequence, belonging to the human centromeric satellite III showed a pH-dependent intercalation topology, as discussed above [88].

C-RICH REGIONS IN GENE SEQUENCES

Apart from telomeric regions, potential i-motif-forming sequences have been found in more than 40% of gene controlling regions. Table 2 summarizes the studied sequences. It has been speculated that the formation of i-motif structures near the promoter regions of these genes could have a biological function, such as chromosomal translocation [91]. As example, the expression of *c-myc* gene produces a protein that activates telomerase, altering the activity of transcription factors and forcing cell growth forward [170]. The formation of different tumours may be caused by over expression of the *c-myc* gene. Therefore, one of the first studied i-motif structures near promoter regions was that of *c-myc* [171]. It was shown that a 33-base-long sequence adopts several different i-motif structures, being remarkable its stability at pH 7 [170]. Since then, many additional studies have been performed on this or similar sequences. Dettler *et al.* studied a shorter sequence and found that it can adopt multiple "i-motif-like", i-motif and single-stranded structures as a function of pH. The i-motif is predominant in the pH range near the pK_a of C, whereas the "i-motif-like" is the most significant species at higher pH values [172]. Recent results on the sequence $5'-TC_4AC_2T_2C_4AC_3TC_4AC_3T-3'$ [173] indicate that the intercalative $C·C^+$ base pairs are not always necessary for an intramolecular i-motif, a similar conclusion proposed for cyclic nucleotides [28]. It seems that the dynamic character of the i-motif formed by this sequence is intrinsic to it and appears to provide additional stability to the i-motif. This dynamic character may be the responsible for protein binding. The C-rich sequences in *c-myc* gene have been also used to study the competition between intramolecular Watson-Crick vs. intermolecular (G-quadruplex and i-motif) structures. As commented above, the presence of the G-rich complementary sequence induces the formation of the Watson-Crick duplex at neutral pH values. However, it has been shown that mutation of C to T bases may shift the equilibrium [174]. On the other hand, it has been proposed that negative superhelicity may also shift the equilibrium, inducing the formation of the intramolecular G-quadruplex and i-motif structures, even at neutral pH values [175].

Another well-studied C-region is that found near the promoter region of the *bcl-2* gene. The expression of this gene produces a protein that inhibits cell death and causes their survival. It has been demonstrated, through mutational studies coupled with bromine footprinting, that the sequence $CAGC_4GCTC_3GC_5T_2C_2T$ $C_3GCGC_3GC_4T$ may form an i-motif showing a 8:5:7 loop folding pattern [176]. The presence of these long, lateral loops, also observed in the *n-myc* gene (see below), raises the question about their biological significance or advantage in relation to the stability of the i-motif. A possible reason for the larger loop sizes in the *bcl-2* promoter is that these are needed to provide additional stability, perhaps through the formation of capping

structures. Another reason could lie on a role as recognition scaffolds for specific interaction with nuclear proteins or other molecules. A smaller sequence was also studied by means of spectroscopic techniques and multivariate data analysis methods [79]. It was postulated the existence of a intermediate species between the i-motif (at pH near the pK_a of C) and the single strand (at neutral pH values) [79]. This is in accordance with the later proposal of the "i-motif-like" species for the *c-myc* gene above mentioned.

Guo *et al.* studied the formation of i-motif structures in 24-bases long sequence belonging to the human vascular endothelial growth factor (VEGF) [177,178]. The structure was proposed to have six $C \cdot C^+$ intercalated base pairs with a dynamic equilibrium between a 2:3:3 and a 2:3:2 loop configuration. At pH 8, bromine footprinting assays revealed that the studied sequence seems to form partially unfolded i-motif structures.

INTERACTION OF C-RICH SEQUENCES WITH PROTEINS

The interaction of C-rich DNA strands with proteins has been a matter of study for many years because the geometry and charge distribution of i-motifs make it an attractive model for specific structural recognition of DNA by proteins [179]. However, it is not clear that the interaction of a protein with a given C-rich sequence is produced through the previous formation of the i-motif structure. Many of the published studies have been carried out at neutral or slightly basic pH values and the formation of the i-motif has not been demonstrated.

Poly-C-binding proteins (PCBP) are ubiquitous oligonucleotide-binding proteins in eukaryotic cells that play a fundamental role in the regulation of gene expression via interaction with C-rich oligonucleotides. The family consists of the archetypal hnRNP K (heterogeneous nuclear ribonucleoprotein K) and isoforms of αCP1 (also known as PCBP) including αCP1-4 and αCP-KL [180]. In addition to their more recognized ability to bind RNA, αCPs have also been shown to bind single stranded DNA. The closely related hnRNP K is established as a transcription factor, binding to the CT element in the promoter region of *c-myc* [181]. The αCPs, on the other side, have also been found to recognize the C-rich strand of human telomeric DNA with high affinity [179]. Of all the αCPs, αCP1 in particular, showed remarkable specificity for the telomeric $(CCCTAA)_n$ repeat motif [182].

In one of the pioneer works, a 39-kDa polypeptide from *Trypanosoma brucei* was shown to have selective affinity for the C-rich strands of the telomere repeats [183]. Unfortunately, the protein was not identified at that time. In an almost parallel work, Marsich *et al.* showed that HeLa nuclear extracts contain a protein that binds with high specificity to the single-stranded $(C_3TA_2)_n$ repeat. Electrophoretic mobility shift assays showed that the oligonucleotide $(C_3TA_2)_3C_3T$ forms a stable complex with this protein when incubated at pH 8 [184]. On the other hand, the interaction of a well-characterized protein, DDP1, a homologous to the multi-KH domain proteins, with a 12-nt long C-rich sequence was also observed [185]. Lacroix *et al.* studied the interaction of two proteins (hnRNP K and ASF/SF2) with the telomeric C-rich sequence throughout a relatively wide pH range (6-9.2). Their results did not proof that the oligonucleotide, once bound to the protein, was still in the i-motif configuration. However, they pointed out to the fact that the protein binds to the unfolded sequence and is able to open the i-motif at acidic pH [179]. In a later work, Bandiera *et al.* studied the interaction between the C-rich telomeric sequence and a series of heterogeneous nuclear ribonucleoprotein subgroups (hnRNP). Once again, their research was done at experimental conditions where the i-motif structure is quite unstable, like pH values near 8. The work demonstrated that the interaction of the proteins takes place with the single-stranded [182]. Similar studies have showed the interaction of hnRNP proteins with a C-rich sequence that belongs to the human VEGF promoter [178]. In this case, the chosen protein was hnRNP K, which had been previously shown to bind C-rich sequences [179,186]. The results suggested that hnRNP K binds to the VEGF C-rich sequence in a different conformation than that of

the i-motif structure. Very recently, the interaction of an i-motif structure near the promoter region of the *bcl-2* gene with the hnRNP LL transcriptional factor has been studied [187]. As the protein unfolds the i-motif structure to yield a stable single-stranded DNA-protein complex, it is concluded that the i-motif may act as a molecular switch that controls gene expression. The interaction of appropriate i-motif-specific ligands may be a way to shift the equilibrium involving the formation of the DNA-protein complex, thus modulating gene expression.

The crystal structures of the KH1 domain of the PCBP-2 protein in complex with the sequences A_2C_3TA and $(A_2C_3T)_2$ were obtained [188]. The KH-domain has a specific fold in which a stable three-stranded antiparallel β-sheet is packed against three α-helices on one face. In both cases, the protein: DNA ratio is 2:1. Also, the structure does not show the formation of $C \cdot C^+$ base pairs, neither the folding of the C-rich strand into an i-motif structure. Using SPR, it has been demonstrated that the KH1 domain makes the most stable interactions with both RNA and DNA [180]. SPR experiments, with a series of poly-C-sequences revealed that C is preferred at all four positions in the oligonucleotide binding cleft and that a C-tetrad binds KH1 with 10 times higher affinity than a C-triplet.

It has been also reported the association of proteins of *Saccharomyces cerevisiae* with a model sequence of the C-rich telomeric strand, $(C_3ACA)_3C_3$. [189]. A gel retardation assay of the yeast protein extract, in conditions where the DNA fragment folds into an intramolecular i-motif, showed the formation of one major retarded band. Differentially bound proteins were identified as Imd2p, Imd3p and Imd4p. These similar proteins are analogs of the two human NAD-dependent inosine 5′-monophosphate dehydrogenases (IMPDH) which occur as tetramers. Most of the assays gave identical results at pH 8 than at pH 6, at which the C-rich sequence folds into an i-motif. However, the results did not allow confirming that the interaction protein-DNA involves the formation of the i-motif structure. Finally, the interaction of C-rich DNA with proteins that recognize single stranded DNA may be also the basis of an analytical method for the detection of i-motif formation by SPR [190]. In this case, a single-stranded DNA binding protein from E. coli was used.

DNAZYMES

Finally, another biological field where C-rich regions have been studied is that of DNAzymes [191,192], which are catalytic nucleic acids. Choosing the adequate cytosine-rich sequences active DNAzymes can be produced at either acidic or basic pH values. This pH-induced interchangeable activation and deactivation of a cation (preferably Mg^{2+})-dependent DNAzyme is due to the self-assembly of the i-motif structures.

CONCLUSIONS

The interest of researchers working in DNA for the i-motif structure seemed to decrease in past years due to the general thought that an acidic pH was necessary to stabilize the structure *in vivo*. However, the interest for this structure has increased again. The reasons for this change could be the discovery of proteins that may bind C-rich sequences, despite the fact that it is not still clear whether the interaction takes place through the i-motif or through the single unstacked strand. On the other hand, the low pH of endosomes and the acid pH of the tumor microenvironment as a result of the active metabolism of cancer cells increase the interest for pH responsive systems for selective delivery and i-motif is an interesting pH sensitive DNA scaffold. In this sense, i-motif sequences have been used to cap mesoporous silica nanoparticles generating a versatile delivery device that can open and close the pores by changing the pH [193] . Moreover, the i-motif structure have great potential applications in nanotechnology. The open / closed nature of the structure,

together with the fact that the pH-range where this opening takes place may be modulated by choosing the right sequence, allows for new applications such as molecular switches, biosensors and nanomachines [13].

ACKNOWLEDGEMENTS

REFERENCES

1.   A. Bacolla, M. Wojciechowska, B. Kosmider, J. E. Larson, and R. D. Wells, *DNA Repair (Amst).*, 2006, 5, 1161–1170.

2.   G. Biffi, D. Tannahill, J. McCafferty, and S. Balasubramanian, *Nat. Chem.*, 2013, 5, 182–6.

3.   G. Biffi, M. Di Antonio, D. Tannahill, and S. Balasubramanian, *Nat. Chem.*, 2014, 6, 75–80.

4.   E. Y. N. Lam, D. Beraldi, D. Tannahill, and S. Balasubramanian, *Nat. Commun.*, 2013, 4, 1796.

5.   K. Gehring, J. L. Leroy, and M. Gueron, *Nature*, 1993, 363, 561–565.

6.   C. Kang, I. Berger, C. Lockshin, R. Ratliff, R. Moyzis, and A. Rich, *Proc. Natl. Acad. Sci. U. S. A.*, 1994, 91, 11636–11640.

7.   P. Alberti, A. Bourdoncle, B. Saccà, L. Lacroix, and J. L. Mergny, *Org. Biomol. Chem.*, 2006, 4, 3383–3391.

8.   D. Liu and S. Balasubramanian, *Angew. Chemie - Int. Ed.*, 2003, 42, 5734–5736.

9.   D. E. Gilbert and J. Feigon, *Curr. Opin. Struct. Biol.*, 1999, 9, 305–314.

10.   M. Gueron and J. L. Leroy, *Curr. Opin. Struct. Biol.*, 2000, 10, 326–331.

11.   S. Kendrick and L. H. Hurley, *Pure Appl. Chem.*, 2010, 82, 1609–1621.

12.   J. Choi and T. Majima, *Chem. Soc. Rev.*, 2011, 40, 5893–5909.

13.   J. Choi and T. Majima, *Photochem. Photobiol.*, 2013, 89, 513–522.

14.   A. L. Lieblein, M. Kramer, A. Dreuw, B. Furtig, and H. Schwalbe, *Angew. Chemie - Int. Ed.*, 2012, 51, 4067–4070.

15.   B. Yang and M. T. Rodgers, *J. Am. Chem. Soc.*, 2014, 136, 282–290.

16.   I. Berger, M. Egli, and A. Rich, *Proc. Natl. Acad. Sci.*, 1996, 93, 12116–12121.

17.   J. L. Leroy, K. Snoussi, and M. Gueron, *Magn. Reson. Chem.*, 2001, 39, S171–S176.

18.   V. A. Bloomfield, D. M. Crothers, and I. J. Tinoco, *Nucleic acids. Structures, properties, and functions*, University Science Books, Sausalito, CA, 2000.

19.   A. T. Phan and J. L. Leroy, *J. Biomol. Struct. Dyn.*, 2000, 17, 245–251.

20.   L. Cai, L. Chen, S. Raghavan, A. Rich, R. Ratliff, and R. Moyzis, *Nucleic Acids Res.*, 1998, 26, 4696–4705.

21.   A. L. Lieblein, J. Buck, K. Schlepckow, B. Furtig, and H. Schwalbe, *Angew. Chemie - Int. Ed.*, 2012, 51, 250–253.

22.   D. Collin and K. Gehring, *J. Am. Chem. Soc.*, 1998, 120, 4069–4072.

23.   F. Geinguenaud, J. Liquier, M. G. Brevnov, O. V Petrausken, Y. I. Alexeev, E. S. Gromova, and E. Taillandier, *Biochemistry*, 2000, 39, 12650–12658.

24.   J. L. Leroy, K. Gehring, A. Kettani, and M. Gueron, *Biochemistry*, 1993, 32, 6019–6031.

25.   S. Ahmed, A. Kintanar, and E. Henderson, *Nat. Struct. Biol.*, 1994, 1, 83–88.

26.   J. L. Leroy and M. Gueron, *Structure*, 1995, 3, 101–120.

27.   M. Canalia and J. L. Leroy, *Nucleic Acids Res.*, 2005, 33, 5471–81.

28.   N. Escaja, J. Viladoms, M. Garavás, A. Villasante, E. Pedroso, and C. González, *Nucleic Acids Res.*, 2012, 40, 11737–11747.

29.   J. Gallego, S. H. Chou, and B. R. Reid, *J. Mol. Biol.*, 1997, 273, 840–856.

30.   M. Canalia and J. L. Leroy, *J. Am. Chem. Soc.*, 2009, 131, 12870–12871.

31.   J. Dai, A. Ambrus, L. H. Hurley, and D. Yang, *J. Am. Chem. Soc.*, 2009, 131, 6102–6104.

32.   L. Chen, L. Cai, X. Zhang, and A. Rich, *Biochemistry*, 1994, 33, 13540–13546.

33.   I. Berger, L. Cai, L. Chen, and A. Rich, *Biopolym. - Nucleic Acid Sci. Sect.*, 1997, 44, 257–267.

34.   J. L. Mergny, L. Lacroix, X. Han, J. L. Leroy, and C. Helene, *J. Am. Chem. Soc.*, 1995, 117, 8887–8898.

35.   G. Manzini, N. Yathindra, and L. E. Xodo, *Nucleic Acids Res.*, 1994, 22, 4634–4640.

36.   H. Liu, Y. Xu, F. Li, Y. Yang, W. Wang, Y. Song, and D. Liu, *Angew. Chemie - Int. Ed.*, 2007, 46, 2515–2517.

37.   A. I. S. Holm, L. M. Nielsen, B. Kohler, S. V Hoffmann, and S. B. Nielsen, *Phys. Chem. Chem. Phys.*, 2010, 12, 3426–3430.

38.   I. J. Lee, J. W. Yi, and B. H. Kim, *Chem. Commun.*, 2009, 5383–5385.

39.   J. L. Mergny, *Biochemistry*, 1999, 38, 1573–1581.

40.   T. Simonsson and R. Sjoback, *J. Biol. Chem.*, 1999, 274, 17379–17383.

41.   Y. Xu, Y. Hirao, Y. Nishimura, and H. Sugiyama, *Bioorganic Med. Chem.*, 2007, 15, 1275–1279.

42.   A. Dembska, P. Rzepecka, and B. Juskowiak, *J. Fluoresc.*, 2013, 23, 807–812.

43.   Y. Wang, X. Li, X. Liu, and T. Li, *Chem. Commun.*, 2007, 4369–4371.

44.   J. Choi, S. Kim, T. Tachikawa, M. Fujitsuka, and T. Majima, *J. Am. Chem. Soc.*, 2011, 133, 16146–16153.

45.   B. Cohen, M. H. Larson, and B. Kohler, *Chem. Phys.*, 2008, 350, 165–174.

46.   S. Dhakal, J. L. Lafontaine, Z. Yu, D. Koirala, and H. Mao, *PLoS One*, 2012, 7, e39271.

47.   S. Dhakal, J. D. Schonhoft, D. Koirala, Z. Yu, S. Basu, and H. Mao, *J. Am. Chem. Soc.*, 2010, 132, 8991–8997.

48.   S. Dhakal, Z. Yu, R. Konik, Y. Cui, D. Koirala, and H. Mao, *Biophys. J.*, 2012, 102, 2575–2584.

49.   Z. Yu and H. Mao, *Chem. Rec.*, 2013, 13, 102–116.

50.   K. Guo, A. Pourpak, K. Beetz-Rogers, V. Gokhale, D. Sun, and L. H. Hurley, *J. Am. Chem. Soc.*, 2007, 129, 10220–10228.

51.   M. Kaushik, N. Suehl, and L. A. Marky, *Biophys. Chem.*, 2007, 126, 154–164.

52.   J. Völker, H. H. Klump, and K. J. Breslauer, *Biopolymers*, 2007, 86, 136–147.

53.   V. Mathur, A. Verma, S. Maiti, and S. Chowdhury, *Biochem. Biophys. Res. Commun.*, 2004, 320, 1220–1227.

54.   A. Laisné, D. Pompon, and J. L. Leroy, *Nucleic Acids Res.*, 2010, 38, 3817–3826.

55.   J. L. Leroy, M. Gueron, J. L. Mergny, and C. Helene, *Nucleic Acids Res.*, 1994, 22, 1600–1606.

56.   E. Guittet, D. Renciuk, and J.-L. Leroy, *Nucleic Acids Res.*, 2012, 40, 5162–5170.

57.   J. M. Benevides, C. Kang, and G. J. Thomas Jr, *Biochemistry*, 1996, 35, 5747–5755.

58.   K. S. Jin, S. R. Shin, B. Ahn, Y. Rho, S. J. Kim, and M. Ree, *J. Phys. Chem. B*, 2009, 113, 1852–1856.

59.   K. S. Jin, S. R. Shin, B. Ahn, S. Jin, Y. Rho, H. Kim, S. J. Kim, and M. Ree, *J. Phys. Chem. B*, 2010, 114, 4783–4788.

60.   P. M. Keane, M. Wojdyla, G. W. Doorley, J. M. Kelly, A. W. Parker, I. P. Clark, G. M. Greetham, M. Towrie, L. M. Magno, and S. J. Quinn, *Chem. Commun.*, 2014, 50, 2990–2992.

61.   F. Rosu, V. Gabelica, L. Joly, G. Grégoire, and E. De Pauw, *Phys. Chem. Chem. Phys.*, 2010, 12, 13448–13454.

62.   A. R. Moehlig, K. E. Djernes, V. M. Krishnan, and R. J. Hooley, *Org. Lett.*, 2012, 14, 2560–2563.

63.   D. M. Hatters, L. Wilson, B. W. Atcliffe, T. D. Mulhern, N. Guzzo-Pernell, and G. J. Howlett, *Biophys. J.*, 2001, 81, 371–381.

64.   S. Wu, X. Wang, X. Ye, and G. Zhang, *J. Phys. Chem. B*, 2013, 117, 11541–11547.

65.   J. B. Chaires, *Top. Curr. Chem.*, 2005, 253, 33–53.

66.   P. Bucek, R. Gargallo, and A. Kudrev, *Anal. Chim. Acta*, 2010, 683, 69–77.

67.   S. Benabou, R. Ferreira, A. Aviñó, C. González, S. Lyonnais, M. Solà, R. Eritja, J. Jaumot, and R. Gargallo, *Biochim. Biophys. Acta - Gen. Subj.*, 2014, 1840, 41–52.

68.   N. J. Greenfield, *Nat. Protoc.*, 2006, 1, 2527–2535.

69.   S. Fernandez, R. Eritja, A. Aviño, J. Jaumot, and R. Gargallo, *Int. J. Biol. Macromol.*, 2011, 49, 729–736.

70.   R. D. Gray and J. B. Chaires, *Curr. Protoc. Nucleic Acid Chem.*, 2011, Unit 17.4.

71.   I. Haq, B. Z. Chowdhry, and J. B. Chaires, *Eur. Biophys. J.*, 1997, 26, 419–426.

72.   R. Gargallo, R. Tauler, and A. Izquierdo-Ridorsa, *Biopolymers*, 1997, 42, 271 – 283.

73.   T. Vojtylova, D. Dospivova, O. Triskova, I. Pilarova, P. Lubal, M. Farkova, L. Trnkova, and P. Taborsky, *Chem. Pap.*, 2009, 63, 731–737.

74.   J. Gallego, E. B. Golden, D. E. Stanley, and B. R. Reid, *J. Mol. Biol.*, 1999, 285, 1039–1052.

75.   T. E. Malliavin, J. Gau, K. Snoussi, and J. L. Leroy, *Biophys. J.*, 2003, 84, 3838–3847.

76.   J. Smiatek, C. Chen, D. Liu, and A. Heuer, *J. Phys. Chem. B*, 2011, 115, 13788–13795.

77.   R. P. Singh, R. Blossey, and F. Cleri, *Biophys. J.*, 2013, 105, 2820–2831.

78.   D. J. Cashman, R. Buscaglia, M. W. Freyer, J. Dettler, L. H. Hurley, and E. A. Lewis, *J. Mol. Model.*, 2008, 14, 93–101.

79.   N. Khan, A. Aviño, R. Tauler, C. Gonzalez, R. Eritja, and R. Gargallo, *Biochimie*, 2007, 89, 1562–1572.

80.   R. A. Zager, B. A. Schimpf, and D. J. Gmur, *Circ. Res.*, 1993, 72, 837–846.

81.   J. Zhou, C. Wei, G. Jia, X. Wang, Z. Feng, and C. Li, *Mol. Biosyst.*, 2012, 6, 580–586.

82.   H. A. Day, C. Huguin, and Z. A. E. Waller, *Chem. Commun.*, 2013, 49, 7696–7698.

83.   P. Fojtik and M. Vorlickova, *Nucleic Acids Res.*, 2001, 29, 4684–4690.

84.   T. Li and M. Famulok, *J. Am. Chem. Soc.*, 2013, 135, 1593–1599.

85.   M. Canalia and J.-L. Leroy, *J. Am. Chem. Soc.*, 2009, 131, 12870–12871.

86.   A. L. Lieblein, B. Fürtig, and H. Schwalbe, *ChemBioChem*, 2013, 14, 1226–1230.

87.   T. A. Brooks, S. Kendrick, and L. Hurley, *FEBS J.*, 2010, 277, 3459–3469.

88.   S. Nonin-Lecomte and J. L. Leroy, *J. Mol. Biol.*, 2001, 309, 491–506.

89.   X. Han, J. L. Leroy, and M. Gueron, *J. Mol. Biol.*, 1998, 278, 949–965.

90.   J. W. Park, Y. J. Seo, and B. H. Kim, *Chem. Commun.*, 2014, 50, 52–54.

91.   M. Kaushik, M. Prasad, S. Kaushik, A. Singh, and S. Kukreti, *Biopolymers*, 2010, 93, 150–160.

92.   N. Esmaili and J. L. Leroy, *Nucleic Acids Res.*, 2005, 33, 213–224.

93.   J. Weil, T. Min, C. Yang, S. Wang, C. Sutherland, N. Sinha, and C. Kang, *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 1999, 55, 422–429.

94.   P. Kumar, A. Verma, S. Maiti, R. Gargallo, and S. Chowdhury, *Biochemistry*, 2005, 44, 16426–16434.

95.   J. Choi, S. Kim, T. Tachikawa, M. Fujitsuka, and T. Majima, *J. Am. Chem. Soc.*, 2011, 133, 16146–16153.

96.   D. Monchaud and M.-P. Teulade-Fichou, *Org. Biomol. Chem.*, 2008, 6, 627–636.

97.   X. Li, Y. Peng, J. Ren, and X. Qu, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, 103, 19658–19663.

98.   Y. Peng, X. Wang, Y. Xiao, L. Feng, C. Zhao, J. Ren, and X. Qu, *J. Am. Chem. Soc.*, 2009, 131, 13813–13818.

99.   Y. Peng, X. Li, J. Ren, and X. Qu, *Chem. Commun.*, 2007, 5176–5178.

100. Y. Chen, K. Qu, C. Zhao, L. Wu, J. Ren, J. Wang, and X. Qu, *Nat. Commun.*, 2012, 3, 1074.

101. T. S. Dexheimer, S. S. Carey, S. Zuohe, V. M. Gokhale, X. Hu, L. B. Murata, E. M. Maes, A. Weichsel, D. Sun, E. J. Meuillet, W. R. Montfort, and L. H. Hurley, *Mol. Cancer Ther.*, 2009, 8, 1363–1377.

102. O. Y. Fedoroff, A. Rangan, V. V Chemeris, and L. H. Hurley, *Biochemistry*, 2000, 39, 15083–15090.

103. S. Shi, X. Geng, J. Zhao, T. Yao, C. Wang, D. Yang, L. Zheng, and L. Ji, *Biochimie*, 2010, 92, 370–377.

104. H. Xu, H. Zhang, and X. Qu, *J. Inorg. Biochem.*, 2006, 100, 1646–1652.

105. X. Chen, X. Zhou, T. Han, J. Wu, J. Zhang, and S. Guo, *ACS Nano*, 2012, 7, 531–537.

106. L. Wang, Y. Wu, T. Chen, and C. Wei, *Int. J. Biol. Macromol.*, 2013, 52, 1–8.

107. C. Xu, C. Zhao, J. Ren, and X. Qu, *Chem. Commun.*, 2011, 47, 8043–8045.

108. D. L. Ma, M. H. T. Kwan, D. S. H. Chan, P. Lee, H. Yang, V. P. Y. Ma, L. P. Bai, Z. H. Jiang, and C. H. Leung, *Analyst*, 2011, 136, 2692–2696.

109. X. Ren, F. He, and Q. H. Xu, *Chem. - An Asian J.*, 2010, 5, 1094–1098.

110. S. Kendrick, H.-J. Kang, M. Alam, M. Madathil, P. Agrawal, V. Gokhale, D. Yang, S. M. Hecht, and L. H. Hurley, *J. Am. Chem. Soc.*, 2014, DOI: 10.1021/ja410934b.

111. R. Hoshyar, S. Z. Bathaie, A. Kyani, and M. F. Mousavi, *Nucleosides, Nucleotides and Nucleic Acids*, 2012, 31, 801–812.

112. A. Latorre and Á. Somoza, *ChemBioChem*, 2012, 13, 951–958.

113. L. Lacroix and J. L. Mergny, *Arch. Biochem. Biophys.*, 2000, 381, 153–163.

114. C. P. Fenna, V. J. Wilkinson, J. R. P. Arnold, R. Cosstick, and J. Fisher, *Chem. Commun.*, 2008, 3567–3569.

115. S. Robidoux and M. J. Damha, *J. Biomol. Struct. Dyn.*, 1997, 15, 529–535.

116. J. L. Mergny and L. Lacroix, *Nucleic Acids Res.*, 1998, 26, 4797–4803.

117. H. Kanehara, M. Mizuguchi, K. Tajima, K. Kanaori, and K. Makino, *Biochemistry*, 1997, 36, 1790–1797.

118. K. Kanaori, S. Sakamoto, H. Yoshida, P. Guga, W. Stec, K. Tajima, and K. Makino, *Biochemistry*, 2004, 43, 5672–5679.

119. J. A. Brazier, J. Fisher, and R. Cosstick, *Angew. Chemie - Int. Ed.*, 2006, 45, 114–117.

120. R. Cosstick, J. Buckingham, J. Brazier, and J. Fisher, *Nucleosides, Nucleotides and Nucleic Acids*, 2007, 26, 555–558.

121. E. E. Swayze, A. M. Siwkowski, E. V Wancewicz, M. T. Migawa, T. K. Wyrzykiewicz, G. Hung, B. P. Monia, and C. F. Bennett, *Nucleic Acids Res.*, 2007, 35, 687–700.

122. N. Kumar, J. T. Nielsen, S. Maiti, and M. Petersen, *Angew. Chemie - Int. Ed.*, 2007, 46, 9220–9222.

123. N. Kumar, M. Petersen, and S. Maiti, *Chem. Commun.*, 2009, 1532–1534.

124. A. Pasternak and J. Wengel, *Bioorganic Med. Chem. Lett.*, 2010, 21, 752–755.

125. P. Perlikova, K. K. Karlsen, E. B. Pedersen, and J. Wengel, *ChemBioChem*, 2014, 15, 146–156.

126. U. Diederichsen, *Angew. Chemie - Int. Ed.*, 1998, 37, 2273–2276.

127. Y. Krishnan-Ghosh, E. Stephens, and S. Balasubramanian, *Chem. Commun.*, 2005, 5278–5280.

128. N. K. Sharma and K. N. Ganesh, *Chem. Commun.*, 2005, 4330–4332.

129. R. Bonnet, P. Murat, N. Spinelli, and E. Defrancq, *Chem. Commun.*, 2012, 48, 5992–5994.

130. S. Modi, A. H. Wani, and Y. Krishnan, *Nucleic Acids Res.*, 2006, 34, 4354–4363.

131. T. Li, D. Liu, J. Chen, A. H. F. Lee, J. Qi, and A. S. C. Chan, *J. Am. Chem. Soc.*, 2001, 123, 12901–12902.

132. D. Liu, J. Chen, A. H. F. Lee, L. M. C. Chow, A. S. C. Chan, and T. Li, *Angew. Chemie - Int. Ed.*, 2003, 42, 797–799.

133. J. S. Hartig and E. T. Kool, *Nucleic Acids Res.*, 2004, 32, 19, e152.

134. T. Zhou, X. Li, M. T. T. Ng, Y. Wang, N. M. Quek, J. Luo, W. Yuan, C. H. Tan, H. Zeng, and T. Li, *Bioconjug. Chem.*, 2009, 20, 644–647.

135. Y. Yang, Y. Sun, Y. Xing, T. Zhang, Z. Wang, Z. Yang, and D. Liu, *Macromolecules*, 2012, 45, 2643–2647.

136. A. A. El-Sayed, E. B. Pedersen, and N. A. Khaireldin, *Nucleosides, Nucleotides and Nucleic Acids*, 2012, 31, 872–879.

137. A. W. I. Stephenson, A. C. Partridge, and V. V Filichev, *Chem. – A Eur. J.*, 2011, 17, 6227–6238.

138. S. Robidoux, R. Klinck, K. Gehring, and M. J. Damha, *J. Biomol. Struct. Dyn.*, 1997, 15, 517–527.

139. F. Seela, S. Budow, and P. Leonard, *Org. Biomol. Chem.*, 2007, 5, 1858–1872.

140. G. Cimino-Reale, E. Pascale, E. Alvino, G. Starace, and E. D'Ambrosio, *J. Biol. Chem.*, 2003, 278, 2136–2140.

141. Y. Zhao, Z. X. Zeng, Z. Y. Kan, Y. H. Hao, and Z. Tan, *ChemBioChem*, 2005, 6, 1957–1960.

142. S. Modi, C. Nizak, S. Surana, S. Halder, and Y. Krishnan, *Nat Nano*, 2013, 8, 459–467.

143. D. Liu and S. Balasubramanian, *Angew. Chem. Int. Ed. Engl.*, 2003, 42, 5734–5736.

144. J.-L. Leroy, *Nucleic Acids Res.*, 2009, 37, 4127–4134.

145. C. Chen, M. Li, Y. Xing, Y. Li, C. C. Joedecke, J. Jin, Z. Yang, and D. Liu, *Langmuir*, 2012, 28, 17743–17748.

146. A. T. Phan and J. L. Mergny, *Nucleic Acids Res.*, 2002, 30, 4618–4625.

147. W. Li, P. Wu, T. Ohmichi, and N. Sugimoto, *FEBS Lett.*, 2002, 526, 77–81.

148. M. del Toro, P. Bucek, A. Aviñó, J. Jaumot, C. González, R. Eritja, and R. Gargallo, *Biochimie*, 2009, 91, 894–902.

149. P. Bucek, J. Jaumot, A. Aviñó, R. Eritja, and R. Gargallo, *Chem. - A Eur. J.*, 2009, 15, 12663–12671.

150. H. T. Lee, C. M. Olsen, L. Waters, H. Sukup, and L. A. Marky, *Biochimie*, 2008, 90, 1052–1063.

151. W. Li, D. Miyoshi, S. I. Nakano, and N. Sugimoto, *Biochemistry*, 2003, 42, 11736–11744.

152. S. L. B. Konig, J. L. Huppert, R. K. O. Sigel, and A. C. Evans, *Nucleic Acids Res.*, 2013, 41, 7453–7461.

153. D. Miyoshi, S. Matsumura, W. Li, and N. Sugimoto, *Nucleosides, Nucleotides and Nucleic Acids*, 2003, 22, 203–221.

154. C. Zhao, J. Ren, and X. Qu, *Chemistry*, 2008, 14, 5435–5439.

155. D. Miyoshi, S. Matsumura, S. I. Nakano, and N. Sugimoto, *J. Am. Chem. Soc.*, 2004, 126, 165–169.

156. H. B. Ghodke, R. Krishnan, K. Vignesh, G. V. P. Kumar, C. Narayana, and Y. Krishnan, *Angew. Chemie - Int. Ed.*, 2007, 46, 2645–2649.

157. Y. Yang, C. Zhou, T. Zhang, E. Cheng, Z. Yang, and D. Liu, *Small*, 2012, 8, 552–556.

158. H. Mei, S. Budow, and F. Seela, *Biomacromolecules*, 2012, 13, 4196–4204.

159. L. Lacroix, J. L. Mergny, J. L. Leroy, and C. Helene, *Biochemistry*, 1996, 35, 8715–8722.

160. K. Snoussi, S. Nonin-Lecomte, and J. L. Leroy, *J. Mol. Biol.*, 2001, 309, 139–153.

161. S. Chakraborty and Y. Krishnan, *Biochimie*, 2008, 90, 1088–1095.

162. S. Chakraborty, S. Modi, and Y. Krishnan, *Chem. Commun.*, 2008, 70–72.

163. B. A. Webb, M. Chimenti, M. P. Jacobson, and D. L. Barber, *Nat Rev Cancer*, 2011, 11, 671–677.

164. A. Rajendran, S. I. Nakano, and N. Sugimoto, *Chem. Commun.*, 2010, 46, 1299–1301.

165. J. Cui, P. Waltman, V. Le, and E. Lewis, *Molecules*, 2013, 18, 12751–12767.

166. J. J. Wenzel, H. Rossmann, C. Fottner, S. Neuwirth, C. Neukirch, P. Lohse, J. K. Bickmann, T. Minnemann, T. J. Musholt, B. Schneider-Raetzke, M. M. Weber, and K. J. Lackner, *Clin. Chem.*, 2009, 55, 1361–1371.

167. J. L. Huppert, *FEBS J.*, 2010, 277, 3452–3458.

168. A. T. Phan, M. Gueron, and J. L. Leroy, *J. Mol. Biol.*, 2000, 299, 123–144.

169. R. D. Wells, D. A. Collier, J. C. Hanvey, M. Shimizu, and F. Wohlrab, *FASEB J.*, 1988, 2, 2939–2949.

170. T. Simonsson, M. Pribylova, and M. Vorlickova, *Biochem. Biophys. Res. Commun.*, 2000, 278, 158–166.

171. T. Simonsson, P. Pecinka, and M. Kubista, *Nucleic Acids Res.*, 1998, 26, 1167–1172.

172. J. M. Dettler, R. Buscaglia, J. Cui, D. Cashman, M. Blynn, and E. A. Lewis, *Biophys. J.*, 2010, 99, 561–567.

173. J. Dai, E. Hatzakis, L. H. Hurley, and D. Yang, *PLoS One*, 2010, 5, e11647.

174. K. Halder, V. Mathur, D. Chugh, A. Verma, and S. Chowdhury, *Biochem. Biophys. Res. Commun.*, 2005, 327, 49–56.

175. D. Sun and L. H. Hurley, *J. Med. Chem.*, 2009, 52, 2863–2874.

176. S. Kendrick, Y. Akiyama, S. M. Hecht, and L. H. Hurley, *J. Am. Chem. Soc.*, 2009, 131, 17667–17676.

177. K. Guo, V. Gokhale, L. H. Hurley, and D. Sun, *Nucleic Acids Res.*, 2008, 36, 4598–4608.

178. D. J. Uribe, K. Guo, Y. J. Shin, and D. Sun, *Biochemistry*, 2011, 50, 3796–3806.

179. L. Lacroix, H. Lienard, E. Labourier, M. Djavaheri-Mergny, J. Lacoste, H. Leffers, J. Tazi, C. Helene, and J. L. Mergny, *Nucleic Acids Res.*, 2000, 28, 1564–1575.

180. Y. M. K. Yoga, D. A. K. Traore, M. Sidiqi, C. Szeto, N. R. Pendini, A. Barker, P. J. Leedman, J. A. Wilce, and M. C. J. Wilce, *Nucleic Acids Res.*, 2012, 40, 5101–5114.

181. E. F. Michelotti, G. A. Michelotti, A. I. Aronsohn, and D. Levens, *Mol. Cell. Biol.*, 1996, 16, 2350–2360.

182. A. Bandiera, G. Tell, E. Marsich, A. Scaloni, G. Pocsfalvi, A. Akindahunsi, L. Cesaratto, and G. Manzini, *Arch. Biochem. Biophys.*, 2003, 409, 305–314.

183. J. E. Eid and B. Sollner-Webb, *Mol. Cell. Biol.*, 1995, 15, 389–397.

184. E. Marsich, A. Piccini, L. E. Xodo, and G. Manzini, *Nucleic Acids Res.*, 1996, 24, 4029–4033.

185. A. Cortés, D. Huertas, L. Fanti, S. Pimpinelli, F. X. Marsellach, B. Piña, and F. Azorín, *EMBO J.*, 1999, 18, 3820–3833.

186. S. Fenn, Z. Du, J. K. Lee, R. Tjhen, R. M. Stroud, and T. L. James, *Nucleic Acids Res.*, 2007, 35, 2651–2660.

187. H.-J. Kang, S. Kendrick, S. M. Hecht, and L. H. Hurley, *J. Am. Chem. Soc.*, 2014, DOI: 10.1021/ja4109352.

188. Z. Du, J. K. Lee, R. Tjhen, S. Li, H. Pan, R. M. Stroud, and T. L. James, *J. Biol. Chem.*, 2005, 280, 38823–38830.

189. J. F. Cornuel, A. Moraillon, and M. Gueron, *Biochimie*, 2002, 84, 279–289.

190. Z. X. Zeng, Y. Zhao, Y. H. Hao, and Z. Tan, *J. Mol. Recognit.*, 2005, 18, 267–271.

191. C. Teller and I. Willner, *Curr. Opin. Biotechnol.*, 2010, 21, 376–391.

192. J. Elbaz, S. Shimron, and I. Willner, *Chem. Commun.*, 2010, 46, 1209–1211.

193. C. Chen, F. Pu, Z. Huang, Z. Liu, J. Ren, and X. Qu, *Nucleic Acids Res.*, 2011, 39, 1638–1644.

194. K. Kanaori, N. Shibayama, K. Gohda, K. Tajima, and K. Makino, *Nucleic Acids Res.*, 2001, 29, 831–840.

195. J. A. Brazier, A. Shah, and G. D. Brown, *Chem. Commun.*, 2012, 48, 10739–10741.

196. S. Saxena, A. Bansal, and S. Kukreti, *Arch. Biochem. Biophys.*, 2008, 471, 95–108.

197. S. S. Pataskar, D. Dash, and S. K. Brahmachari, *J. Biomol. Struct. Dyn.*, 2001, 19, 307–313.

198. P. Catasti, X. Chen, L. L. Deaven, R. K. Moyzis, E. M. Bradbury, and G. Gupta, *J. Mol. Biol.*, 1997, 272, 369–382.

199. V. V Jolad, F. K. Murad, J. R. P. Arnold, and J. Fisher, *Org. Biomol. Chem.*, 2005, 3, 2234–2236.

200. Y. Xu and H. Sugiyama, *Nucleic Acids Res.*, 2006, 34, 949–954.

Table 1. Overview of the telomere sequences studied.

Telomere	Sequence	Instrumental techniques	Relevant features	Reference
Vertebrate telomere	C_3TA_2	NMR, pH 4.5, 100 mM, 10mM DNA	Three distinct tetramers slowly exchange, differing in the intercalation topology	[194]
	$C_3TA_2C_3$	UV, SEC, CD	Dimeric form	[91]
	$C_3TA_2C_3TA_2$	UV, NMR, PAGE (50 mM sodium phosphate or acetate)	Dimeric form	[25]
		UV, SEC, CD (20 mM sodium cacodylate or acetate buffers	Dimer an tetrameric forms in equilibrium	[91]
	$5mCCT_3CCT_3ACCT_3CC$		Additional T·A and T·T base pairs. The two-base loops are sufficient to span the narrow grooves of the i-motif core. 3'E intercalation topology.	[89]
	$C_3TA_2C_3TA_2C_3TA_2C_3$	NMR, SEC, UV	One of the first studies done on i-motif structures by NMR and SEC. The four different configurations in which all cytosines are base-paired and all base pairs are intercalated are discussed.	[55]
	$C_3TA_2C_3TA_2C_3TA_2C_3$	NMR	Study of the influence of loop nucleotides on stability, structure and kinetics of folding	[86]
	$(C_3TA_2)_3C_3T$	NMR	Intramolecular i-motif with 5'E intercalation topology. The second TA_2 linker loops across one of the narrow grooves, while the first and third linkers loop across the wide grooves. Motional averaging between at least two structures of each bottom loop.	[168]
		CD, PAGE, melting experiments	Intramolecular structure, use of a partition function to determine the number of protons involved in the formation of the i-motif structure.	[35]
	$(C_3TA_2)_4$	NMR, PAGE	Dimer	[25]
		DSC, CD and melting experiments in 10 mM buffer and 100 mM NaCl	Sequential melting: bimolecular complex --> intramolecular complex --> random coil	[51]
Tetrahymena telomere	A_2C_4 (PDB codes 294D and 1YBL)	NMR	This is the first example of a 3' terminal C·C+ pair. There are four grooves - two broad and flat major grooves and two extremely narrow minor grooves. The helical twist between covalently linked C·C+ pairs is 12-16°.	[20]
		NMR	The base A2 forms an A2·A2 base pair stacked to C3·C3+ and cross-strand stacked to A1.	[92]
	C_4A_2 (PDB codes 1YBN and 1YBR	NMR	The tetramer adopts two distinct intercalation topologies in slow conformational exchange. One, whose outermost C·C+ pairs are built by the cytidines of the 5' end and the other by those of the 3' end. In both topologies, the adenosine bases are fairly well stacked to the adjacent C·C+ pairs.	[92]
	$(C_4A_2)_3C_4$	NMR, SEC, UV	One of the first studies done on i-motif structures by NMR and SEC. The four different configurations in which all cytosines are base-paired and all base pairs are intercalated are discussed.	[55]

Table 2. C-rich sequences corresponding to the promoter regions of genes that have been studied.

Sequence	Promoter region	Instrumental techniques used in the study	Relevant features	Reference
$C_3GC_4\ T_2C_2TC_3GCGC_3G$	*bcl-2*	CD, NMR, melting experiments, acid-base titrations, multivariate analysis	Two different i-motif structures, differing in the protonation of bases at the loops, are proposed	[79]
$CAGC_4GCTC_3GC_5T_2C_2TC_3GCGC_3GC_3T$	*bcl-2*	CD, FRET, bromine footprinting	One major intramolecular i-motif with a transition pH of 6.6 and a 8:5:7 loop conformation was observed. A novel assay involving the sequential incorporation of a fluorescent thymine analogue at each thymine position is used to provide evidence of a capping structure within the top loop region.	[176]
$C_4TC_3TCGCGC_2GC_3G$	*c-kit*	CD, NMR, melting experiments, acid-base titrations, multivariate analysis	The competition with the Watson-Crick duplex was investigated through a wide pH range.	[149]
$C_3TC_2TC_3AGCGC_3AC_3T$	*c-kit*	Melting experiments	A Class I i-motif which formation was not observed at pH 7.0	[195]
$TA_2T_3C_5TC_5TC_5A_2T$	*c-jun*	Molecular absorption, PAGE	Na^+ favored the duplex formation over the G- and C-rich intramolecular structures.	[196]
$TC_4AC_2T_2C_4AC_3TC_4AC_3TC_4A$	*c-myc*	CD, PAGE	C-rich sequences, differing in their length, form bimolecular or intramolecular i-motif structures. Determination of the number of hemiprotonated $C\cdot C^+$ base pairs from changes in ellipticity upon pH.	[170]
$CT_3C_2TAC_3TC_3TAC_3TA_2$	*c-myc*	NMR	Alternative approach to determine the folding structure of the i-motif. Different intramolecular i-motif structures were observed.	[31]
$CT_3C_2TAC_3TC_3TAC_3TA_2$	*c-myc*	DSC, CD and molecular modelling	An "i-motif-like" structure is proposed at pH higher than 5.2, approximately.	[172]
$TC_4AC_2T_2C_4AC_3TC_4AC_3TC_4A$	*c-myc*	Footprinting	Negative superhelicity induces the formation of i-motif structures from Watson-Crick duplex.	[175]
$TC_4AC_2T_2C_4AC_3TC_4AC_3T$	*c-myc*	NMR	Different intramolecular i-motif structures were observed.	[173]
$C_4AC_2T_2C_4AC_3TC_4AC_3TC_4$	*c-myc*	PAGE, melting experiments	Determination of ΔH, ΔS and ΔG associated to the unfolding of i-motif structures, as well as the number of counterions uptaken or released.	[53]
$(TC_2)_4TTC(TC_2)_4GTG(TC_2)_4$	*c-myb*	Melting experiments	No i-motif formation was observed at pH 7.0	[195]
$C_4GC_4GCG)_{2-3}$	*EPM1 disorder*	CD, PAGE	The stability of the structure increases with the increase in the length of the repeat.	[197]
Several sequences	*HIF-1α*	Melting experiments	Stable near physiological pH and temperature.	[195]
Several sequences	*hTERT*	Melting experiments	Stable near physiological pH and temperature	[195]
$GC(TC_3)_3TC_2T(TC_3)_3$	*k-ras*	CD, PAGE, UV melting experiments	The number of gained protons along the formation of the i-motif is calculated from the standard Hill plot.	[35]
$C_4TGTC_4ACA_4TGTC$; $(C_4TGTC_4ACA)_n$ **(n=6)**	*human insulin*	NMR, molecular dynamics, replication assays	The replication assay shows that for n = 6 a structure (probably an i-motif) that blocks progression of replication even in the presence of its complementary strand is formed.	[198]
$(C_4ACAC_4TGT)_2$	*human insulin*	NMR, UV melting experiments	An intramolecular i-motif that is present at pH 7 and $5\,^oC$	[199]
$AC_5TGCATCTGCATGC_5TC_3AC_5T$	*n-myc*	CD, NMR, PAGE, UV melting experiments, acid-base titrations, multivariate analysis	A Watson-Crick hairpin is formed in the long loop of the i-motif	[67]
$AC_2GCG\ C_4TC_5GC_5GC_5GC_{13}$	*PDGF-A*	Melting experiments	Stable near physiological pH and temperature	[195]
$GC_2GC_3A_4C_6G$	*Rb*	CD, 2-aminopurine fluorescence	Fluorescence changes accompanying a 2-aminopurine-labelled G- quadruplex to duplex transition from addition of the C-rich complementary strand were monitored. The double-helix form was predominant at neutral pH.	[200]
$C_2GC_4CGC_4GC_4GC_4TA$	*RET*	CD, polymerase stop assay, Br_2 footprinting, molecular modelling	A 2:3:2 loop configuration was proposed. The competition with the G-rich sequence was also studied.	[50]
$GAC_4GC_5G_2C_3GC_4G_2$	*VEGF*		A 2:3:2 loop configuration was proposed.	[177]