Published in final edited form as:

*Angew Chem Int Ed Engl*. 2023 June 26; 62(26): e202301666. doi:10.1002/anie.202301666.

# Crystal Structure of an i-Motif from the *HRAS* Oncogene Promoter

**Kevin S. Li**[a], **Deondre Jordan**[a], **Linda Y. Lin**[a], **Sawyer E. McCarthy**[a], **John S. Schneekloth Jr.**[b], **Liliya A. Yatsunyk**[a]

[a]Department Chemistry and Biochemistry, Swarthmore College, 500 College Ave, Swarthmore PA 19081 (USA)

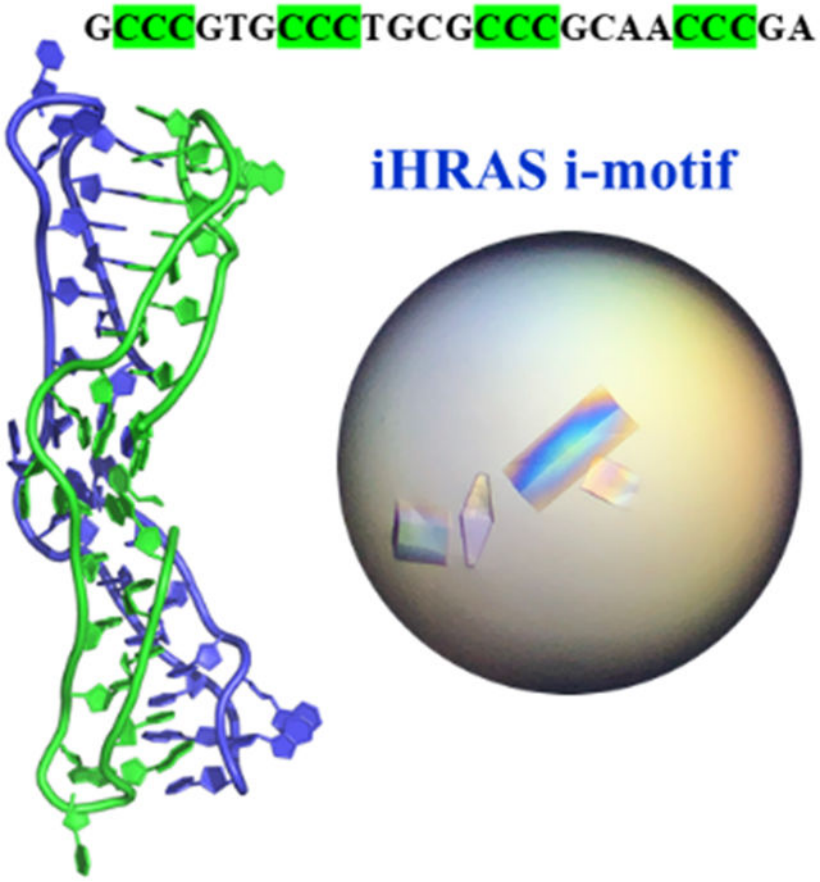
[b]Chemical Biology Laboratory, National Cancer Institute, National Institute of Health, Frederick, MD 21702 (USA)

## Abstract

An i-motif is a non-canonical DNA structure implicated in gene regulation and linked to cancers. The C-rich strand of the *HRAS* oncogene, 5'-CGCCCGTGCCCTGCGCCCGCAACCCGA-3' (here referred to as iHRAS), forms an i-motif *in vitro* but its exact structure was unknown. *HRAS* is a member of the *RAS* proto-oncogene family. About 19% of US cancer patients carry mutations in *RAS* genes. We solved the structure of iHRAS at 1.77 Å resolution. The structure reveals that iHRAS folds into a double hairpin. The two double hairpins associate in an antiparallel fashion, forming an i-motif dimer capped by two loops on each end and linked by a connecting region. Six C-$C^+$ base pairs form each i-motif core, and the core regions are extended by a G-G base pair and C stacking. Extensive canonical and non-canonical base pairing and stacking stabilizes the connecting region and loops. The iHRAS structure is the first atomic resolution structure of an i-motif from a human oncogene. This structure sheds light on i-motifs folding and function in the cell.

## Graphical Abstract

lyatsun1@swarthmore.edu .




The structure of a biologically relevant i-motif from the *HRAS* oncogene was solved to 1.8-Å resolution. The structure is a dimer of two i-motifs formed by six C-$C^+$ pairs. The structure contains 20 base pairs, of which only two are canonical. The extensive network of capping and connecting interactions is unprecedented and suggests that auxiliary interactions are essential for i-motif stability *in vivo*.

**Keywords**

i-motifs; X-ray crystallography; non-canonical base pairing; gene regulation; dimers

## Introduction

An i-motif (iM) is a non-canonical secondary structure formed by C-rich DNA.[1-3] It is composed of two parallel duplexes intercalated in an antiparallel fashion and stabilized by hemi-protonated C-$C^+$ base pairs. Depending on whether the outermost cytosine base pair is formed at the 5’ or the 3’ end, iMs are classified as 5’E or 3’E, respectively.[1,3] *In vitro,* formation of hemi-protonated cytosine base pairs requires acidic conditions.[1] However, molecular crowding, negative superhelicity, chemical modifications, the presence of long C-tracts, interactions with ligands, crystal packing forces, and other factors may favor iM formation even under neutral pH.[4-9]

Strong evidence for the existence of iMs in nuclei of living cells was obtained using an iM-specific antibody, iMab.[10] Interestingly, the iMab foci peak in intensity during the G1/S boundary phase, a period of high transcriptional activity in the cell cycle, potentially linking iMs to gene regulation.[10] C-rich sequences with high iM-forming potential are interspersed throughout the genome and are overrepresented in centromeres, telomeres, and promoter regions. They may act as recognition sites for transcription factors.[11,12] For example, transcription factors hnRNP LL and hnRNP A1 bind to and unfold iMs, leading to transcriptional activation of downstream genes.[11,12]

Sequences with iM-forming potential in *c-Myc, c-Myb, VEGF, PDGF, Hif-1α, BCL2, RET,* and *RAS* promoters have been characterized by biophysical and structural methods, most often NMR.[11,13-17] One oncogene of great pharmaceutical interest is *HRAS*, which belongs to the *RAS* family of proto-oncogenes that also includes *KRAS* and *NRAS*. *RAS* is the most commonly mutated gene family in cancer; about 19% of the US cancer patients carry *RAS* mutations.[18,19] *HRAS*, which codes for a small GTPase, lies upstream of multiple cell proliferation pathways.[20] Mutations in the HRAS protein can cause constitutive activation of downstream pathways, leading to uncontrolled cell growth.[21] Such mutations are present in bladder cancers (6%) and a subset of head and neck squamous cell carcinoma (5%).[18] Targeting of RAS proteins with small-molecule drugs is a challenge, however, because they lack deep pockets for inhibitor binding.[22] Only recently has there been success in targeting KRAS protein with covalent ligands, exemplified by the FDA approval of sotorasib for metastatic non-small-cell lung cancer in 2021.[23] There are no therapies for *HRAS*-driven cancers. An alternative strategy to targeting HRAS protein is to inhibit *HRAS* gene expression. Therefore, there is great interest in understanding the regulation of the *HRAS* oncogene and the nature of non-canonical DNA structures possibly involved in such regulation.

In 2015, two sequences with iM-forming potential, *hras-1* and *hras*-2, were characterized in the *HRAS* promoter region immediately upstream of the major transcription start site.[11] These sequences are complementary to two previously characterized G-rich sequences that adopt another non-canonical DNA structure called the G-quadruplex. Formation of G-quadruplex structures in the *HRAS* gene was suggested to repress transcription of *HRAS*.[24] Both C-rich sequences fold *in vitro* into stable iMs under slightly acidic conditions. In their folded forms, *hras-1* and *hras*-2 reportedly regulate the expression of *HRAS* by interacting with hnRNP A1, an abundant nuclear proteins that regulates mRNA biogenesis.[11] It has been suggested that hnRNP A1 binds to the lateral loops of the iM.[11] Drug screening studies using *hras-1* yielded two small molecules that bind to *hras-1* iM with sub-micromolar affinity and high selectivity.[25] To evaluate potential ligand binding pocket in the absence of *hras-1* structure, molecular docking studies were done using homology models. Due to complexity of modeling loops, the molecular basis for ligand recognition remains unknown. Any further development of these ligands or discovery of new scaffolds would be greatly aided by a high-resolution *hras-1* structure.

Atomic-resolution structural information on iMs is extremely scarce. Analysis of the Protein Data Bank (PDB)[26] indicates that the first iM structure was deposited in 1993.[27] Overall, the PDB contains 27 NMR and 10 crystal structures of unique iMs (i.e., unique sequence or

conformation). Of these 37 structures, 23 are tetrameric (14 are NMR and nine are crystal structures), seven are dimeric (six are NMR and one is a crystal structure), and seven are monomeric (all are NMR structures). Monomeric structures (PDB ID 8BV6, 8BQY,[28] 7O5E,[29] 1A83,[30] 1G22,[31] 5OGA,[32] and 1ELN/1EL2[33]) are formed by either human telomere and centromere fragments or by designed sequences that fold into minimal iM or iM/B-DNA junctions. Thus, to date, there are no structures of biologically relevant iMs, whether from oncogene promoter regions or unmodified telomeric or centromeric DNA, which emphasizes the need for new structural information on iMs.

In this study, we present the first crystal structure of an iM formed by a C-rich region within the *HRAS* promoter, 5'-CGCCCGTGCCCTGCGCCCGCAACCCGA-3' (here referred to as iHRAS). The structure was solved to a 1.77 Å resolution and is an antiparallel dimer of two iHRAS strands, which leads to two covalently connected nearly identical iMs. Each iM contains six C-$C^+$ base pairs further stabilized by stacking with a G-G homopurine base pair and a cytosine. The loops and connecting region between the two iMs have extensive canonical and non-canonical base pairing and stacking. The entire assembly is highly compact. The iHRAS structure greatly advances our understanding of the iM topology and of the roles that loops and connecting regions play in its folding and stabilization.

## Results and Discussion

### Biophysical characterization of Br-iHRAS

In this work, we aimed to solve the crystal structure of iHRAS C-rich DNA sequence from the *HRAS* oncogene promoter. Initial crystallization trials resulted in diffraction-quality crystals; however, solving the phase problem was a barrier to completing the structure. To overcome this barrier, we designed five variants of iHRAS in which one or two of the cytosines were replaced with 5-bromocytosine (5Br-C) (Table S1). The geometry of the 5Br-C base is such that the base modification should not interfere with iM formation.

To demonstrate that introduction of brominated nucleobases did not broadly affect the iM structure of iHRAS we performed biophysical characterization of iHRAS and Br-iHRAS mutants. The CD spectrum of iHRAS at pH 6.0 has a characteristic peak at 288 nm (Figure S1A), and the spectrum is identical to the CD signature reported previously.[11] The Br-iHRAS mutants have CD signatures similar to that of iHRAS but with somewhat lower intensity and, in some cases, with the main peak at a slightly shorter wavelength between 282 and 285 nm. Br-iHRAS2, with 5Br-C at position 4, has a CD signature that resembles closely that of iHRAS albeit with lower intensity.

Thermal difference spectra (TDS) of all sequences contain a negative peak at approximately 304 nm (Figure S1B) indicative of iM formation. Based on a previous study, this negative peak was expected to occur at approximately 295 nm.[34] Why it is redshifted by 9 nm is not entirely clear. We speculate that the red shift may be due to formation of non-canonical base pairs in addition to the expected C-$C^+$ pairs. We have observed a similar location of TDS negative peak for other iM structures studied in our laboratory. iHRAS displays the highest intensity of the 304 nm feature in TDS, followed by Br-iHRAS1 and Br-iHRAS2. For other mutants, the negative peal at 304 nm is weak, suggesting a lower extent of iM folding.

Analysis using polyacrylamide gel electrophoresis (PAGE) revealed that iHRAS and all Br-iHRAS mutants share a major band that migrates below the T24 marker, corresponding to a monomolecular species, and a fainter dimer band above the T30 marker (Figure S1C).

The melting temperature ($T_m$) of iHRAS is 48 ± 2 °C in 10 mM MES pH 6.0 buffer. This value is within experimental error of the melting temperature of 47.6 °C determined in 50 mM sodium cacodylate pH 6.0, 50 mM KCl buffer reported previously.[11] The thermal stabilities of Br-iHRAS mutants were similar with all $T_m$ values in the range of 44-48 °C (Figure S1D). However, melting transitions of brominated mutants were less well defined, and the resulting $T_m$ values have higher associated errors.

In sum, all brominated mutants behave similarly to iHRAS with Br-iHRAS2 displaying the greatest similarity. We attempted crystallization of all mutants but obtained diffraction-quality crystals only for Br-iHRAS2. We solved the phase problem using the single-wavelength anomalous diffraction method, relying on the anomalous bromine signal.

## Overall architecture of iHRAS iM

We solved the crystal structure of native iHRAS to a resolution of 2.02 Å (PDB ID 8DHC) and Br-iHRAS2 mutant to a resolution of 1.77 Å (PDB ID 8CXF). Alignment of the native and brominated structures yielded an RMSD of 0.227 Å, indicating that structures are nearly identical. The rest of this report will only discuss the higher resolution Br-iHRAS2 structure unless otherwise mentioned.

Contrary to our expectation of a monomolecular iM, we observed a tail-to-tail dimer of two 3'E iMs each with six C-$C^+$ base pairs. In the 3'E iMs, the outermost C-$C^+$ pair is at the 3' end. 3'E iMs are more common than are 5'E iMs as they are thermodynamically more stable.[35] Each DNA strand (labeled here A and B) forms two hairpins, one at the 5' end and another at the 3' end. Such folding results in two loops, one at the 'head' of each hairpin, bringing the 5' and 3' ends near the center of the molecule. Strands A and B associate in an antiparallel fashion to form a dimer, which contains two intermolecular iMs linked by a connecting region (Figure 1 and Figure S2). Nucleotides C1-G13 from chain A and C14'-G26' from chain B (nucleotides of chain B are designated by ') form the first iM, iM-1; C1'-G13' and C14-G26 form the second iM, iM-2. Alignment of iM-1 and iM-2 revealed that they are nearly identical with an RMSD of 0.125 Å. Thus, in the discussion that follows the values for the two iMs were averaged.

The structure contains 20 base pairs, and only two of them are canonical. There are two Watson-Crick G-C pairs, two Hoogsteen G-$C^+$ pairs, twelve C-$C^+$ pairs, three homopurine G-G pairs, and one homopyrimidine T-T pair (Figure 2). The overall structure of iHRAS can be divided into three components: the iM cores, the loops (GTG and GCAA), and the connecting region formed by the 5'-CG, the 3'-G overhang, and the middle TGCG stretch (Figure 1C).

Each iM core contains six intercalated C-$C^+$ pairs: [C3-C16']$^+$, [5Br-C4-C17']$^+$, and [C5-C18']$^+$ form one stack, and [C9-C23']$^+$, [C10-C24']$^+$, and [C11-C25']$^+$ intercalate between them. All cytosines adopt the *anti* glycosidic conformation as is commonly observed for iM

structures. C-$C^+$ base pairs show little shearing, stretching, or staggering. The buckle and propeller twist angles are also small (less than 10 ° for all but propeller value for $[C3\text{-}C16]^+$, Table S2) indicating that C-$C^+$ base pairs are nearly planar. The pair that involves the 5Br-C at C4 has the same geometric parameters as a typical unmodified C-$C^+$ base pair because the Br atom points out, avoiding steric clashes.

Each folded strand forms a minor groove with an average P-P distance of 7.0 ± 0.9 Å (Table S3A). The minor groove is narrower at the center of each iM, with an average P-P distance of 6.6 ± 0.5 Å, and wider at the ends of each iM (both at the loop side and in the connecting region side), with the average P-P distance of 8.4 ± 0.1 Å (Table S3). This widening likely occurs to accommodate the specific geometries of the adenine stack (see more on the adenine stack below) in the loops and of the sharp turns into the connecting region. Association of two DNA strands results in the formation of a major groove with an average width of 13.5 ± 1.6 Å (Table S3B). The average dimensions of the minor groove in other representative iM structures are between 6.3 and 9.1 Å, and the average dimension of the major grooves varies between 12.8 and 16.6 Å.[30,33,36-39] The values observed here thus fall within the lower to middle portion of these ranges. Sugar-phosphates between successive cytosines have an average helical rise of 6.4 ± 0.9 Å (Table S4A), allowing room for an average stacking interval between intercalated pairs of 3.2 ± 0.2 Å (Table S4B). This distance is similar to 3.1 Å observed in iM structures formed by dCCCT (PDB ID 191D),[38] dCCCC (PDB ID 190D),[37] and $d(CCCTA_2)_3CCCT$ (PDB ID 1ELN/1EL2)[33] but shorter than a typical distance of 3.4 Å reported for double stranded DNAs or G-quadruplexes where the base pairs or G-tetrads, respectively, π stack. In an iM, the C-$C^+$ base pairs stack in such fashion that only exocyclic atoms O2 and N4 overlap but the bases do not; this results in a narrower spacing between base pairs. The helical twist in our structure is 22 ± 4° (Table S5), although the distribution is bimodal with average values of the two peaks at 26.8 ± 0.4 and 19.3 ± 1.7°. The reported helical twists in other structures of iMs range between 10° and 23° (PDB ID 191D, 1C11, 190D, 294D, 1YBL, 1YBR, 1YBR, 1YBN, 1A83). However, in the iM structure formed by CCG triplet repeats (PDB ID 4PZQ), the helical twist is higher, 30°.[39] It is possible that helical twists were measured differently in different reports, and thus the numbers may be more consistent than they appear.

Stacking on top of the terminal C5-C18’ base pair is a parallel, sheared homopurine G6-G19’ base pair (Figure 1C). The stacking interval between the G-G and the C-$C^+$ pairs is 3.0 Å, matching the interval between intercalated cytosine pairs. This G-G pair is maintained by reciprocal N2-N3 hydrogen bonds (Figure 2) and is characterized by a large propeller twist of 36.3° and a buckle angle of 14.6° (Table S2), thus, the two bases are not coplanar. In addition, the C-G steps leading to the G6-G19’ base pair have large helical twists of 56.5° for C5-G6 and 52.3° for C18’-G19’ (Table S4).

The C-$C^+$ capping interactions observed in iHRAS is similar to those seen in other iM structures.[28,29,32,39-41] For example, the iM core formed by CCG triplet repeats (PDB ID 4PZQ) is flanked by one G-G base pair at each end.[39] As is the case for iHRAS, that structure displays a large propeller twist (23° and 39°) and a large buckle angle (16° and 27°) for the two G-G base pairs as well as large helical twists for the C-G steps (56° and 53°).[39] There are iM cores that are extended by minor groove slipped tetrads

such as G-C-G-C,[28] G-T-G-T[28,29,40,41], and G-C-G-T[32]. In all these cases guanines are engaged in reciprocal N2-N3 hydrogen bonds closely resembling G-G homopurine base pair observed in iHRAS structure including its high propeller twist and buckle angles. For example, the C-rich strand of the human centromeric region (PDB ID 1C11) has an iM core capped by a G-T-G-T minor groove tetrad with G-G propeller twist of 31° and buckle angel of 11°.[40] In addition to the G-G base pair, T-T,[29,30,33,38] A-A,[42] and T-A[30] base pairs and an A-A-T base triple [36] have been shown to cap or extend iM cores.

Finally, the C-$C^+$/G-G stack in iHRAS is capped by C20' from the GCAA lateral loop (Figure 3A). This nucleotide stacks onto G6 and participates in several other stabilizing interactions. N4 of C20' is involved in a bifurcated hydrogen bond with a phosphate of T7 and O6 of G19* (* denotes a symmetry-generated molecule). In addition, O4' of its sugar interacts with N1 of G6.

Capping interactions have been suggested to provide thermal and pH stabilization of iM structures, which is best demonstrated in the case of so-called minimal iMs. In minimal iMs, the iM core is formed by only two C-$C^+$ base pairs, yet the structures are stable because the short iM core is capped by minor groove slipped tetrads at both ends.[28,32,41]

iHRAS structure has an average B-factor of 58.2 $Å^2$. Five nucleotides (T7, G8, and A21 from lateral loops and G15 and G26 from the connecting region) display B-values higher than 60 $Å^2$, Figure S3. B-factors for native iHRAS and Br-iHRAS2 are rather similar with some local variations.

## Conformations of lateral loops

There are two loops, G6-T7-G8 and G19-C20-A21-A22, at each end of the iHRAS dimer which cap each narrow groove (Figure 1C). In addition to G6-G19 base pairing and C20 stacking, loops are stabilized by the intermolecular contacts with nucleotides from other asymmetric units. Specifically, G8 from the GTG loop π-stacks with G8*. T7 projects outward and is the only residue in the molecule that does not engage in any contacts. As such, T7 has a high B-factor of 78.2 $Å^2$ (average B-factor for DNA is 58.2 $Å^2$). A21 and A22 from the GCAA loop project away from the molecule nearly perpendicular to the iM core. Their bases π-stack with each other and with A21*/A22* (Figure 3B). The twist angle of 37.1° between A21 and A22 (Table S3) leads to nearly perfect pairwise overlap of the 6-membered rings, partial overlap of the 5-membered rings, and an overall right-handed helical arrangement of the adenine stack. The A21-A22-A22*-A21* π-stacks interact with each other via a polyethylene glycol from the crystallization solution contributing to the 3D arrangement of the crystal lattice (Figure S4A). Similar, although not identical, adenine clusters were observed in the crystal structure of dAACCCC tetramolecular iM from *Tetrahymena thermophila* telomeric DNA (PDB ID 294D).[43]

## Architecture of the connecting region

The two iM cores in the iHRAS dimer are connected via a four nucleotide T12-G13-C14-G15 linker. In addition, the 5'-C1-G2 and 3'-G26 overhangs are located near the connecting region (Figure S2). The terminal hydroxides of the 5' ends of chains A and B are close enough to form a hydrogen bond; however, due to a lack of electron density, we did not

include the 5’ terminal hydroxides in our model. The 3’ ends of the DNA strands are at the opposite sides of the connecting region and point away from the iM core and from each other (Figure S2). The 3’ end is the most disordered region in iHRAS where G26 has a high B-value of 91.2 $Å^2$ and A27 has poor electron density and was not modeled.

The TGCG linker nucleotides are engaged in extensive base pairing and stacking interactions, running roughly perpendicular to the iM core, and involving chains A and B as well as the symmetry partners. At the center are two G-C pairs, Hoogsteen C1-G15’ and Watson-Crick G2-C14’ (Figure 3C). The Hoogsteen C1-G15’ base pair is maintained by two hydrogen bonds between N3 and N4 of C1 and N7 and O6 of an *anti* oriented G15’ (Figure 2). The C1-G15’ base pair is non-planar and displays an approximately −7° propeller twist as well as an approximately 11° buckle (Table S2). The base pair is further stabilized by the stacking between G15’ and a *syn*-oriented G26’ as well as by the hydrogen bond between N4 of C1 and N7 of G2’. G2 adopts a *syn* glycosidic conformation as expected for the Watson-Crick base pairing. The G2-C14’ base pair is also non-planar and displays a 15.8° propeller twist (Table S2). Overall, the stacking pattern consists of G26’/C1-G15’/G2-C14’. An identical arrangement of bases is repeated at the base of iM-2 with the stack of bases running nearly parallel to the stack described above (Figure S2B).

The two stacks, G26’/C1-G15’/G2-C14’ and G26/C1’-G15/G2’, are capped by T12 and T12’, respectively, such that each T stacks on top of each G2-C14 base pair with virtually no π-system overlap (Figure 3D). T12 and T12’ form a homopyrimidine base pair via N3-O2 reciprocal hydrogen bonds (Figure 2). The T12-T12’ base pair is stabilized by stacking with the G13-G13’ base pair such that the T and G bases are engaged in extensive π-π interactions (Figure 3E). T12-T12’ and G13-G13’ base pairs repeat only once in the iHRAS dimer, unlike all other base pairs that repeat twice. The T12-T12’ base pair has a strong propeller twist of 32.4°, whereas G13-G13’ is nearly planar (propeller twist of 6.1°). Due to its planarity, the G13-G13’ base pair stacks neatly with the same pair from the symmetry generated set of molecules connecting two dimers to each other and contributing to 3D crystal packing (Figure S4B).

## Water networks

The Br-iHRAS2 structure contains 74 defined water molecules, whereas the structure of native iHRAS contains seven waters due to its lower resolution (2.02 vs. 1.77 Å). Five waters are common to both structures suggesting that water networks are likely similar.

Waters form fragmented networks across the major grooves and around the loops (Figure 4). The minor grooves are too narrow (average P-P distance of 7.0 Å) to harbor water molecules, an observation made for other crystallized iMs. The connecting region is tightly packed and connects to water molecules only on its outskirts. In the major groove, water molecules hydrogen bond to phosphates, to N4 atoms of cytosines, or to each other to form a groove-spanning network (Figure 4A). It is worth noting that N4 atoms of most cytosines are engaged in a bifurcated hydrogen bond to O2 of a cytosine partner and either a water molecule or a phosphate (Table S6). This bifurcated hydrogen bond is common in other iMs and plays an important role in stabilization of the iM core. Upon careful examination, N4 atoms of five cytosines in chain A and eight cytosines in chain B interact with one water

molecule each, whereas three cytosines in each chain form hydrogen bonds to a phosphate from the opposite chain. Only one cytosine in chain B (C17') and four cytosines in chain A do not form hydrogen bonds with waters or phosphates, although weak electron density is detected at the expected distances for some of them. A similar water network is observed in the high-resolution crystal structures of dTCACCC (1.85 Å, PDB ID 200D)[44] and dCCCT (1.4 Å, PDB ID 191D)[38], which form tetramolecular iMs.

In the loops of iHRAS, water molecules provide backbone stabilization connecting phosphates to each other and to nearby bases. There are two water molecules with unusually high electron density (one is shown in Figure 4B and is marked with asterisk). This water is in a penta-coordinated environment of N2 and N3 of G8, N1 and N2 of G19' and another water molecule. The second water is found in the loop of iM-2 in a nearly identical environment. Similar water molecules are present in the native iHRAS structure. We considered ions ($K^+$, $Na^+$, $Mg^{2+}$, or $Cl^-$) for this position; however, none were good choices based on the value of B-factor, charge, or preferred type and geometry of ligand environment (see Materials and Methods in SI).

## Oligomeric state of iHRAS

PAGE results indicated that under dilute conditions (50 μM DNA, 10 mM MES 6.0, 10 mM KCl) iHRAS primarily exists in a monomeric form (Figure S1C). In contrast, the crystal structure revealed an iHRAS dimer (Figure 1). We set out to determine whether it is the high concentration of DNA (1 mM) in the crystallization sample, the components of the crystallization solution, or crystal packing forces that drive the dimer formation. We prepared samples of iHRAS with concentrations ranging from 5 μM to 2 mM and analyzed them using PAGE. Regardless of the concentration, the major band on the PAGE was that of a monomer, although a prominent dimer band was seen in the 2 mM samples (Figure S5A). Moreover, neither the CD signature nor the thermal stability changed with concentration (Figure S5B-C). These data indicate that concentration alone does not drive the dimerization process, at least not at concentrations below 1 mM. Next, we performed PAGE on samples prepared under the crystallization condition and on crystals; significant dimerization was not detected (Figure S6). A sample of the Br-iHRAS2 crystals, on the other hand, did run at the position of the dimer band, suggesting that it is crystal packing forces that lead to iHRAS dimerization. Collectively, our data suggest that although iHRAS exists predominantly in a monomeric form in solution, it has an intrinsic ability to dimerize.

To understand the monomer-dimer equilibrium we built a model for the iHRAS monomer by removing iM-2 (residues C14-G26 and C1'-G13'), connecting residues G13 and C14' (which were found in close proximity), and regularizing the region between residues T12 and C14. The resulting structure (Figure 5A, right) has small deviations in the positions of T12 and C14 bases and new location of the G13 base compared to the iHRAS structure. The only interactions that are lost in the monomeric iM are T12-T12' and G13-G13' base pairing. Therefore, our X-ray-determined iHRAS structure, although dimeric, is an excellent model for a monomeric iHRAS.

It is straightforward to envision the pathway from the iHRAS dimer to the monomer. First, the dimer must dissociate into its constituents, chains A and B, which adopt a double

hairpin structure connected via the TGCG connecting region (Figure 5A, center). Next, the individual chains must fold intramolecularly converting the connecting TGCG region into loop 2 and the two original loops to loop 1 (GTG) and loop 3 (GCAA). The schematic of the proposed monomolecular iHRAS structure is shown in Figure 5B.

### Biological relevance of iHRAS dimer

In dilute conditions, iHRAS exists primarily as a monomer (here and [11]). Since the 27-nucleotide sequence of iHRAS appears once in the human reference genome (hg38),[45] the region is predicted to adopt a monomeric fold *in vivo*. The dimeric crystal structure of iHRAS provides us with a robust understanding of the folding of monomeric iHRAS. However, we do not want to completely dismiss the possibility of a biologically relevant dimer in the *HRAS* gene. Downstream of the iHRAS sequence is a second characterized iM, *hras-2*,[11] which contains four stretches of four or five cytosines connected by a single guanine, CCCCCGCCCCCGCCCCGCCCC (C-stretches are underlined). The C-rich nature of *hras-2* and reported biophysical studies[11] suggest that it folds into a stable iM. The high-resolution structure of *hras-2* has not been determined, although structural studies are underway in our laboratory. It was originally suggested that iMs formed by *hras-1* (here iHRAS) and *hras-2* independently inhibit *HRAS* transcription,[11] but we speculate that iHRAS and *hras-2* may also function in tandem through the formation of a heterodimer (Figure S7). C-rich stretches in *hras-2* are longer than in iHRAS increasing the probability of dimer formation because pairing of three cytosines in each stretch of iHRAS can happen with any of the four or five cytosines in each stretch of *hras-2*. Heterodimer formation could occur when sufficient negative superhelicity results in melting of the region between iHRAS and *hras-2*. In addition, the 46-nucleotide linker between the 3'-A in iHRAS and the first C-stretch in *hras-2,* may adopt a secondary structure with two or three hairpins, one of which is predicted to have a stem of six G-C base pairs (Figure S7C). Such folding would bring the two C-rich sequences into proximity increasing the probability of heterodimer formation. Cooperative folding of nearby secondary structure elements has been demonstrated in other oncogenes including in the human *TERT* promoter where a 31-base pair hairpin is suggested to initiate cooperative folding of two nearby G-quadruplexes.[46] Biologically relevant dimeric iMs are known. For example, dimeric iMs formed by the C-rich DNA sequences from centromeric regions of human and *Drosophila melanogaster* chromosome have been suggested to play important roles in nucleosome organization.[40,47,48] Dimerization of iMs at the entrance and exit of nucleosomes are predicted to allow precise positioning of nucleosomes in the lateral direction. Further support for biologically relevant dimerization comes from a recent study that identified the first human protein with an exceptional selectivity for heterodimeric G-quadruplexes lending credibility to their *in vivo* existence.[49] If heterodimeric G-quadruplexes have biological importance, so could heterodimeric iMs.

## Conclusion

We report the first crystal structure of iHRAS iM formed by the 27-nucleotide DNA sequence from the *HRAS* oncogene promoter. Previously, only structures of human telomeric or centromeric iM-forming regions had been determined at high resolution by NMR or X-ray crystallography, and all either contain multiple strands (i.e., are

tetramolecular[42,43] or bimolecular[40]) or have base modifications (e.g., 5-methyl cytosine or uridine)[30,33] to improve sample homogeneity.

In our structure, one iHRAS strand folds into a double hairpin connected by the TGCG connecting region. The two strands associate in an antiparallel fashion, forming a dimer with 5' and 3' ends in close proximity in the middle of the structure between the two iM cores. The dimer contains two nearly identical 3'E iMs with six C-$C^+$ intercalated base pairs. The iM core is extended by a homopurine G-G base pair and a C that stacks on top. The two iMs are linked by a connecting region that forms many canonical and non-canonical base pairing and base stacking interactions. Overall, every observed base in the iHRAS structure except T7 is engaged in extensive hydrogen bonding and/or π-stacking interactions leading to a very compact elongated structure. These interactions stabilize loops and the connecting region. The extensive networks of interaction that stabilize the iM core in our crystal structure are unprecedented. Our structure provides structural evidence supporting the idea that auxiliary base pairing is an essential aspect of iM stability *in vivo*, in addition to molecular crowding, negative superhelicity, interactions with ligands, and other conditions.

Importantly, the structural features of the loops and connecting region are unique structural elements that can be targeted by small-molecule ligands or proteins as potential cancer therapies. For example, binding of nuclear protein hnRNP A1 to *hras-1* and *hras-2* was suggested to occur through the lateral loops.[11] Such features will allow design of highly gene-selective ligands and will be useful for discovery of highly selective proteins.

Others[11,24] and we have shown that DNA from the *HRAS* promoter may adopt a number of non-canonical secondary structures. Each of the folded states of the *HRAS* promoter (i.e., G-quadruplex formed by the G-rich strand, the iMs formed by *hras-1* and *hras-2*, and an *hras-1/hras-2* heterodimer) may correspond to a different state of *HRAS* oncogene regulation thus allowing for very precise control of *HRAS* expression. The first crystal structure of an iM from a human oncogene promoter presented in this work thus advances our understanding of iM targeting, gene regulation, and other biological functions of iMs. In addition, it will help elucidate the mode of binding of existing iM ligands and will facilitate design of new scaffolds that selectively bind to iHRAS and iMs in general.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

The authors would like to thank Dr. Kay Perry, a Staff Scientist at NE-CAT beamline at APS for her help and advice with data processing and structure refinement. We would like to thank Prof. Hurley (University of Arizona) for the fruitful discussions. We would like to thank our research assistant Kailey Martin for her help with repeating some of the biophysical experiments. Finally, we would like to thank our research student David (Ming) Ye for his help with figures and some data analysis.

## References

[1]. Abou Assi H, Garavís M, González C, Damha MJ, Nucleic Acids Res. 2018, 46, 8038–8056. [PubMed: 30124962]

[2]. Guéron M, Leroy J-L, Curr. Opin. Struct. Biol 2000, 10, 326–331. [PubMed: 10851195]
[3]. Day HA, Pavlou P, Waller ZAE, Bioorg. Med. Chem 2014, 22, 4407–4418. [PubMed: 24957878]
[4]. Assi HA, Harkness RW, Martin-Pintado N, Wilds CJ, Campos-Olivas R, Mittermaier AK, González C, Damha MJ, Nucleic Acids Res. 2016, 44, 4998–5009. [PubMed: 27166371]
[5]. Rajendran A, Nakano S, Sugimoto N, Chem. Commun 2010, 46, 1299–1301.
[6]. Cui J, Waltman P, Le HV, Lewis AE, Molecules 2013, 18, 12751–12767. [PubMed: 24132198]
[7]. Yu. Fedoroff O, Rangan A, Chemeris VV, Hurley LH, Biochemistry 2000, 39, 15083–15090. [PubMed: 11106486]
[8]. Sun D, Hurley LH, J. Med. Chem 2009, 52, 2863–2874. [PubMed: 19385599]
[9]. Wright EP, Huppert JL, Waller ZAE, Nucleic Acids Res. 2017, 45, 2951–2959. [PubMed: 28180276]
[10]. Zeraati M, Langley DB, Schofield P, Moye AL, Rouet R, Hughes WE, Bryan TM, Dinger ME, Christ D, Nat. Chem 2018, 10, 631–637. [PubMed: 29686376]
[11]. Miglietta G, Cogoi S, Pedersen EB, Xodo LE, Sci Rep 2015, 5, 18097. [PubMed: 26674223]
[12]. Kang H-J, Kendrick S, Hecht SM, Hurley LH, J. Am. Chem. Soc 2014, 136, 4172–4185. [PubMed: 24559432]
[13]. Kaiser CE, Van Ert NA, Agrawal P, Chawla R, Yang D, Hurley LH, J. Am. Chem. Soc 2017, 139, 8522–8536. [PubMed: 28570076]
[14]. Guo K, Gokhale V, Hurley LH, Sun D, Nucleic Acids Res. 2008, 36, 4598–4608. [PubMed: 18614607]
[15]. Dai J, Hatzakis E, Hurley LH, Yang D, PloS One 2010, 5, e11647. [PubMed: 20657837]
[16]. Guo K, Pourpak A, Beetz-Rogers K, Gokhale V, Sun D, Hurley LH, J. Am. Chem. Soc 2007, 129, 10220–10228. [PubMed: 17672459]
[17]. Brazier JA, Shah A, Brown GD, Chem. Commun 2012, 48, 10739–10741.
[18]. Moore AR, Rosenberg SC, McCormick F, Malek S, Nat. Rev. Drug Discov 2020, 19, 533–552. [PubMed: 32528145]
[19]. Prior IA, Hood FE, Hartley JL, Cancer Res. 2020, 80, 2969–2974. [PubMed: 32209560]
[20]. Aoki Y, Niihori T, Kawame H, Kurosawa K, Ohashi H, Tanaka Y, Filocamo M, Kato K, Suzuki Y, Kure S, Matsubara Y, Nat. Genet 2005, 37, 1038–1040. [PubMed: 16170316]
[21]. Simanshu DK, Nissley DV, McCormick F, Cell 2017, 170, 17–33. [PubMed: 28666118]
[22]. O'Bryan JP, Pharmacol. Res 2019, 139, 503–511. [PubMed: 30366101]
[23]. Skoulidis F, Li BT, Dy GK, Price TJ, Falchook GS, Wolf J, Italiano A, Schuler M, Borghaei H, Barlesi F, Kato T, Curioni-Fontecedro A, Sacher A, Spira A, Ramalingam SS, Takahashi T, Besse B, Anderson A, Ang A, Tran Q, Mather O, Henary H, Ngarmchamnanrith G, Friberg G, Velcheti V, Govindan R, Engl N. J. Med 2021, 384, 2371–2381.
[24]. Cogoi S, Shchekotikhin AE, Xodo LE, Nucleic Acids Res. 2014, 42, 8379–8388. [PubMed: 25013182]
[25]. Journey SN, Alden SL, Hewitt WM, Peach ML, Nicklaus MC, Schneekloth JS Jr, Med. Chem. Commun 2018, 9, 2000–2007.
[26]. Berman H, Henrick K, Nakamura H, Nat. Struct. Mol. Biol 2003, 10, 980–980.
[27]. Gehring K, Leroy J-L, Guéron M, Nature 1993, 363, 561–565. [PubMed: 8389423]
[28]. Serrano-Chacón I, Mir B, Cupellini L, Colizzi F, Orozco M, Escaja N, González C, J. Am. Chem. Soc 2023, 145, 3696–3705. [PubMed: 36745195]
[29]. Serrano-Chacón I, Mir B, Escaja N, González C, J. Am. Chem. Soc 2021, 143, 12919–12923. [PubMed: 34370473]
[30]. Han X, Leroy J-L, Guéron M, J. Mol. Biol 1998, 278, 949–965. [PubMed: 9600855]
[31]. Nonin-Lecomte S, Leroy JL, J. Mol. Biol 2001, 309, 491–506. [PubMed: 11371167]
[32]. Mir B, Serrano I, Buitrago D, Orozco M, Escaja N, González C, J. Am. Chem. Soc 2017, 139, 13985–13988. [PubMed: 28933543]
[33]. Phan AT, Guéron M, Leroy JL, J. Mol. Biol 2000, 299, 123–144. [PubMed: 10860727]
[34]. Mergny J-L, Li J, Lacroix L, Amrane S, Chaires JB, Nucleic Acids Res. 2005, 33, e138. [PubMed: 16157860]

[35]. Malliavin TE, Gau J, Snoussi K, Leroy J-L, Biophys. J 2003, 84, 3838–3847. [PubMed: 12770889]

[36]. Weil J, Min T, Yang C, Wang S, Sutherland C, Sinha N, Kang C, Acta Crystallogr. D Biol. Crystallogr 1999, 55, 422–429. [PubMed: 10089350]

[37]. Chen L, Cai L, Zhang X, Rich A, Biochemistry 1994, 33, 13540–13546. [PubMed: 7947764]

[38]. Kang C, Berger I, Rich A, Proc. Natl. Acad. Sci 1994, 91, 11636–11640. [PubMed: 7972115]

[39]. Chen Y-W, Jhan C-R, Neidle S, Hou M-H, Angew. Chem. Int. Ed 2014, 53, 10682–10686.

[40]. Gallego J, Chou S-H, Reid BR, J. Mol. Biol 1997, 273, 840–856. [PubMed: 9367776]

[41]. Escaja N, Viladoms J, Garavís M, Villasante A, Pedroso E, González C, Nucleic Acids Res. 2012, 40, 11737–11747. [PubMed: 23042679]

[42]. Esmaili N, Leroy JL, Nucleic Acids Res. 2005, 33, 213–224. [PubMed: 15647504]

[43]. Cai L, Chen L, Raghavan S, Ratliff R, Moyzis R, Rich A, Nucleic Acids Res. 1998, 26, 4696–4705. [PubMed: 9753739]

[44]. Kang C, Berger I, Loskshin C, Ratliff R, Moyzis R, Rich A, Proc. Natl. Acad. Sci 1995, 92, 3874–3878. [PubMed: 7731999]

[45]. Kent WJ, Genome Res. 2002, 12, 656–664. [PubMed: 11932250]

[46]. Song JH, Kang H-J, Luevano LA, Gokhale V, Wu K, Pandey R, Sherry Chow H-H, Hurley LH, Kraft AS, Cell Chem. Biol 2019, 26, 1110–1121.e1-e4. [PubMed: 31155510]

[47]. Garavís M, Escaja N, Gabelica V, Villasante A, González C, Chem. – Eur. J 2015, 21, 9816–9824. [PubMed: 26013031]

[48]. Garavís M, Méndez-Lago M, Gabelica V, Whitehead SL, González C, Villasante A, Sci. Rep 2015, 5, 13307. [PubMed: 26289671]

[49]. Liano D, Monti L, Chowdhury S, Raguseo F, Di Antonio M, Chem. Commun 2022, 58, 12753–12762.

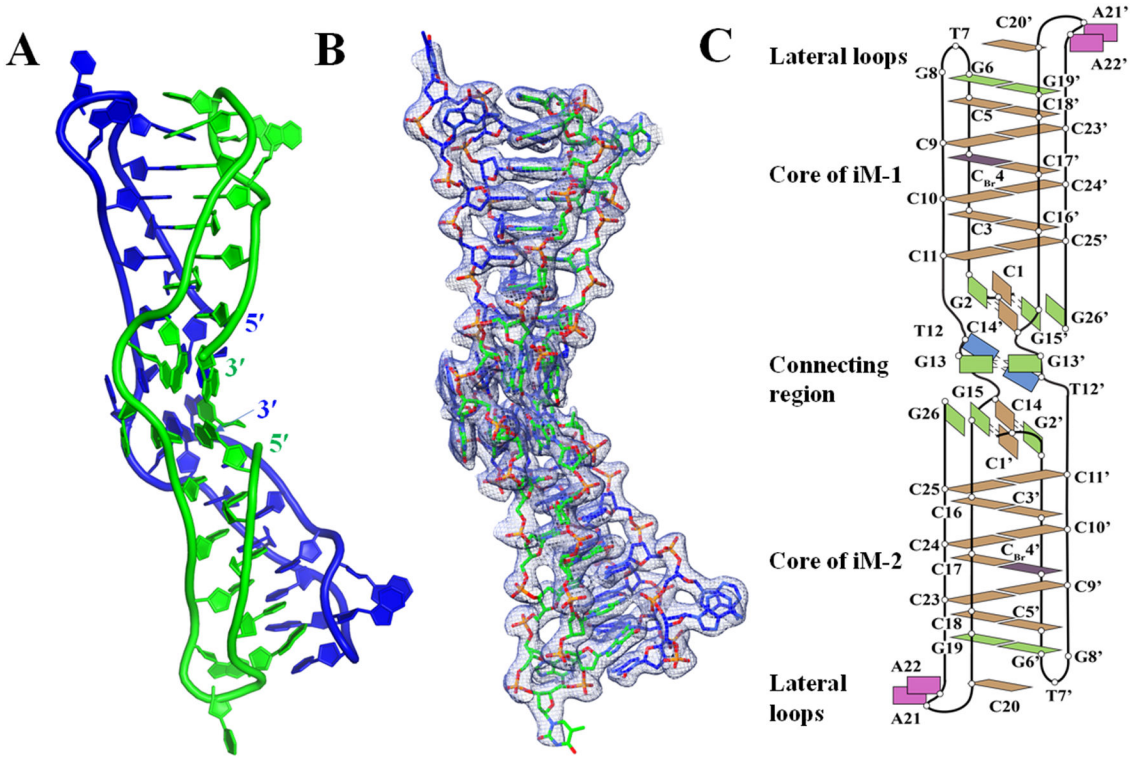



**Figure 1.**

Structure of the dimeric iHRAS iM. **A)** Cartoon representation with purines, pyrimidines, and sugars shown as filled rings. Chains A and B are colored green and blue, respectively. **B)** iHRAS structure surrounded by the electron density at I/σ = 1.0. **C)** iHRAS schematics with nucleotide numbering. Nucleotides are colored by base: thymines are blue, guanines are green, cytosines are brown, adenines are magenta, and the 5Br-C located at position 4 is colored dark brown. Nucleotides from chain B are marked with an apostrophe (').

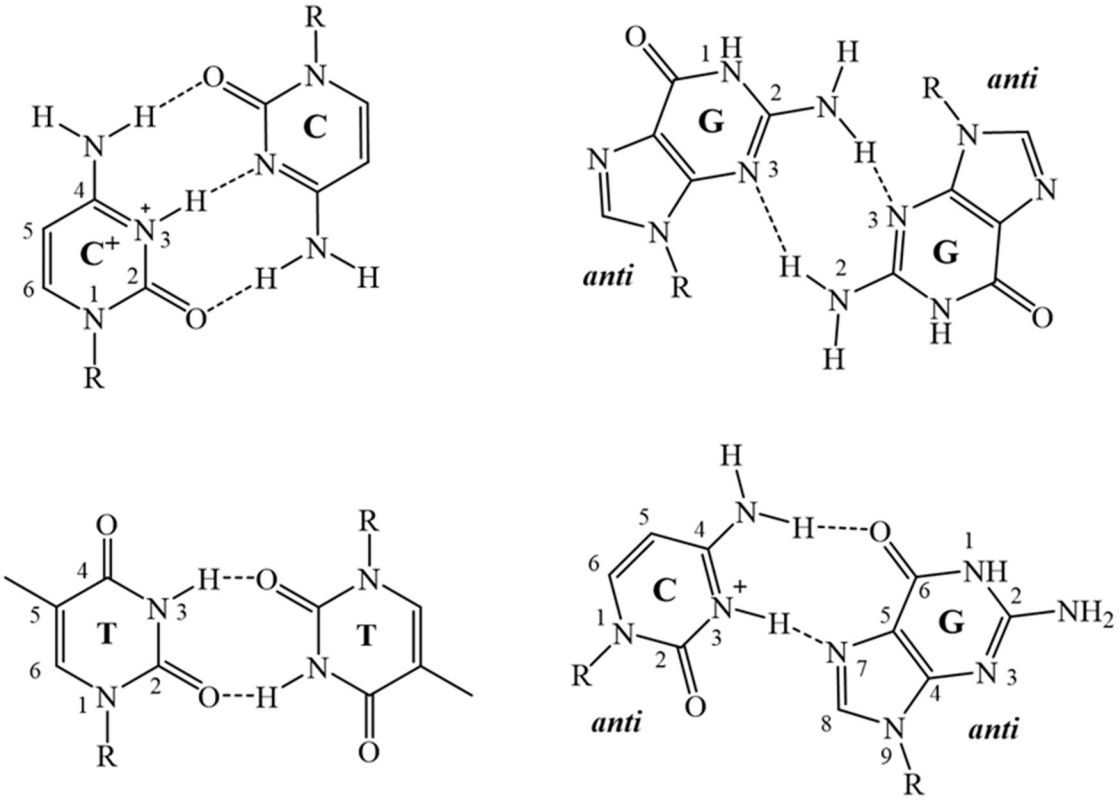



**Figure 2.**
Illustration of non-canonical base pairs observed in the structure of iHRAS.

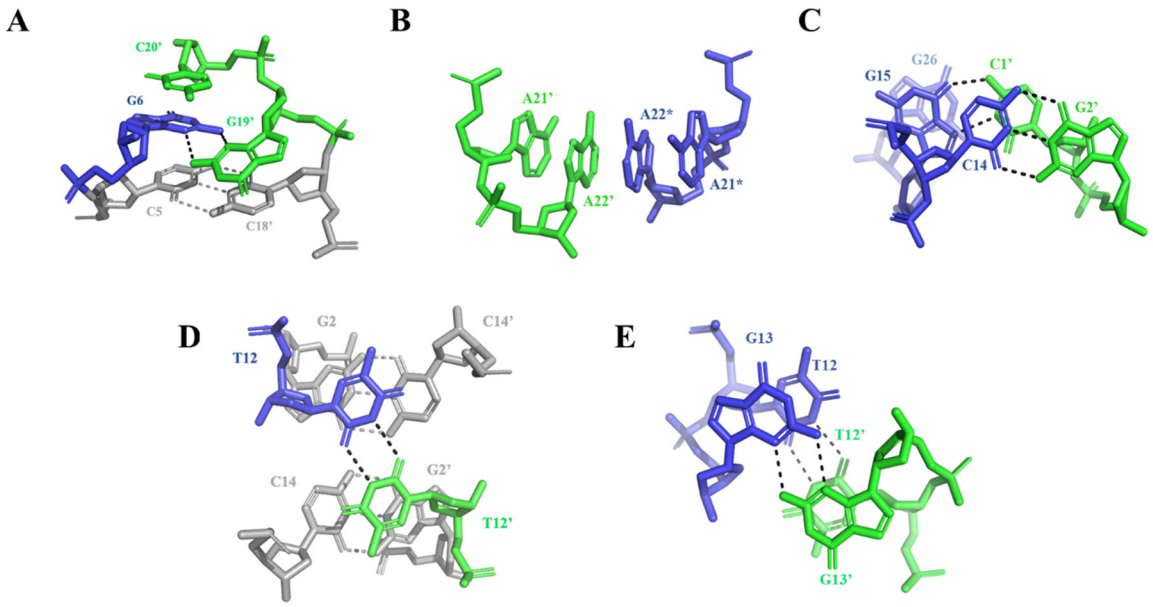



**Figure 3.**

Base pairing and stacking interactions in iHRAS. **A**) The homopurine pair, G6-G19', stacks onto the cytosine pair, C5-C18'. This arrangement is further stabilized by stacking with C20'. **B**) A four-adenine π-stack is formed by A21'-A22'-A22*-A21*. **C**) The G2'-C14 Watson-Crick pair and the C1'-G15 Hoogsteen pair interact in the connecting region. **D**) T12-T12' homopyrimidine pair stacks onto the G2-C14' and the G2'-C14 base pairs. **E**) G13-G13' stacks onto T12-T12'. Note, interactions listed in A-C refer to iM-1. Identical set of interactions are found in iM-2. Bases marked with an apostrophe belong to chain B and those marked with an asterisk are from a symmetry mate.

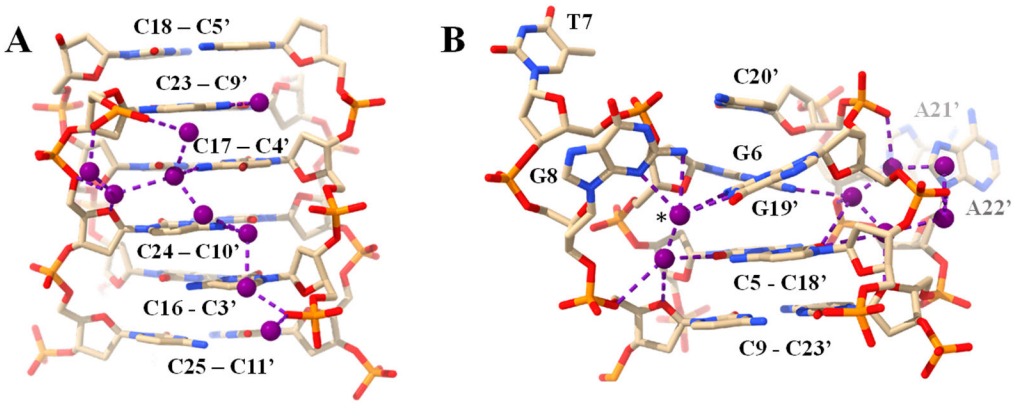



**Figure 4.**

Water networks in **A**) the major groove and **B)** loop region of iHRAS. Water molecules are colored purple. Hydrogen bonds are indicated by dashed lines. The water molecule with strong electron density is marked with an asterisk in panel B. Nucleotides are colored by atoms with P in orange, O in red, N in blue, and C in beige. The stacking of the G6-G19' base pair and C20' onto the iM-1 C5-C18' base pair is clearly visible in panel B.

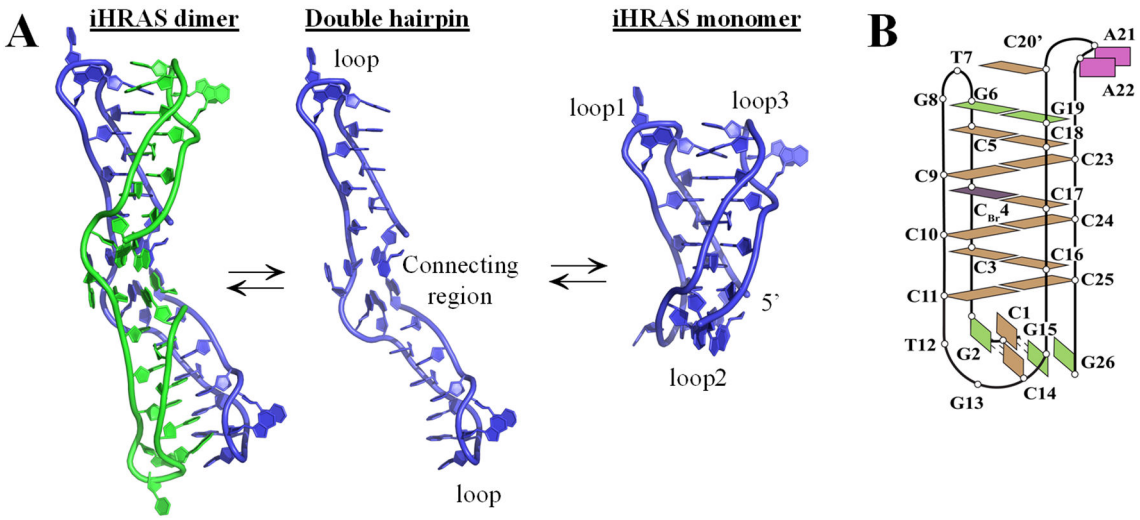



**Figure 5.**

The pathway to the iHRAS monomer. **A)** A proposed transition between iHRAS iM dimer and monomer. The intramolecular model of iHRAS, depicted in blue (right), is a regularized structural model and not a real structure. **B)** Schematics of the monomeric iHRAS