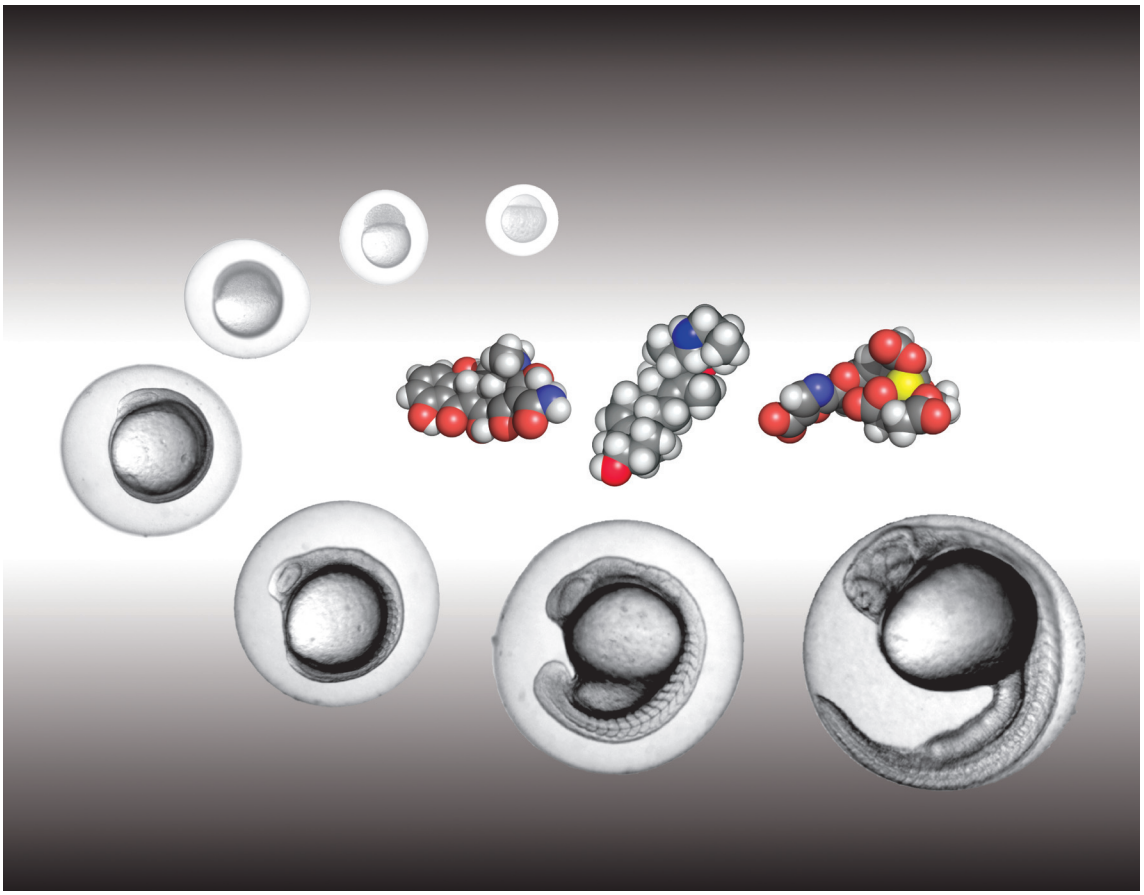
Chem Soc Rev

This article was published as part of the

# 2008 Chemistry–Biology Interface Issue

Reviewing research at the interface where chemistry meets biology

Please take a look at the full table of contents to access the other papers in this issue



# Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes†

**Julian Leon Huppert**

***Received 26th February 2008***
***First published as an Advance Article on the web 6th May 2008***
**DOI: 10.1039/b702491f**

There are many structures that can be adopted by nucleic acids other than the famous Watson–Crick duplex form. This *tutorial review* describes the guanine rich G-quadruplex structure, highlighting the chemical interactions governing its formation, and the topological variants that exist. The methods that are used to study G-quadruplex structures are described, with examples of the information that may be derived from these different methods. Next, the proposed biological functions of G-quadruplexes are discussed, highlighting especially their presence in telomeric regions and gene promoters. G-quadruplex structures are the subject of considerable interest for the development of small-molecule ligands, and are also the targets of a wide variety of natural proteins.

## Introduction to G-quadruplexes

### Structures

The double helix structure of duplex DNA is well known. The two antiparallel strands are held together by complementary basepairing between adenine (A) and thymine (T), and between guanine (G) and cytosine (C), using two hydrogen bonds in AT, and three in GC (see Fig. 1a). However, this is not the only basepairing arrangement that can occur between bases, and alternative basepairing leads to alternative structures, including triple-stranded and four-stranded structures.

*Cavendish Laboratory, University of Cambridge, JJ Thomson Ave, Cambridge, UK CB3 0HE. E-mail: jlh29@cam.ac.uk; Fax: +44 1223 337000; Tel: +44 1223 337256*
† Part of a thematic issue examining the interface of chemistry with biology.

**Julian Leon Huppert**

*Julian earned his PhD in biological chemistry with Shankar Balasubramanian, and won a Research Fellowship from Trinity College, Cambridge, which he used to work with Manolis Dermitzakis at the Wellcome Trust Sanger Institute. He is currently an RCUK Academic Fellow in Computational Biology at the Cavendish Laboratory, Cambridge. His group focuses on understanding the structure and function of complex nucleic acid structures, using a range of biophysical and bioinformatic tools. Julian also founded an award-winning biotechnology company and is actively involved in UK politics.*

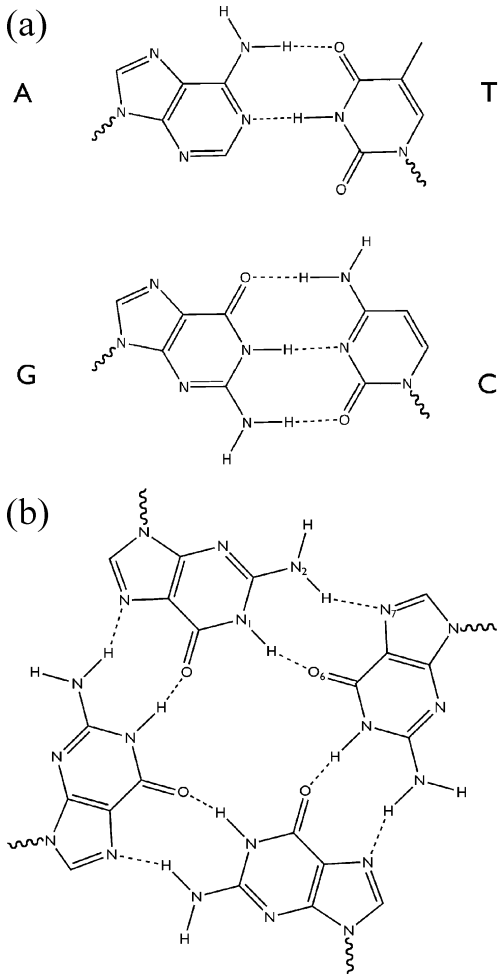



**Fig. 1** (a) Watson–Crick basepairing. Adenine and thymine form two hydrogen bonds, whereas guanine and cytosine form three hydrogen bonds. (b) Four guanines can hydrogen bond in a square arrangement to form a G-quartet. There are two hydrogen bonds on each side of the square.

This review will describe one of these four-stranded structures, called a G-quadruplex.[1–4]

G-quadruplexes, as the name suggests, have a core that is made up of guanine bases only, with four guanines arranged in a rotationally symmetric manner, making hydrogen bonds from $N_1$–$O_6$ and $N_2$–$N_7$ around the edges of the resulting square (see Fig. 1b).[5] These planar structures are called G-quartets, and are stabilized by monovalent cations, in particular $K^+$ and to a lesser extent $NH_4^+$ and $Na^+$, which interact with the lone pairs on the $O_6$ atoms surrounding the central core. They can form spontaneously at sufficiently high concentrations of guanine, and indeed were discovered in this way in 1910,[6] though the structure was not determined until 1962.[7]

These G-quartets have large π-surfaces, and hence tend to stack on each other due to π–π stacking, as well as to enable cations to intercalate between the G-quartets. In particular, oligonucleotides with contiguous runs of guanine, such as d(TGGGT) can form stacked structures with the G-quartets linked by the sugar–phosphate backbone. These are called G-quadruplexes and can form from DNA or RNA strands (or other variants, such as PNA). They are helical in nature due to the constraints of π–π stacking, although for convenience they are often depicted without the helicity, as shown in Fig. 2.

Since there is a directionality to the strands, customarily described as from the 5′ end to the 3′ end, there are topological variants possible for these four strands. All four strands may be parallel, three parallel and one in the opposite direction (antiparallel), or there may be two in one direction and two in the other, either with the parallel pairs adjacent to each other or opposite each other. A shorthand has arisen which describes all the arrangements with at least one antiparallel strand as 'antiparallel', although this does not give a full description of the structures. These are depicted in Fig. 3.

At a molecular level, the different directionality of the strands relates to the conformational state of the glycosidic bond between the guanine base and the sugar. This may be either *syn* or *anti*, as depicted in Fig. 4. When all four strands are parallel, all the bases are in the *anti* conformation and the grooves between the backbones are all of equal size—the system is entirely $C_4$ symmetric. When any of the strands are antiparallel, the bases must be in the *syn* form in order for the hydrogen bonds to be formed correctly. This then affects the orientation of the backbone relative to the G-quartets, and hence results in grooves of different sizes. When successive guanines (starting with the guanine contributing N1 and N2) are both *anti* or both *syn*, the groove is medium in size; if the first is *anti* and the second *syn*, the groove is wider, and if the first is *syn* and the second *anti*, then the groove is narrower. Thus a G-quadruplex with adjacent parallel strands will be arranged with glycosidic bonds *anti–syn–syn–anti*, and will have grooves that are wide, medium, narrow and medium. In contrast, a structure with alternating strands will have glycosidic bonds *anti–syn–anti–syn*, with grooves wide, narrow, wide and narrow.

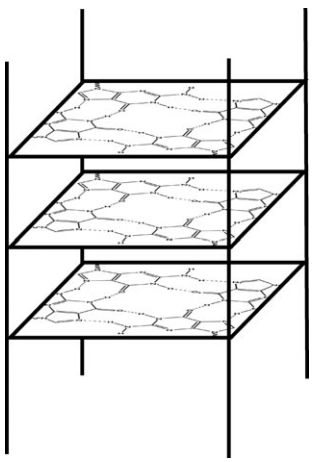


**Fig. 2** G-quartets can stack on top of each other, forming G-quadruplex structures. These are held together by π-stacking and the sugar–phosphate backbone.

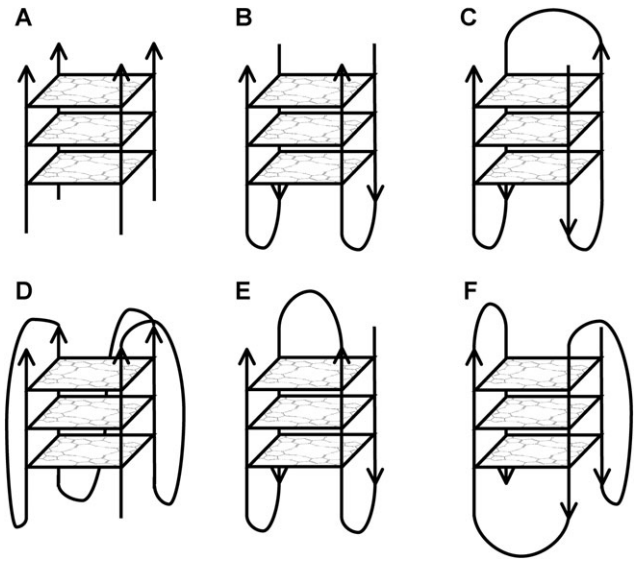



**Fig. 3** G-quadruplexes can adopt a range of different stoichiometries and folding patterns. (A) Tetramolecular structure with all strands parallel; (B) bimolecular antiparallel structure with adjacent parallel strands; (C) unimolecular antiparallel structure with alternating parallel strands; (D) unimolecular parallel structure with three double chain reversal loops; (E) unimolecular antiparallel structure with adjacent parallel strands and a diagonal loop; (F) unimolecular mixed structure with three parallel and one antiparallel strands. All three structures (D), (E) and (F) have been observed for the human telomeric repeat.

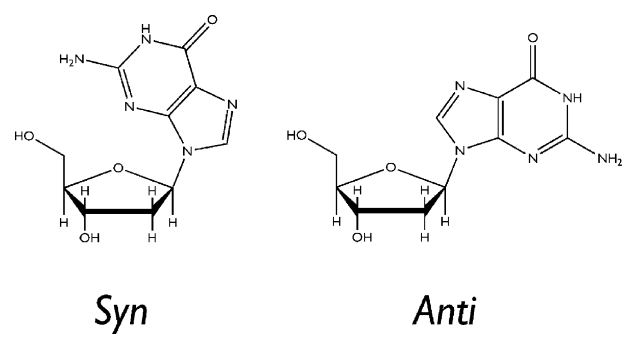



**Fig. 4** The bond between the base and sugar can rotate. It has two preferred conformations, *syn* and *anti*.

G-quadruplexes may be comprised of four separate strands, as in the example above, forming tetramolecular G-quadruplexes, which are always found in the all-*anti* parallel form. Alternatively, they may be formed from two strands, each with two sets of contiguous guanines, or just from one strand, folding back on itself to form an intramolecular structure. In either of these cases, there will be loops that serve to connect the strands of the structure together. Depending on which strands are connected, these loops may cross diagonally across the top of the structure, joining diagonally opposed antiparallel strands; go across a side, linking adjacent antiparallel strands; or may loop around the side of the structure linking parallel strands and forming a double-strand reversal loop. Some examples of these are shown in Fig. 3.

### Methods for studying G-quadruplexes

There are a number of different experimental techniques used to study G-quadruplex formation, each examining different aspects of the structures, and hence reporting on different aspects of their formation. The majority of these techniques are principally descriptive, and complete structure determination requires the use of either NMR structure determination or X-ray crystallography. X-Ray crystallography[8] was the first technique used for complete structure determination, and requires the production of single crystals of the structure to be studied. Despite the development of high throughput screens to assist with crystal growth, this is still a relatively slow and uncertain procedure. Once crystals have been obtained, however, a wealth of detail about the structures can be obtained, and over 50 crystal structures of G-quadruplexes are now available in the protein data bank/nucleic acid data bank (PDB and NDB). These include both nucleic acid-only structures and structures with bound ligands. The resolution for some of these structures is extremely good, below 1 Å in some instances. One significant disadvantage of X-ray crystal structures is that by their very nature they can only report on the structure adopted in the solid state, which may not be the same as the structure adopted in solution. In particular, if a substance is polymorphic, the crystal structure will generally describe the form that crystallizes most easily, rather than that which would otherwise be favoured. This is a particular concern for G-quadruplex structures, as they are frequently highly polymorphic.

The other technique for detailed structure determination is NMR spectroscopy.[9] This requires much less sample preparation than crystallography, but does require very pure and high-concentration samples. With the increasing ease of custom synthesis, this is relatively straightforward. At the simplest level, it is possible to gain much information even from a 1-D $^{1}$H NMR spectrum, as there are a relatively small number of protons in nucleic acids and the guanine $NH_1$ imino protons have a characteristic shift when hydrogen bonded. In addition, they exchange relatively slowly with the deuterated solvent when compared to non-hydrogen-bonded protons. This may therefore be used to show G-quadruplex formation. In order to provide more detailed analysis, multi-dimensional techniques are needed, which allow the complete assignment of resonances to the sequence being studied. Correlations may then be used to reveal backbone conformations and sugar pucker angles, and ultimately the full structure. In some instances, partial labelling (with $^{15}$N, $^{13}$C and $^{31}$P) is necessary for improved clarity, and base substitution may also be used, such as replacing particular guanines with 8-bromoguanine. NMR techniques may also be used to study dynamics and kinetics, and can also report on polymorphism. Some 30 G-quadruplex structures from NMR studies are currently in the PDB/NDB.

Other techniques report on particular features of the G-quadruplex structure and hence are supportive of the structure, rather than explicit about its details. One of the earliest techniques used was dimethylsulfate (DMS) footprinting, which was part of the Maxam and Gilbert protocol for DNA sequencing. DMS methylates the $N_7$ position of guanine, which then leads to facile depurination, as shown in Fig. 5. The addition of piperidine then leads to cleavage at the now abasic site, and gel electrophoresis allows visualization of the length of the cleaved fragments, resulting in a ladder with peaks corresponding to every guanine in a sequence. However, in a G-quadruplex the $N_7$ is hydrogen bonded, and hence is protected from methylation, resulting in little or no cleavage at the guanines involved in G-quadruplex formation.

Nucleic acids absorb ultraviolet light, in a manner that varies with their base stacking and with temperature. One consequence of this is that it is possible to study G-quadruplexes by changing the temperature from a temperature at which they are stable to a higher temperature where they are unstable. They undergo a melting transition between these two temperatures, which may be followed by monitoring the absorbance of UV light at 295 nm.[10,11] At this wavelength, there is a marked hypochromic shift (*i.e.* lower absorbance)

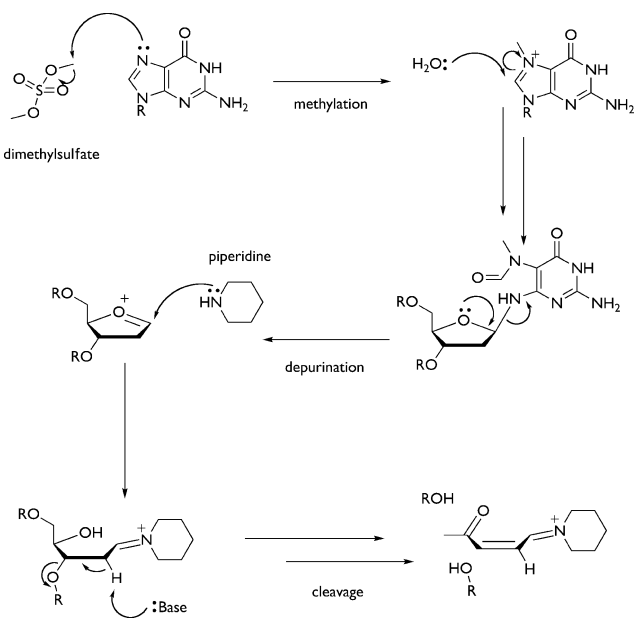



**Fig. 5** Dimethylsulfate treatment followed by addition of piperidine leads to cleavage of DNA sequences at positions containing guanine. Protection of $N_7$ *via* hydrogen bonding protects those guanines from cleavage.

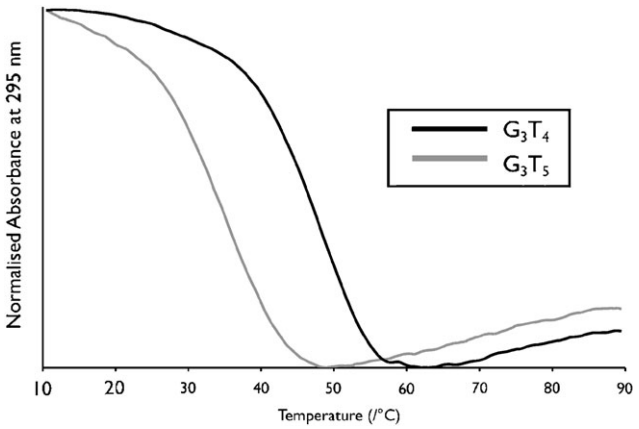



**Fig. 6** G-quadruplexes undergo a hypochromic shift at 295 nm upon melting. The midpoint of the melting transition is the melting temperature, $T_m$. Sequences with longer loops are less stable. Sequences shown in this plot are $G_3T_4$ = d($TG_3T_4G_3T_4G_3T_4G_3T$) and $G_3T_5$ = d($TG_3T_5G_3T_5G_3T_5G_3T$). Data taken from ref. 12.

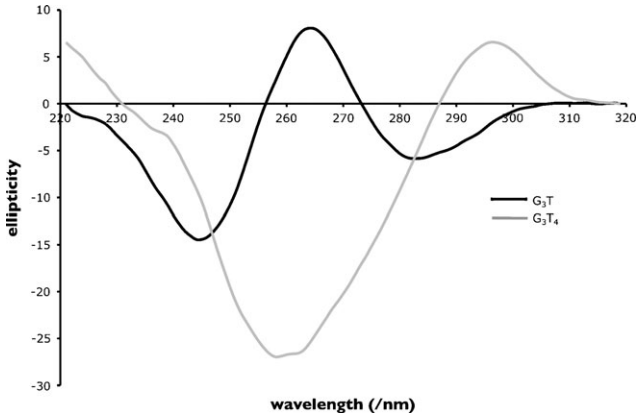



**Fig. 7** Circular dichroism spectra provide a diagnostic tool to predict the structure of a G-quadruplex. A peak around 260 nm and a trough around 240 nm implies the presence of a parallel G-quadruplex structure and a peak around 295 nm with a trough around 260 nm implies an antiparallel G-quadruplex. Sequences shown are $G_3T$ = d($TG_3TG_3TG_3TG_3T$) and $G_3T_4$ = d($TG_3T_4G_3T_4G_3T_4G_3T$). Data taken from ref. 12.

upon melting, as seen in Fig. 6. This allows ready determination of the melting temperature ($T_m$, the temperature at which half the G-quadruplexes have denatured), and, by making the assumption that melting is a two-state process, it is possible to perform a detailed van't Hoff analysis to extract values for thermodynamic variables such as $\Delta G$, $\Delta H$ and $\Delta S$. In essence, since $\Delta G = \Delta H - T\Delta S = -RT \ln K_{eq}$, a plot of $\ln K_{eq}$ against $1/T$ should give a straight line with gradient $-\Delta H/R$, and intercept $\Delta S/R$.

These results have shown that G-quadruplexes are very stable under pseudo-physiological conditions (100 mM KCl, pH 7.4), with many sequences having melting temperatures above 60 °C. The melting temperature is affected by the sequence of the loops, and also by their length.[12,13] Longer loops are significantly less stable, whereas shorter loops, and particularly single-base loops are especially stable. Some sequences with very short loops, such as the sequence d(TGGGTGGGTGGGTGGG) with all three loops a single-base T, melt in excess of 95 °C and only show a measurable melting transition when the $K^+$ concentration is lowered significantly so as to destabilize them. RNA G-quadruplexes are generally significantly more stable than the equivalent sequences formed of DNA, and have longer lifetimes as a result.[14]

Circular dichroism (CD) spectroscopy is also used to identify G-quadruplex structures and in particular to distinguish all-parallel structures from antiparallel structures.[15] In CD spectroscopy, circularly polarized light is shone through a solution, and if there is a chiral species in the solution, it will generally interact asymmetrically with the enantiomeric forms of light, and the asymmetry varies with wavelength. This may then be used to produce difference plots, which are characteristic of different structures, as exemplified in Fig. 7. However, there is only limited theory to date[16] to predict the form of the CD spectrum from a molecular structure, nor the structure from the CD spectrum. Nonetheless, CD still represents one of the simplest ways of predicting the folding structure of a G-quadruplex. In general, a peak in CD at 260 nm wavelength, and a trough at 240 nm is descriptive of an all-parallel structure, whereas a peak at 295 nm and a trough at 260 nm describes an antiparallel structure. Polymorphic forms will contain a superposition of the CD spectra deriving from the pure forms. Interestingly, sequences with single-base loops, as well as being especially thermally stable, also form almost exclusively parallel structures, with double-chain reversal loops. Although it seems counter-intuitive that these loops, linking the top and bottom of a G-quadruplex stack, are favoured for such short lengths, the helicity of the structure actually means that the distance involved is relatively small.

Fluorescence spectroscopy has also been used to study G-quadruplex folding. One approach uses dye–quencher pairings and performing thermal melting in such a way that in the folded structure the fluorescent dye is quenched, whereas upon melting the quencher is separated from the dye.[17] These results may be analysed in much the same way as the UV melting data described earlier. Alternatively, a FRET pair of dyes may be used, as shown schematically in Fig. 8.[18] FRET (fluorescence resonance energy transfer) is a phenomenon by which, when two appropriate dyes are in proximity to each other, the excitation of one (at lower wavelengths) may lead to non-radiative transfer of energy to the second dye, rather than fluorescent emission of the first. This second dye then emits fluorescently at a higher wavelength. The amount of energy transferred is extremely sensitive to the distance between the two fluorophores, on a 1–10 nm scale, and hence can be used to report the separation of any two points in a G-quadruplex structure. Using fluorescence measurements has the advantage that there is a very strong signal-to-noise ratio and hence small amounts of material may be effectively studied. A number of single-molecule studies of G-quadruplex structures have been performed making use of this sensitivity.[19] However, fluorescence use does require that the nucleic acid sequence be dual-labelled, which can be expensive or chemically challenging, and potentially could alter the structure formed.

In addition to these principal methods, there are a wide range of other techniques that have been used. These include

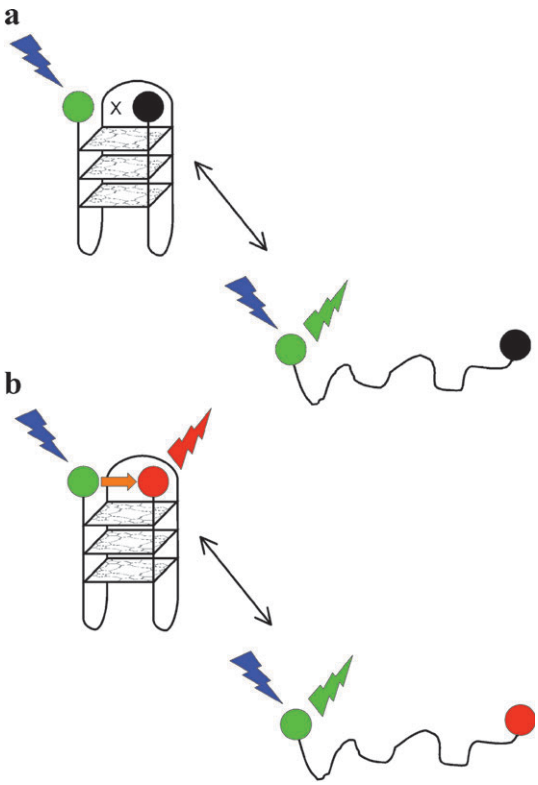



**Fig. 8** (a) A DNA sequence labelled at one end with a fluorescent dye and at the other end with a quencher, which absorbs energy but does not re-emit it, can be used to monitor folding of a G-quadruplex, as the dye emits fluorescence in the unfolded form, but not the folded form. (b) An alternative strategy uses fluorescence resonance energy transfer. Instead of a quencher, another dye is used. When the two dyes are close, energy may be transferred from the excited dye to the other dye, which then re-emits at a longer wavelength. The amount of direct emission versus FRET emission indicates how close the dyes are.

both physical measurements, such as mass spectrometry,[20] atomic force microscopy (AFM),[21] Raman spectroscopy[22] and infrared spectroscopy,[23] as well as biochemical techniques, such as a polymerase stop assay,[24] which relies on the fact that G-quadruplexes can induce pausing in DNA polymerases, and computational methods such as molecular dynamics.[25] Gel electrophoresis is also widely used as a method of approximately describing molecular size.

### G-quadruplex predictions

Based on the biophysical experiments described above, it is possible to broadly predict which sequences of single-stranded DNA will form G-quadruplex structures *in vitro*. Essentially, sequences with at least four runs of GGG, with loops of 1–7 bases between them, of any sequence, form stable G-quadruplexes. If the runs of GGG are shorter or mutated, the stability decreases significantly, and similarly longer loops result in less stable sequences. This rule has been formally expressed and implemented as an algorithm called *quadparser*, which is freely available.[26] This tool allows the prediction of putative G-quadruplex sequences (PQS) within any DNA sequence and has been widely used. Other variants and approaches have also been used for this purpose.[27]

Applying *quadparser* to the entire human genome reveals 376 000 PQS, which is less than would be expected by chance.[26] There is no suggestion that all of these PQS actually form physiologically at the same time, as other factors such as chromatin structure and supercoiling will affect their formation. Additionally, some of the identified sequences are probably non-functional elements present by chance. However, *quadparser* still provides a useful analytical tool to predict functional G-quadruplexes, by detailed analysis of particular regions and the use of methods such as conservation to support the predictions.

## G-quadruplex functions

### Telomeres and telomerase

It is important for eukaryotic cells, which have linear chromosomes, to be able to distinguish between chromosome ends and unexpected breaks in the DNA. In order to facilitate this discrimination, they have repeated sequences at the ends, called telomeres.[28] In all vertebrates, this repeated sequence has the pattern d(GGGTTA)$_n$, and other organisms generally have very similar sequences, characterized by runs of GGG with intervening bases, often thymine. In humans, there are typically more than 1000 repeats of this sequence in double-stranded form, followed by a smaller overhanging single-stranded region with 1–200 bases, but only consisting of the G-rich strand. The reason for the overhang is that polymerases cannot replicate the extreme 5′ end of a DNA strand, because of the need for an RNA primer in this position, which is then degraded. As a result of this so-called end replication problem, the length of the double-stranded region of the telomere becomes shorter with every cell division. Ultimately, the telomere becomes too short, leading to chromosome fusion, senescence and apoptosis. This therefore imposes a finite lifetime on cells, unless something acts to elongate the telomeres. Stem cells, which need to be immortal and hence bypass this limit, have an enzyme called telomerase,[29] which elongates telomeres using an internal RNA template, laying down the d(GGGTTA) repeats. In around 85% of cancers, the cancerous cells also bypass the limit on cell divisions by expressing telomerase, and there is therefore a great deal of interest in developing approaches to reduce the activity of telomerase for therapeutic purposes.[30–33]

The human telomeric sequence, d(GGGTTA)$_n$ folds spontaneously into an intramolecular G-quadruplex form, with the GGG runs forming the G-quartet core, and TTA forming the loops of the structure.[34] This structure is stable under physiological conditions, with a thermal melting temperature of around 65 °C. All telomeric sequences studied to date can also form G-quadruplex structures with comparable thermal stability. It has therefore been proposed that the physiologically relevant structure of the telomeric overhang has a series of G-quadruplexes, much like beads on a string. This proposal has been elegantly supported by biochemical studies, in particular by the demonstration that an antibody specific for parallel G-quadruplexes binds very tightly to the telomeres in the macronuclei of the ciliate *Stylonychia lemnae*, which has the telomeric repeat sequence d(G$_4$T$_4$).[35,36] This unusual ciliate was used for the studies because it forms macronuclei containing millions of

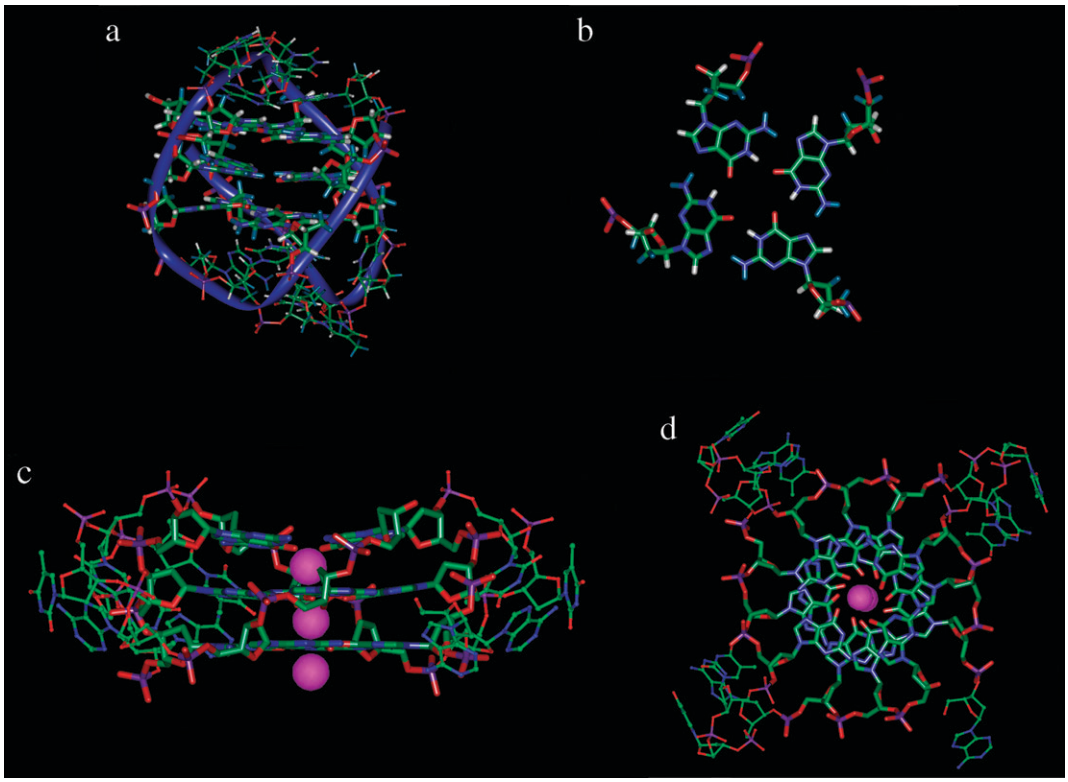



**Fig. 9** (a) Side view of the antiparallel human telomeric G-quadruplex structure solved by Wang and Patel using NMR spectroscopy, from PDB entry 143D. (b) Detailed view of the central G-quartet from PDB entry 143D. (c) Side view of the parallel human telomeric G-quadruplex structure solved by Parkinson, Lee and Neidle using X-ray crystallography, from PDB entry 1KF1. (d) Top view of the parallel structure from PDB entry 1KF1. In all cases, guanines are shown as cylinders, other bases as balls and sticks. Potassium ions are shown in magenta. Pictures were drawn using iMol.

minichromosomes, each around the size of one gene with telomeres on each end. They therefore have a much greater concentration of telomeres than could be found in humans or most other species. This telomeric repeat sequence has become an important target for drug development,[30–33] as it has been shown that by binding to and stabilizing telomeric G-quadruplexes, it is possible to block telomerase from acting and extending the telomeres, hence preventing the immortalisation of cancerous cells.

The structures formed by the telomeric repeat have been the subject of considerable study, and various structures have been solved for the telomeric repeat, using slightly different sequences and conditions. The first of these was by NMR spectroscopy in $Na^+$ solution, solved by Wang and Patel in 1993,[37] which revealed an adjacent antiparallel structure, with the central loop travelling diagonally over one face, and the other two loops on the same side as each other, linking adjacent strands (Fig. 9a and b). The next significant structure was solved by Parkinson, Lee and Neidle in 2002 using X-ray crystallography. They used the same sequence but crystallized the structure from a solution containing $K^+$ ions.[38] In contrast to that observed by NMR, they obtained a parallel structure, with all three loops forming double chain reversal loops to link the adjacent parallel strands together (Fig. 9c and d). More recent studies have shown that the telomeric sequence can form a wide variety of different structures, which all seem to exist in equilibrium with each other.[39–41]

## Transcription regulation

Gene transcription is tightly regulated, by a variety of methods. One method that is used in some cases is based on the presence of G-quadruplexes located in the promoter region of a gene, broadly speaking the kilobase upstream of the transcription start site (TSS).[33,42] This model was originally demonstrated by Hurley and co-workers for the oncogene *c-myc*,[43] an important transcription factor involved in regulating around 15% of all human genes. As a result of this, overexpression of *c-myc* has been implicated in a wide range of cancers including colorectal cancer. Within its promoter there is a region, 115–142 basepairs upstream of the TSS, which is highly sensitive to nucleases, suggesting that it forms an accessible structure free from histone proteins. This region controls the vast majority of the transcription of the gene, and studies *in vitro* of the sequence d(***GGGG***A***GGG***T***GGG-GA******GGG***T***GGGG***AAGG) show that it is capable of forming into a family of polymorphic G-quadruplexes, using various combinations of the guanine runs underlined.[44] It has further been shown that the G-quadruplex ligand TMPyP4 (see below) binds to this element leading to downregulation of *c-myc* expression.[45]

This clear proof of principle led to the proposal that this may be a general mechanism for gene regulation. The simplest form of the model (Fig. 10) proposes that there is an equilibrium between two forms of the DNA. On one side of the

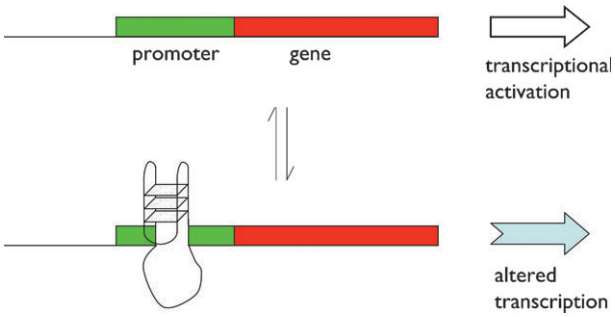



**Fig. 10** The formation of a G-quadruplex in a promoter can affect the level and nature of transcription from that gene. At the simplest level, it may act as a steric block to the transcription machinery.

equilibrium is double helix DNA, and transcription occurs as normal; on the other side, one strand is separated, and has folded up into a G-quadruplex. This structure then acts as a steric block to transcription. Addition of a G-quadruplex ligand, whether a small molecule or a protein, will energetically favour the G-quadruplex form, and hence move the equilibrium towards that side and reduce the transcriptional activity. This dynamic equilibrium has been experimentally demonstrated.[46]

Although this model is presented in terms of steric blockage leading to a reduction in transcriptional activity, as was found for *c-myc*, it is also possible that the G-quadruplex form could be an activating domain, either because of putative protein recognition of the G-quadruplex, or if the accessibility of the other strand leads to increased transcriptional activity. However, although a wide variety of different genes have now been shown to have promoter G-quadruplexes,[33] such as *VEGF*, *HIF-1a*, *Bcl-2*, *Ret*, *c-kit* and *KRAS*, none have yet had G-quadruplex formation leading to increased transcriptional activity.

Although in Fig. 10 the complementary C-rich strand is drawn as an unstructured sequence, it is possible that it could form an alternative four-stranded structure called an i-motif (Fig. 11).[47] An i-motif has four strands forming a structure somewhat like two interleaved ladders, with pairs of strands held together by diagonal $C \equiv C^+$ bonds. These rely on the protonation of the $N_3$ of cytosine, which has a $pK_a$ of 4.2. As a result, these structures are generally only stable under acidic conditions, but the stability will clearly be different in the context of chromosomal DNA, especially if a G-quadruplex structure forms and holds the ends of the i-motif together in the correct orientation. Nonetheless, there is still controversy as to the biological relevance of the i-motif structure.[48]

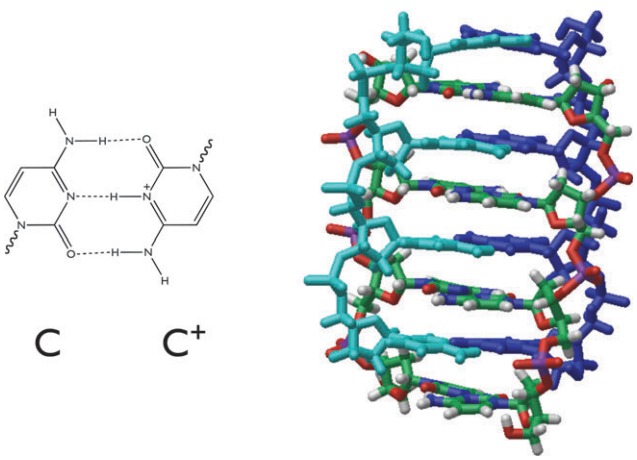



**Fig. 11** Left: two cytosines can pair up, forming three hydrogen bonds, when one of them has been protonated. Right: four cytosine-containing strands can form an i-motif, with interleaved $C:C^+$ bonds. One pair of binding strands is shown in CPK colouring, the other in blue/cyan. Data was taken from PDB entry 190D, and the figure was produced using iMol.

Using the predictive algorithm *quadparser* described earlier,[26] it is possible to investigate how many human genes contain G-quadruplex motifs in their promoter regions, both to identify novel genes for experimental testing, but also to see if there is evidence of over- or under-representation of such genes. This analysis has shown that almost half of all genes (43%) contain putative G-quadruplex structures in their promoters, which is considerably more than would be expected by chance, based on the rest of the genome.[49] The enrichment of G-quadruplex motifs occurs increasingly nearer the TSS, with the first few hundred bases seeming to be particularly important for G-quadruplex motifs to be present.

Interestingly, genes involved in cancer are even more likely to have such promoter G-quadruplexes, with 67% having such sequences. Using the online Gene Ontology database, which describes the functions of every human gene, it is possible to see if there is any general bias for types of genes that have promoter G-quadruplexes. This has revealed that genes involved in tightly-regulated processes such as development, neurogenesis and cell differentiation are more likely than other classes of genes to have promoter G-quadruplexes, whereas genes involved in processes such as protein biosynthesis, olfaction and immune response are much less likely to have promoter G-quadruplexes.

## Other physiological functions

G-quadruplexes have been shown to exist in other physiologically important locations and have been shown experimentally to affect important biological processes. The formation of G-quadruplexes in mRNA sequences is likely to be especially stable, both because RNA G-quadruplexes are inherently more stable, and also because there is no complementary strand present to compete with G-quadruplex formation, although there are other single-stranded structures that RNA can form. It has been shown experimentally that a G-quadruplex at the 5′ end of the 5′ untranslated region of the mRNA of the signal transduction gene *NRAS* has the effect of reducing translational efficiency—removing or mutating the G-quadruplex forming sequence resulted in a 4-fold increase in translation. Genomic searching suggests this may be true for thousands of other genes.[50]

Normally after transcription has occurred, the RNA strand separates entirely from the DNA duplex, but in some areas, particularly those involved in immunoglobulin class switch recombination, this does not occur, and the RNA remains bound to the template DNA strand, leaving the coding DNA strand to form a loop. In some instances, these loops are

sufficiently large to be directly visualized by electron microscopy. It has been shown by the binding of the G-quadruplex binding protein nucleolin that the looped out coding strand forms G-quadruplexes, suggesting that one function of G-quadruplexes may be to stabilize this unusual looped RNA–DNA duplex arrangement.[51]

It has also been suggested that G-quadruplexes may play a role in splicing, meiosis, and replication. Many other functional roles may also be suggested, and significant further experimental work is required to find out which of these are in fact related to G-quadruplexes.

## Binding G-quadruplexes

### Small molecule ligands

Understanding the functional roles of G-quadruplexes is interesting in its own right, but there is also interest in trying to develop ligands that can bind to and hence stabilize them, leading to the possibility of novel therapeutics. There has been considerable work on developing G-quadruplex ligands, especially to target the human telomeric repeat and hence block the action of telomerase.[31–33]

G-quadruplex structures present a large π-surface, roughly twice as large as that found in DNA, since there are four coplanar bases rather than two. As a result the majority of the small molecules that bind G-quadruplexes themselves have a large π-surface, so as to maximize the π–π interactions they can form. Another design feature is that G-quadruplexes, like all nucleic acids, carry a high negative charge, and hence cationic ligands will generally bind more tightly to them, although in a non-specific way.

Using the above principles, it is relatively easy to design compounds that will bind G-quadruplexes, although not necessarily with high affinities. However, developing discrimination such that the compounds do not also target duplex DNA is significantly harder, and many G-quadruplex ligands do also bind duplex DNA. This is a major problem for any work involving cells, as duplex DNA generally far outnumbers any G-quadruplexes present. Nonetheless, some good G-quadruplex binders have been developed, such as those depicted in Fig. 12. These include the cationic porphyrin 5,10,15,20-tetra(*N*-methyl-4-pyridyl)porphin, TMPyP4 (although widely used, this has only limited selectivity for G-quadruplexes over duplexes),[52] and a variety of acridine and acridone compounds, such as the 3,6,9-substituted acridine BRACO-19.[53] The tightest known G-quadruplex binder is the naturally occurring macrocycle telomestatin, found in the bacterium *Streptomyces anulatus*. This has a potent anti-telomerase activity and was reported to have an $EC_{50}$ (half maximal effective concentration) of 5 nM,[54] although the methodology used has been criticised.[55]

There are a number of techniques that have been used to study the binding of these molecules to G-quadruplex sequences, including indirect methods, such as by examining the change in melting temperature of an oligonucleotide, and by observing protection in biochemical assays. Other methods measure the binding more directly and, in particular, surface plasmon resonance (SPR) has been used extensively to measure the binding constants of G-quadruplex ligands, together with their stoichiometries.[56] Since most G-quadruplex ligands are fluorescent, due to their large π-systems, this can be used to measure binding, either through changes in the quantum efficiency of the ligand when bound to the G-quadruplex, or *via* anisotropy measurements.

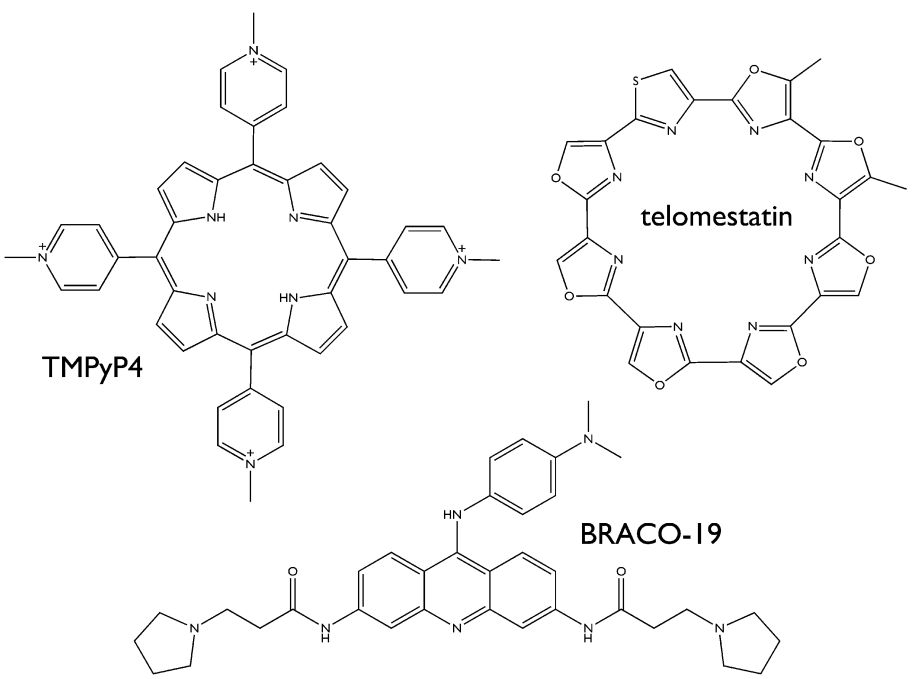



**Fig. 12** π-Rich molecules can bind to G-quadruplexes strongly and selectively. These include the porphyrin TMPyP4, the tri-substituted acridine BRACO-19, and the natural product telomestatin.

One elegant method for studying the binding of ligands to different nucleic acid systems is competition dialysis.[57] In this experiment, a large number of nucleic acid solutions of various forms are placed into microdialysis units, and then dialysed against a solution containing the ligand. The system is then left for a long time to equilibrate. Since the ligand (but not the nucleic acid) is free to move between the wells of the microdialysis plate, as well as the dialysate, its final concentration in each well will reflect the amount of binding to the nucleic acid species in that well, thus allowing relatively straightforward comparison of binding affinities between a large number of species simultaneously.

In order for ligand binding to be therapeutically effective, it is not enough for the ligand to bind to G-quadruplexes, or even to be very highly selective for them over duplex DNA. It must also be able to bind selectively to one G-quadruplex over another. This is a big challenge, since there are relatively few recognition points to discriminate different G-quadruplex structures. Various methods have been used to try to combine targeting of the loops and grooves of each structure with targeting of the G-quartet core, but to date this has only provided limited success.[58]

Despite these concerns, clinical trials of some ligands are proceeding, with considerable success. Cylene Pharmaceuticals Inc developed an anticancer agent called Quorfloxacin,[59] which has shown great promise in clinical trials. It was originally targeted against the *c-myc* promoter, but appears to derive its anti-cancer effect by targeting ribosomal quadruplexes and blocking the binding of the protein nucleolin. G-quadruplexes can also act as drugs in their own right, and Antisoma has completed some clinical trials on a G-quadruplex-forming oligonucleotide, which targets nucleolin directly, by competing for its natural substrate.[60]

### Proteins

In keeping with its proposed biological functions, there are a wide variety of G-quadruplex binding proteins that occur naturally, and over 30 have been reported so far.[61] These include G-quadruplex binding proteins, proteins that promote the formation of G-quadruplexes, helicases that can unwind G-quadruplexes, proteins that destabilize G-quadruplexes, and G-quadruplex-specific nucleases. The completeness of this list of functions fits with the premise that G-quadruplexes can and do form *in vivo*, and that their formation is regulated both positively and negatively. Many of these proteins operate principally at telomeres, where most experimental study has focused. Of particular interest is the family of RecQ helicases, found in a wide variety of species, which unwind G-quadruplexes with great specificity over duplexes. In humans, their absence results in chromosome instability, leading to serious medical conditions, called Bloom's syndrome and Werner's syndrome, depending upon which of the two human RecQ helicases is malfunctioning. This has been attributed to the inability to remove formed G-quadruplexes for recombination and replication.

Artificial proteins have also been engineered to bind G-quadruplexes. In particular, antibodies have been developed that bind highly selectively to G-quadruplexes, and display some discrimination between different forms of G-quadruplex. This has been used to demonstrate experimentally that G-quadruplexes can and do form *in vivo*.[35] Another approach has focused on using phage display libraries to evolve a G-quadruplex binder from the three zinc finger protein Zif-268, which is a naturally sequence specific duplex binder and transcription factor. A mere 12 point mutations sufficed to convert the protein from a duplex binder which doesn't bind G-quadruplexes, to a relatively tight G-quadruplex binder ($K_d$ 25 nM) with no duplex binding.[62]

## Conclusion

G-quadruplex structures are interesting on many levels. Structurally, they display a fascinating array of polymorphic structures, and we are unable as yet to predict their structure or stability theoretically. They seem to play a number of important biological roles, including regulating the critical processes of transcription and translation, and there is pharmaceutical interest in being able to manipulate these processes to develop novel therapeutics. They can also be used for a range of innovative nanotechnological applications—including all the ones no-one has yet envisaged. It is a rapidly growing field, with promise for chemists, biologists, physicists and informaticians alike.

## Acknowledgements

JH is a Research Councils UK Academic Fellow, and a Fellow of Trinity College Cambridge, who financially supported this work. JH would like to thank Dr Caroline Wright of the PHG Foundation for proofreading and advice.

## References

1. *Quadruplex Nucleic Acids*, ed. S Neidle and S Balasubramanian, Royal Society of Chemistry, Cambridge, UK, 2006.
2. T. Simonsson, *Biol. Chem.*, 2001, **382**, 621.
3. V. Dapic, V. Abdomerovic, R. Marrington, J. Peberdy, A. Rodger, J. O. Trent and P. J. Bates, *Nucleic Acids Res.*, 2003, **31**, 2097.
4. J. L. Huppert, *Philos. Trans. R. Soc. London, Ser. A*, 2007, **365**, 2969.
5. S. Burge, G. N. Parkinson, P. Hazel, A. K. Todd and S. Neidle, *Nucleic Acids Res.*, 2006, **34**, 5402.
6. I. Bang, *Biochem. Z.*, 1910, **26**, 293.
7. M. Gellert, M. N. Lipsett and D. R. Davies, *Proc. Natl. Acad. Sci. U. S. A.*, 1962, **48**, 2013.
8. N. H. Campbell and G. N. Parkinson, *Methods*, 2007, **43**, 252.
9. M. W. Da Silva, *Methods*, 2007, **43**, 264.
10. P. A. Rachwal and K. R. Fox, *Methods*, 2007, **43**, 291.
11. J.-L. Mergny, A. Phan and L. Lacroix, *FEBS Lett.*, 1998, **435**, 74.
12. P. Hazel, J. L. Huppert, S. Balasubramanian and S. Neidle, *J. Am. Chem. Soc.*, 2004, **126**, 16405.
13. P. A. Rachwal, T. Brown and K. R. Fox, *FEBS Lett.*, 2007, **581**, 1657.
14. J.-L. Mergny, A. De Cian, A. Ghelab, B. Sacca and L. Lacroix, *Nucleic Acids Res.*, 2005, **33**, 81.
15. S. Paramasivan, I. Rujan and P. H. Bolton, *Methods*, 2007, **43**, 324.
16. D. M. Gray, J. D. Wen, C. W. Gray, R. Repges, C. Repges, G. Raabe and J. Fleishhauer, *Chirality*, 2008, **20**, 431.
17. R. A. J. Darby, M. Sollogoub, C. McKeen, L. Brown, A. Risitano, N. Brown, C. Barton, T. Brown and K. R. Fox, *Nucleic Acids Res.*, 2002, **30**, 39.
18. T. Simonsson and R. Sjöback, *J. Biol. Chem.*, 1999, **274**, 17379.

19. L. M. Ying, J. J. Green, H. T. Li, D. Klenerman and S. Balasubramanian, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 14629.
20. F. Rosu, V. Gabelica, C. Houssier, P. Colson and E. De Pauw, *Rapid Commun. Mass Spectrom.*, 2002, **16**, 1729.
21. L. T. Costa, M. Kerkmann, G. Hartmann, S. Endres, P. M. Bisch, W. M. Heckl and S. Thalhammer, *Biochem. Biophys. Res. Commun.*, 2004, **313**, 1065.
22. T. Miura and G. J. Thomas, Jr, *Biochemistry*, 1995, **34**, 9645.
23. V. Gabelica, F. Rosu, E. De Pauw, J. Lemaire, J.-C. Gillet, J.-C. Poully, F. Lecomte, G. Grégoire, J.-P. Schermann and C. Desfrançois, *J. Am. Chem. Soc.*, 2008, **130**, 1810.
24. H. Han, L. H. Hurley and M. Salazar, *Nucleic Acids Res.*, 1999, **27**, 537.
25. J. Sponer and N. Spackova, *Methods*, 2007, **43**, 278.
26. J. L. Huppert and S. Balasubramanian, *Nucleic Acids Res.*, 2005, **33**, 2908.
27. A. K. Todd, *Methods*, 2007, **43**, 246.
28. E. H. Blackburn, *Nature*, 1991, **350**, 569.
29. M. O'Reilly, S. A. Teichmann and D. Rhodes, *Curr. Opin. Struct. Biol.*, 1999, **9**, 56.
30. J.-L. Mergny, J.-F. Riou, P. Mailliet, M.-P. Teulade-Fichou and E. Gilson, *Nucleic Acids Res.*, 2002, **30**, 839.
31. S. Neidle and G. H. Parkinson, *Nat. Rev. Drug Discovery*, 2002, **1**, 383.
32. L. Oganesian and T. M. Bryan, *Bioessays*, 2007, **29**, 155.
33. D. J. Patel, A. T. Phan and V. Kuryavyi, *Nucleic Acids Res.*, 2007, **35**, 7429.
34. S. Neidle and G. H. Parkinson, *Curr. Opin. Struct. Biol.*, 2003, **13**, 275.
35. C. Schaffitzel, I. Berer, J. Postberg, J. Hanes, H. J. Lipps and A. Plückthun, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 8572.
36. K. Paeschke, T. Simonsson, J. Postberg, D. Rhodes and H. J. Lipps, *Nat. Struct. Mol. Biol.*, 2005, **12**, 847.
37. Y. Wang and D. J. Patel, *Structure*, 1993, **1**, 263.
38. G. H. Parkinson, M. P. H. Lee and S. Neidle, *Nature*, 2002, **417**, 876.
39. J. Dai, C. Punchihewa, A. Ambrus, D. Chen, R. A. Jones and D. Yang, *Nucleic Acids Res.*, 2007, **35**, 2440.
40. A. T. Phan, K. L. Luu and D. J. Patel, *Nucleic Acids Res.*, 2006, **34**, 5715.
41. J. Y. Lee, B. Okumus, D. S. Kim and T. Ha, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 18938.
42. T. S. Dexheimer, M. Fry and L. H. Hurley, DNA quadruplexes and gene regulation, in *Quadruplex Nucleic Acids*, ed. S Neidle and S Balasubramanian, Royal Society of Chemistry, Cambridge, UK, 2006.
43. A. Siddiqui-Jain, C. L. Grand, D. J. Bearss and L. H. Hurley, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 11593.
44. J. Seenisamy, E. M. Rezler, T. J. Powell, D. Tye, V. Gokhale, C. S. Joshi, A. Siddiqui-Jain and L. H. Hurley, *J. Am. Chem. Soc.*, 2004, **126**, 8702.
45. C. L. Grand, H. Han, R. M. Muñoz, S. Weitman, D. D. Von Hoff, L. H. Hurley and D. J. Bearss, *Mol. Cancer Ther.*, 2002, **1**, 565.
46. P. S. Shirude, B. Okumus, L. M. Ying, T. Ha and S. Balasubramanian, *J. Am. Chem. Soc.*, 2007, **129**, 7484.
47. K. Gehring, J. Leroy and M. Guéron, *Nature*, 1993, **363**, 499.
48. A. T. Phan and J.-L. Mergny, *Nucleic Acids Res.*, 2002, **30**, 4618.
49. J. L. Huppert and S. Balasubramanian, *Nucleic Acids Res.*, 2007, **35**, 406.
50. S. Kumari, A. Bugaut, J. L. Huppert and S. Balasubramanian, *Nat. Chem. Biol.*, 2007, **3**, 218.
51. M. L. Duquette, P. Handa, J. A. Vincent, A. F. Taylor and N. Maizels, *Genes Dev.*, 2004, **18**, 1618.
52. H. Han, D. R. Langley, A. Rangan and L. H. Hurley, *J. Am. Chem. Soc.*, 2001, **123**, 8902.
53. M. A. Read and S. Neidle, *Biochemistry*, 2000, **39**, 13422.
54. M.-Y. Kim, H. Vankayalapati, K. Shin-ya, K. Wierzba and L. H. Hurley, *J. Am. Chem. Soc.*, 2002, **124**, 2098.
55. A. De Cian, G. Cristofari, P. Reichenbach, E. De Lemos, D. Monchaud, M.-P. Teulade-Fichou, K. Shin-ya, L. Lacroix, J. Lingner and J.-L. Mergny, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 17347.
56. J. E. Redman, *Methods*, 2007, **43**, 302.
57. P. Ragazzon and J. B. Chaires, *Methods*, 2007, **43**, 313.
58. K. Jantos, R. Rodriguez, S. Ladame, P. S. Shirude and S. Balasubramanian, *J. Am. Chem. Soc.*, 2006, **128**, 13662.
59. K. P. Papadopoulos, D. W. Northfelt, D. M. Hufnagel, A. D. Ricart, P. P. Griffin, M. D. Oslund, D. D. Von Hoff, W. G. Rice, J. K. Lim and R. F. Marschke, *J. Clin. Oncol.*, 2007, **25**, 3585.
60. P. J. Bates, J. B. Kahlon, S. D. Thomas, J. O. Trent and D. M. Miller, *J. Biol. Chem.*, 1999, **274**, 26369.
61. M. Fry, *Front. Biosci.*, 2007, **12**, 4336.
62. M. Isalan, S. D. Patel, S. Balasubramanian and Y. Choo, *Biochemistry*, 2001, **40**, 830.