Supporting Information

G-quadruplex secondary structure from circular dichroism spectroscopy

Rafael del Villar-Guerra, John O. Trent* and Jonathan B. Chaires*
James Graham Brown Cancer Center
University of Louisville
505 S. Hancock St., Louisville, KY 40202 (USA)
* E-mail: j.chaires@louisville.edu, john.trent@louisville.edu

Abstract: A curated library of circular dichroism spectra of 23 G-quadruplexes of known structure was built and analyzed. The goal of this study was to use this reference library to develop an algorithm to derive quantitative estimates of the secondary structure content of quadruplexes from their experimental CD spectra. Principle component analysis and singular value decomposition were used to characterize the reference spectral library. CD spectra were successfully fit to obtain estimates of the amounts of base steps in *anti-anti*, *syn-anti* or *anti-syn* conformations, in diagonal or lateral loops or in other conformations. The results show that CD spectra of nucleic acids can be analyzed to obtain quantitative structural information about secondary structure content in an analogous way to methods used to analyze protein CD spectra.

Table of Contents

Experimental Procedures

1. REFERENCE G-QUADRUPLEX CIRCULAR DICHROISM SPECTRA

Oligonucleotide sequence, length, molecularity, topology and the Protein Data Bank code (ID PDB) for the dataset of 23 G-quadruplex DNA (G4-DNA) used in this work are given in Table S1. Oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA) with standard desalting and dissolved at a concentration of ~1mM. The buffer composition used for the folding and dilution of the oligonucleotides were identical to those reported for the structure determination (Table S1).

CD spectra were measured for each sequence, after appropriate annealing and sample preparation, under solution conditions identical to those used in the original structural determination using published protocols developed in our laboratory.[1] Sample homogeneity was confirmed by sedimentation velocity experiments as previously described.[2] In only one case (sample 186D) was significant heterogeneity observed, requiring additional purification by SEC (size exclusion chromatography) methods developed and reported.[3]

Circular dichroism (CD) and absorbance spectra were recorded in a 1-cm path length cuvette at 20 °C with a Jasco J-810 spectropolarimeter (Jasco, Easton, MD) equipped with a Peltier temperature controller with a concentration of 1-5 μM per quadruplex. The CD spectrum for the oligonucleotide 1I34 correspond to the spectra recorded at 4 ^{o}C and the spectrum for the oligonucleotide 186D correspond to the fraction with a sedimentation coefficient ($s_{20,w}$) of 1.9 obtained after purification by SEC (size exclusion chromatography).The absorbance at 260 nm of the samples used for CD experiments were ~ 0.7-1. The typical instrumental parameters to record the CD spectra were: 220-340 nm measurement range, 0.5 or 1 nm data pitch, 2 nm band width, 0.5 sec response, standard sensitivity, 100 or 200 nm/min of scanning speed. To obtain the CD reference spectra four spectra were averaged and the spectral contribution from the buffer was subtracted.
CD spectra were normalized to $\Delta\varepsilon$ ($M^{-1}\cdot cm^{-1}$) $=\theta/(32980*c*l)$ based on G-quadruplex strand concentration, where θ is the CD ellipticity in millidegrees, c is DNA concentration in mol/L, and l is the path length in cm. The DNA concentrations were calculated from the 260 nm absorbance using the molar extinction coefficient listed in Table S1. The molar concentration of DNA was expressed per quadruplex.

All the CD spectra were arranged as **C** matrix in the form that the each column contain the CD spectra of the corresponding G-quadruplex in the range of 220-330 nm at 1 nm increments. To obtain equal increments of 1 nm between points interpolation of some spectra was necessary.

2. DEFINITION OF SECONDARY STRUCTURE ELEMENTS OF QUADRUPLEX.

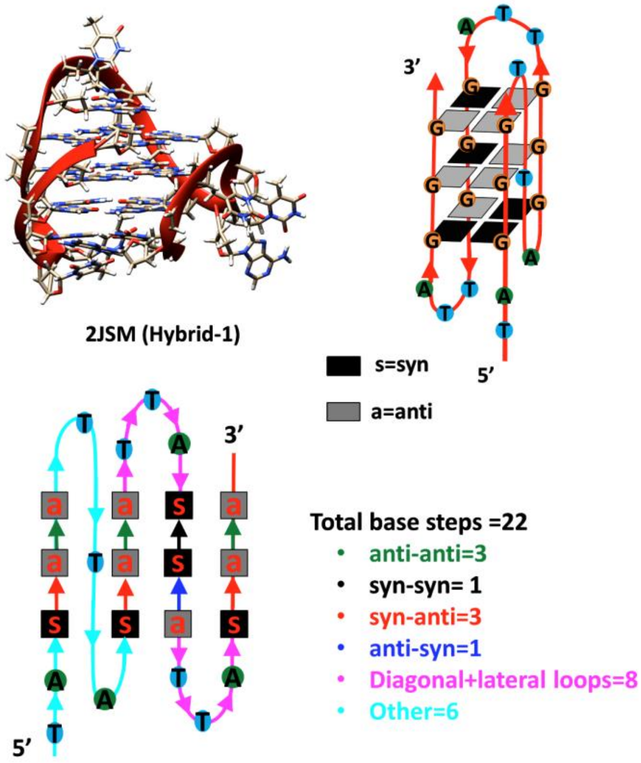


Figure S1. Schematic illustration for the definition of the secondary structure elements of G-quadruplex used in this study.

The fraction of secondary structural elements of G-quadruplex were determined by the visual inspection of the corresponding structures deposited in the PDB data bank (Table S4). These factions were calculated by normalizing the values to the total number of nucleotide base steps in their structures. In the case of NMR structures, the first model of the PDB file was used for the determination of the structural elements. Only the monomeric quadruplex structure observed in the asymmetric unit were considered for the structural analysis when more than one crystal structure was found (PDB files 1S45 and 2AVH). The quantification of the secondary structure elements was done following the geometric formalism for the description of DNA G-quadruplexes topologies previously described.[4] The schematic models of G-quadruplex aligned according to the frame of reference to obtain the fractions of the secondary structural elements are showed in Figure S1. The progression from 5' to 3' end of the first strand (blue line) of the G-quartets aligned according to the frame of reference was used as a criterion to define the polarity of the G-G stacking base steps.[4b] These secondary structural fractions define the structural matrix, **F**, that was used for the calculation of the G-quadruplex secondary structure basis CD spectra (Table S4).

3. PRINCIPAL COMPONENT ANALYSIS (PCA) AND CLUSTER ANALYSIS.

The multivariate data analysis, principal component and cluster analysis of the G-quadruplex CD spectral library was performed with R software [5], R version 3.3.2 (2016-10-31), using the R package called FactoMineR.[6] This software and the package are freely available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org. The CD spectra were scaled to unit variance before the analysis.Hierarchical clustering on principal components (HCPC)[7] was performed onto the first five principal components using the Euclidean distance as a measure of similarity between individuals, and the Ward criterion as agglomeration method. The initial clusters obtained by hierarchical clustering (HC) were further consolidated by partitional clustering using the *k*-means algorithm.

The results of the cluster analysis are presented as a dendrogram. In the dendrogram, the horizontal axis represents the PDB code of each G-quadruplex and the vertical axis represents the degree of similarity of the individuals. The individuals are colored according to their belonging to a cluster and the center of each cluster was represented by larger point size in the biplot graph. The selection of the optimal number of clusters was determined automatically using the gain in within inertia criterion.[7]

The v-test value of the wavelengths (variables) and the associated p-value was obtained as outputs of the function HCPC within statistical package FactoMineR.[6] The v-values where used as a tool to select the most statistically significant set of variables that are able to characterize each cluster (Figure S2).[8]

4. SINGULAR VALUE DECOMPOSITION (SVD)

The mathematical details of the SVD analysis have been already described. [9] For the SVD analysis, the CD spectra of our G-quadruplexes reference library were arranged in a matrix, **C**, of 23 columns. SVD decomposition of the matrix, **C**, was carried out using the software Matlab 7.1.0.246 (The MathWorks, Inc., 2005).

5. MINIMUM NUMBER OF BASIS SPECTRA

The original CD spectra matrix, **C**, was reconstructed by using only the **μ** most significant singular values and CD eigenvectors by $\mathbf{C_\mu = U_\mu S_\mu V_\mu{}^T}$, where $\mathbf{U_\mu}$, $\mathbf{S_\mu}$ and $\mathbf{V_\mu}$ are submatrices of **U, S** and **V**, respectively, corresponding to the first **μ** basis components. This linear combination $\mathbf{C_\mu = U_\mu S_\mu V_\mu{}^T}$ provides the best least-squares approximation to the matrix **C** having a rank **μ**. The number of significant basis spectra and the information contained in the CD spectral matrix, **C**, was determined by evaluating several statistical parameters such as: the relative magnitude of the singular values, their contribution to the total variance of the data set, the values of the autocorrelation function for the columns of the **U** and the **V** matrices, the variance in the dataset unaccounted ($\boldsymbol{\sigma}^2$) for reconstructing the original matrix (**C**),[10] and the spectral root mean squared (RMS) difference in comparison with the noise level of the CD spectra (see Figures S3-7 and Tables S2-3). The noise level of the CD spectra was determined by standard deviation in the range of 320-340 nm, where no CD signal was observed for any of the G-quadruplexes.

The contribution of each singular value to the total variance of the data set (relative variance, RV) for the first ten significant components of the CD spectra were calculated by Equation S1.[9b]

$$RV = \frac{S_i^2}{\sum_i S_i^2}$$

Equation S1

where S_i^2 is the square of the singular value. The relative importance of each column of **U**, as a component of the original protein CD spectra, can be estimated by the magnitude of the singular values.[10c]

The values of the autocorrelation function for the columns of the **U** and the **V** matrices were calculated as described previously.[9b]

The variance in the dataset unaccounted ($\boldsymbol{\sigma}^2$) for reconstructing the original matrix, $\mathbf{C}=\mathbf{U}_\mu\mathbf{S}_\mu\mathbf{V}_\mu^T$, were calculated by Equation S2, when only the $\boldsymbol{\mu}$ most significant basis CD spectra were used.[10b, c]

$$\sigma_\mu^2 = \left(\frac{1}{N(m-\mu)}\right)\sum_{i=\mu+1}^{m} s_i^2$$

Equation S2

where **C** is the matrix containing the reference CD spectra library of G-quadruplexes in its columns, $\boldsymbol{\sigma}^2$ is the variance unaccounted for reconstructing the matrix **C**, s_i is the i th singular value of S, $\boldsymbol{\mu}$ is the number of basis spectra used for the reconstruction of the original spectra, *m* is the number of reference spectra and *N* is the number of spectral data points.[10b]

6. QUANTITATIVE ANALYSIS CD SPECTRA OF G-QUADRUPLEX DNA: ESTIMATION OF SECONDARY STRUCTURE

Secondary structure CD spectra were calculated using singular value decomposition (SVD) and generalized inverse (Moore-Penrose pseudoinverse) methods on the matrix of the reference CD spectra of 23 G-quadruplex (**C)** with known secondary structures (**F**). The method to obtain secondary structure CD spectra of G-quadruplex is similar to that used with proteins and only the basic formulas of these methods will be described.[1].[1c]

The method assumes that the CD spectrum of a G-quadruplex can be represented as a linear combination of secondary structural basis spectra.

$$C_\lambda = \sum\nolimits_i f_i \cdot B_{\lambda,i} + noise$$

Equation S3

where C_λ is the CD spectrum of a G-quadruplex as function of the wavelength, f_i is the fraction of the ith secondary structural element, and $B_{\lambda,i}$ is the basis spectrum corresponding to the ith secondary structure element.

The set of linear equations can be represented in a matrix notation as **C=BF**, where **C** is the λx23 matrix of 23 G-quadruplex CD spectra, **B** is the λ×5 matrix of basis CD spectra for each secondary structure element expressed as column, and **F** is the 5×23 matrix of the fractions of each secondary structure element considered (Table S4).

The matrix equation relating the CD spectra to the secondary structure elements is $\mathbf{F{=}X_{\mu}C,}$ where $\mathbf{C}$ is the reference set of CD spectra of the known G-quadruplexes, $\mathbf{F}$ their corresponding known fractions of secondary structural elements (Table S4) , and $\mathbf{X_{\mu}}$ the unknown matrix of vectors which relate $\mathbf{C}$ and $\mathbf{F}$.[1e] The $\mathbf{X_{\mu}}$ matrix correspond to the generalized inverse of the basis CD spectra of the secondary structures elements calculated as $\mathbf{X_{\mu}{=}FC_{\mu}^{-1}}$. The inverse of the reference G-quadruplex CD spectra matrix, $\mathbf{C_{\mu}^{-1}}$, was calculated using SVD analysis with only the five ($\boldsymbol{\mu}$=5) most significant singular values and CD eigenvectors, $\mathbf{C_{\mu}^{-1}} = \mathbf{V_{\mu}S_{\mu}^{+}U_{\mu}^{T}}$, where $\mathbf{U_{\mu}}$**,** $\mathbf{S_{\mu}}$ and $\mathbf{V_{\mu}}$ are submatrices of **U, S** and **V**, respectively. Once the $\mathbf{X_{\mu}}$ matrix was obtained, $\mathbf{X_{\mu} = FC_{\mu}^{-1}} = \mathbf{FV_{\mu}S_{\mu}^{+}U_{\mu}^{T}}$, the fractions of secondary structure ($\mathbf{F_{calc}}$) corresponding to the test G-quadruplex CD spectrum ($\mathbf{C_{unk}}$) can be estimated.

Constrained least-square fitting of the test G-quadruplex CD spectrum ($\mathbf{C_{unk}}$) to five basis spectra (**B**) was used to predict the secondary structure fraction ($\mathbf{F_{calc}}$) such that

$$\min \left\| CD_{unk} - \sum_{i=1}^{5} f_{calc,i} \cdot B_{\lambda,i} \right\|^{2}$$

$$0.99 \leq \sum_{i=1}^{5} f_{calc,i} \leq 1.01$$

$$0.00 \leq f_{calc,i} \leq 1.00$$

Equation S4

where CD_{unk} is a column vector corresponding to the test G-quadruplex CD spectrum, $f_{calc,i}$ is the estimated fraction of the ith secondary structural element, and $B_{\lambda,i}$ is a column vector corresponding to the ith secondary structure basis spectra. The basis spectra corresponding to the secondary structure elements were calculated by $\mathbf{B}{=}\mathbf{X^{+}_{\mu}}$, (Figure S8) where the columns of the matrix **B** correspond to the five secondary structure basis CD spectra and $\mathbf{X^{+}_{\mu}}$ is the Moore-Penrose pseudoinverse of the matrix $\mathbf{X_{\mu}}$. The minimization process was performed with open source R software[2], R version 3.3.2 (2016-10-31), using the function "lsqlincon" from the R package "pracma" for solving quadratic programming problems in the presence of bound and linear constraints. The linear constraint used was that the sum of the secondary structures cannot be less than 0.99 or higher than 1.01. The bound constraints used was that the values of the secondary structures fractions cannot be negative or higher than 1.00.

The estimated fractions of each secondary structure ($\mathbf{F_{calc}}$) were used to reconstruct the test CD spectrum of the G-quadruplex by using the equation $\mathbf{C_{res}}{=}\ \mathbf{BF_{unk}}$, where $\mathbf{C_{res}}$ is the fitted CD spectrum, **B** is the matrix with the five secondary structure basis CD spectra, and $\mathbf{F_{calc}}$ is the predicted fraction of the secondary structure elements.[12]

The estimated secondary structure fraction reported in this work for all G-quadruplex of the reference library were calculated performing a leave-one-out cross-validation by using the five ($\boldsymbol{\mu}$=5) most significant secondary structure basis spectra (Table S5).

To carry out this leave-one-out cross-validation, the spectrum (**C_{unk}**) and the secondary structural elements (**F_{unk}**) of the test G-quadruplex were removed from the corresponding reference CD spectra matrix, **C**, and from the reference secondary structure fractions matrix, **F**. Then, the generalized inverse CD spectra matrix, **$X_\mu=FC_\mu^{-1}$**, was constructed using the remaining G-quadruplex CD spectra and the corresponding secondary structural element fractions. Once the **X_μ** matrix was obtained, constrained least-square fitting of the test G-quadruplex CD spectrum (**C_{unk}**) to five basis spectra (**B**), calculated using the generalized inverse, **B**=**X_μ^{-1}**, was performed**.** This procedure was repeated one after another for each test G-quadruplex in the data set.

The fitting of G-quadruplex CD spectra described here was implemented in the open source R software environment (https://www.r-project.org/)[2], R version 3.3.2 (2016-10-31),. Our script is available to interested users upon request.

7. ESTIMATION OF THE TOPOLOGICAL CLASS COMPOSITION.

The method to estimate the topological class composition of G-quadruplexes is similar to that describe previously for the estimation of secondary structure. However, in this case, three the basis spectra were obtained using the tertiary structure matrix fractions showed in Table S7.The results of the estimation of the tertiary structure by leave-one-out cross-validation constrained least-squares fitting of the CD spectra of G-quadruplex are shown in Figure S11.

8. ACCURACY OF THE PREDICTION METHOD.

The structural root mean squared difference (RMSD $_{strcutural}$) (Equation S1), the statistical ξ parameter,[13]the slope and the Pearson correlation coefficient (r) between the predicted secondary structural elements and the experimental values determined from the PDB structures was used to evaluate the quality of the prediction method of the structural parameter.

The spectral and/or structural root mean squared (RMSD) difference was calculated by Equation 3 between experimental and the predicted values.

$$RMSD = \sqrt[2]{\left(\frac{\sum_i^N \left(x_{i,experimental} - x_{i,predicted}\right)^2}{N}\right)}$$

Equation S5

where the summation in Equation S5 extends over the number of spectral data points ($N=N_\lambda$) when two CD spectra are compared, and over the number of secondary structure classes ($N=N_{class}$) when the structural RMS is calculated. In the case of the structural RMSD, $x_{i,experimental}$ is the fraction of secondary structure calculated from the PDB structure and $x_{i,predicted}$ is the fraction of secondary structure calculated from CD.

The statistical ξ parameter ($\xi = \sigma_{PDB}/RMSD_{structural}$) was calculated as the standard deviation for a particular secondary structural parameter obtained from the different values observed in the PDB file (σ_{PDB}) divided by the structural root mean squared difference of the parameter (RMSD). [13]

The goodness of the fitting was calculated by spectral root-mean-square deviation (RMSD, Equation S6) and the normalized root-mean-square deviation (NRMSD) parameter defined by Equation S7 [14]

$$RMSD = \sqrt[2]{\frac{\sum_{N}(CD_{exp} - CD_{calc})^2}{N}}$$

Equation S6

$$NRMSD = \frac{RMSD}{CD_{exp,max} - CD_{exp,min}}$$

Equation S7

where C_{exp} is the experimental spectrum being evaluated, CD_{calc} is the calculated spectrum derived from the estimated secondary structure fractions, N is the number of data points in a spectrum, $CD_{exp,max}$ and $CD_{exp,min}$ are the largest and smallest intensity observed in the experimental spectrum being evaluated. A low value for the NRMSD (>0.1) indicates a small error between the experimental CD spectra and the fitted CD spectra, derived the from predicted secondary structure fractions structure. However, a low NRMSD alone does not mean that an analysis is accurate. [15] Therefore, the selection of the best set of secondary structural parameter was based on the overall performance values of these statistical parameters.

9. TABLES AND FIGURES

Table S1. Reference DNA G-quadruplex oligonucleotides dataset used in the present study.

PDB ID	Sequence	ε260 (M^{-1} cm^{-1}) per quadruplex	length	Molecularity	Method	Strand and loop topology	Ref.
1EMQ	5'-TGGTGGC-3'	203805	28	tetramolecular	NMR	Parallel (4+0)	[16]
1EVM	5'-AGGGT-3'	164564	20	tetramolecular	NMR	Parallel (4+0)	[17]
1FQP	5'-GGGTTTTGGG-3'	173190	20	bimolecular	NMR	Antiparallel (2+2) 2 diagonal loops	[18]
1I34	5'-GGTTTTGGCAGGGTTTTGGT-3'	154020	20	unimolecular	NMR	Antiparallel (2+2) 1 propeller loop 2 diagonal loops	[19]
1LVS	5'-GGGGTTTTGGG-3'	183378	22	bimolecular	NMR	Antiparallel (2+2) 2 diagonal loops	[20]
1NP9	5'-TTAGGGT-3'	241628	28	tetramolecular	NMR	Parallel (4+0)	[21]
1XAV	5'-TGAGGGTGGGTAGGGTGGGTAA-3'	190394	22	unimolecular	NMR	Parallel (4+0) 3 propeller loops	[22]
139D	5'-TTGGGGT-3'	247466	28	tetramolecular	NMR	Parallel (4+0)	[23]
148D	5'-GGTTGGTGTGGTTGG-3'	142023	15	unimolecular	NMR	Antiparallel (2+2) 3 lateral loops	[24]
186D	5'-TTGGGGTTGGGGTTGGGGTTGGGG-3'	193875	24	unimolecular	NMR	Hybrid-2 (3+1) 2 lateral loops 1 propeller loop	[25]
201D	5'-GGGGTTTTGGGGTTTTGGGGTTTTGGGG-3'	241822	28	unimolecular	NMR	Antiparallel (2+2) 1 diagonal loop 2 lateral loops	[26]
156D	5'-GGGGTTTTGGGG-3'	182212	24	bimolecular	NMR	Antiparallel (2+2) 2 diagonal loop	[27]
1S45*	5'-TGGGGT-3'	199347	24	tetramolecular	X-Ray	Parallel (4+0)	[28]
2AVH**	5'-GGGGTTTGGGG-3'	193611	22	bimolecular	X-RAY	Antiparallel (2+2) 2 lateral loops	[29]
2GKU	5'-TTGGGTTAGGGTTAGGGTTAGGGA-3'	244300[30]	24	unimolecular	NMR	Hybrid-1 (3+1) 1 propeller loop 2 lateral loops	[31]
2HY9	5'-AAAGGGTTAGGGTTAGGGTTAGGGAA-3'	278200[30]	26	unimolecular	NMR	Hybrid-1 (3+1) 1 propeller loop 2 lateral loops	[32]
2JSM	5'-TAGGGTTAGGGTTAGGGTTAGGG-3'	236500[30]	23	unimolecular	NMR	Hybrid-1 (3+1) 1 propeller loop 2 lateral loops	[33]
2JPZ	5'-TTAGGGTTAGGGTTAGGGTTAGGGTT-3'	261200[30]	26	unimolecular	NMR	Hybrid-2 (3+1) 2 lateral loops 1 propeller loop	[34]
2JSL	5'-TAGGGTTAGGGTTAGGGTTAGGGTT-3'	253100[30]	25	unimolecular	NMR	Hybrid-2 (3+1) 2 lateral loops 1 propeller loop	[33]
2KF8	5'-GGGTTAGGGTTAGGGTTAGGGT-3'	223500[30]	22	unimolecular	NMR	Hybrid-3Antiparallel (2+2) 1 diagonal loop 2 lateral loops	[35]
2KKA***	5’-AGGGTTAGGGTTAGGGTTAGGGT-3’	237000[30]	23	unimolecular	NMR	Hybrid-3 Antiparallel (2+2) 1 diagonal loop 2 lateral loops	[36]
2LD8	5'-TAGGGTTAGGGTTAGGGTTAGGG-3'	236500	23	unimolecular	NMR	Parallel (4+0) 3 propeller loops	[37]
143D	5'-AGGGTTAGGGTTAGGGTTAGGG-3'	228500	22	unimolecular	NMR	Antiparallel (2+2) 1 diagonal loop 2 lateral loops	[38]

* Two PDB files 1S45 and 1S47 were found for this oligonucleotide sequence. Only the monomeric quadruplex structures observed in the asymmetric unit of the file 1S45 was considered in this work.

** Two PDB files 2AVH and 2AVJ were found for this oligonucleotide sequence. Only the monomeric quadruplex structures observed in the asymmetric unit of the file 2AVH was considered in this work.

*** The PDB file of this oligonucleotide present inosine base (I).For recording the CD spectra the iosine (I) was replaced for a guanine. (G).

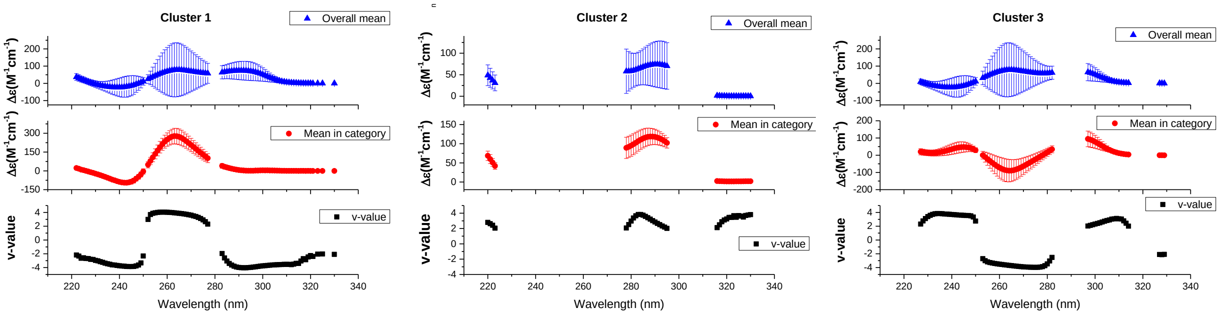


Figure S2. Representation of the average CD signal intensity for the whole data set (overall mean, blue), the average of the CD signal intensity in the cluster (mean in the cluster, red) and the v-values (black) calculated for the most significant wavelengths of clusters 1, 2 and 3 obtained from the agglomerative hierarchical clustering analysis (HCA) performed onto the first five principal components (PCs) of the CD spectra data of 23 reference G-quadruplex oligonucleotides. The associated standard deviations are represented by error bars.

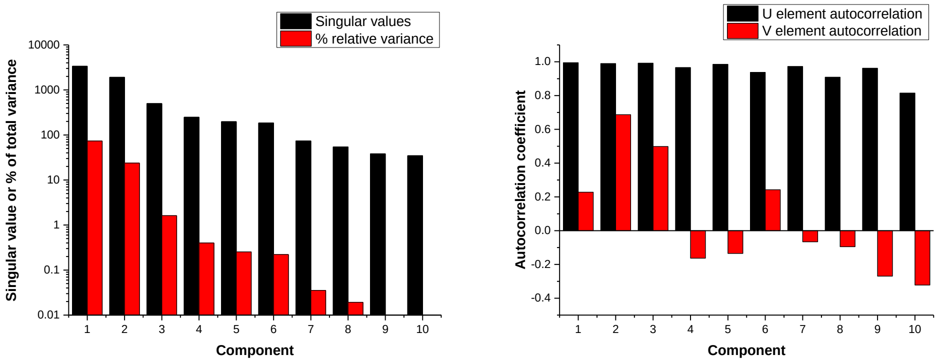


Figure S3 . Bar graphs showing the singular values and their relative total variance for the first ten components (left) and the autocorrelation coefficients (right) estimated for the first ten significant components of the U and V matrices after performing SVD analysis on the CD spectra matrix (**C**).

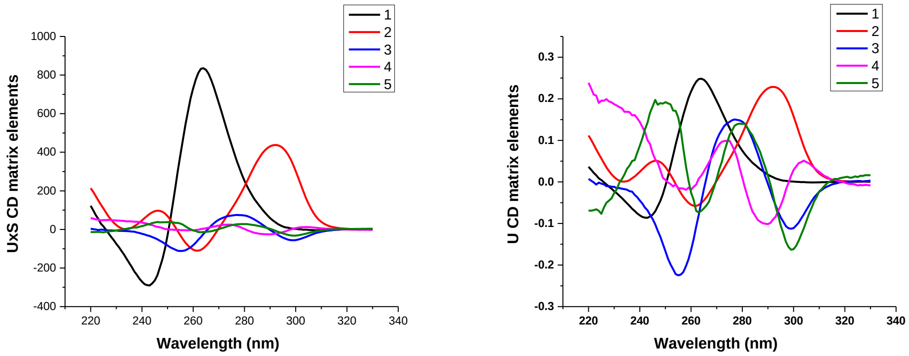


Figure S4. Five most significant basis CD spectra generated after performing SVD analysis on the CD spectra matrix (**C**) of the 23 G-quadruplexes.

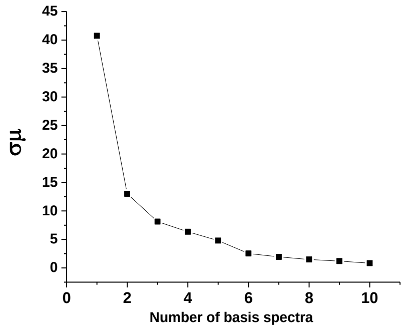


Figure S5. The standard deviation σ_μ unaccounted in the reconstruction of the CD spectra matrix (**C**) of the 23 G-quadruplexes for the first ten significant components.

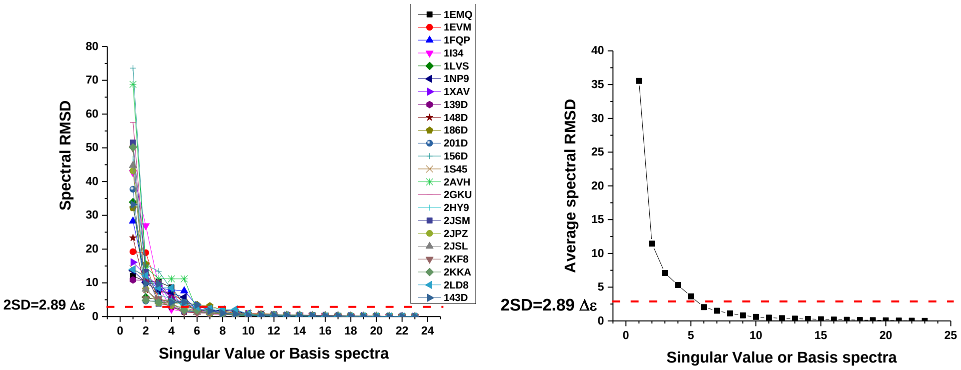
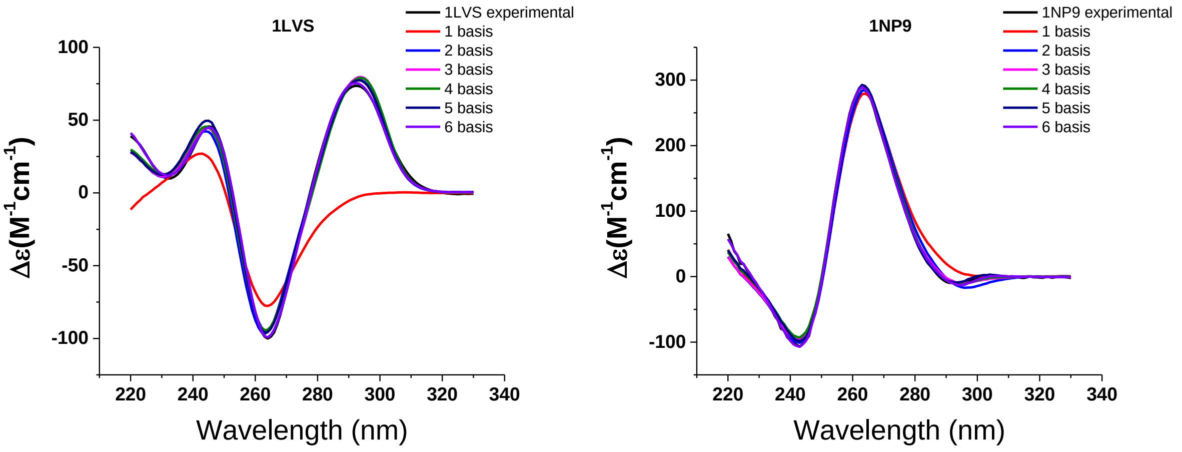


Figure S6. Individual (left) and average (right) spectral RMS (root mean squared difference) between reconstructed and original CD spectra as a function of the number of basis spectra or singular values. The dash red line correspond to two standard deviations of the error level (2SD=2.89 Δε units) determinated in the range of 330nm-340nm.



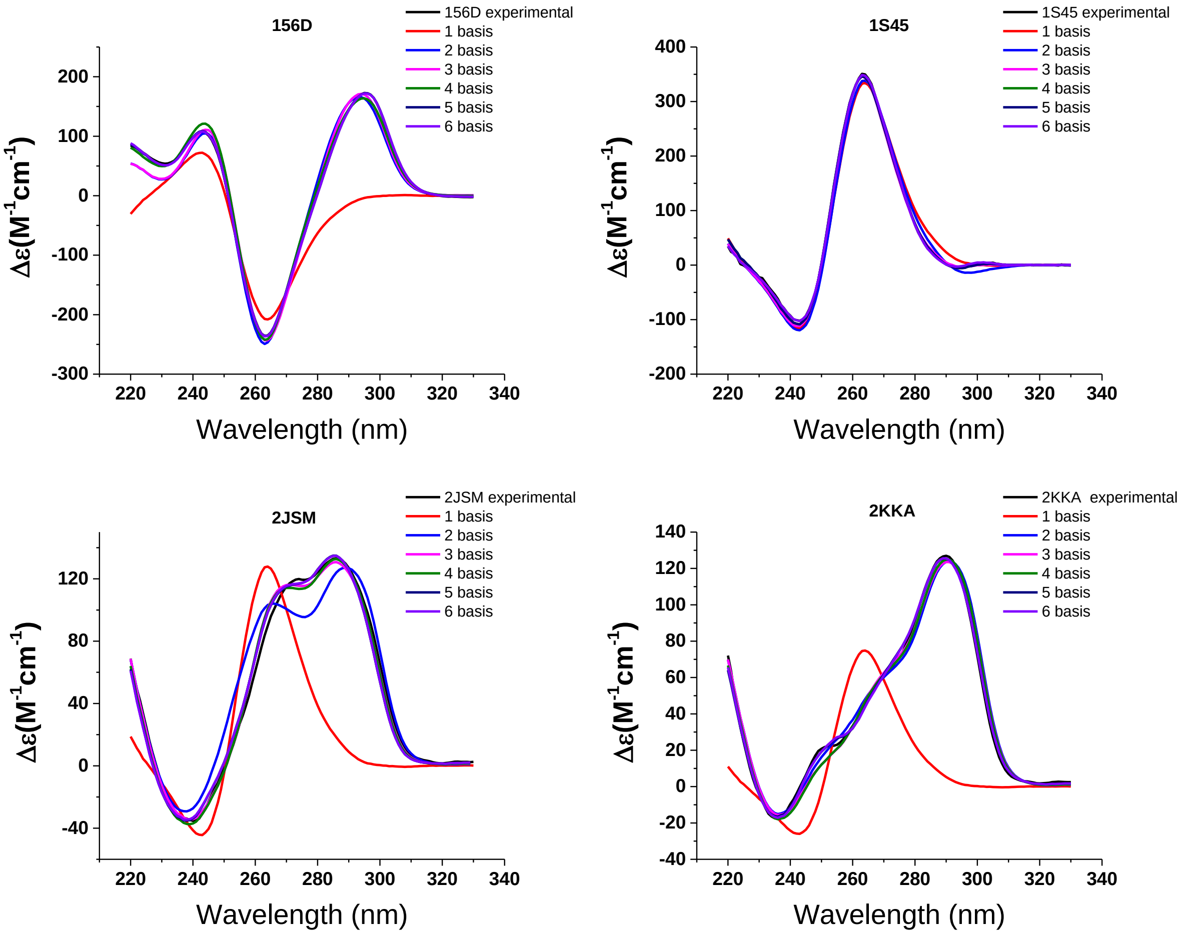


Figure S7. Reconstructed CD spectra of G-quadruplexes 1S45, 1I34, 2JPZ, 201D and 1FQP with various numbers of basis spectra: measured CD spectra (green), 1 basis (red), 2 basis (blue), 3 basis (pink), 4 basis (orange), 5 basis (black) and 6 basis (violet). The intensity of the CD spectra was expressed as Δε (M^{-1} cm^{-1}) = θ(mdeg)/(32980*c(mol/Liter)*L(cm)) where c= concentration per quadruplex in mol/Liter.

Table S2. The largest ten singular values, relative variance of each singular value and the standard deviation σ_μ

Component	S singular values	% relative variance	σ_μ
1	3364.16	73.62	40.75
2	1913.16	23.81	13.01
3	497.97	1.61	8.13
4	248.10	0.40	6.35
5	197.46	0.25	4.80
6	184.46	0.22	2.53
7	73.74	0.04	1.93
8	54.38	0.02	1.48
9	38.28	0.01	1.19
10	34.63	0.01	0.83

Table S3. The autocorrelation coefficients estimated for the first ten significant components of the U and V matrices after performing SVD analysis on the CD spectra matrix (**C**) of the 23 G-quadruplexes.

Component	U element autocorrelation	V element autocorrelation
1	0.994455	0.227678
2	0.989543	0.686918
3	0.991845	0.498414
4	0.966099	-0.16311
5	0.985352	-0.13485
6	0.937543	0.242124
7	0.972398	-0.0659
8	0.908838	-0.09494
9	0.962054	-0.26936
10	0.814879	-0.32183

Table S4. Fractions of the secondary structural elements of the 23 G-quaduplex used in this work. The values were obtained from the visual inspection of the PDB files. The fractions were calculated by normalizing to the total number of base steps in the structure. The base steps G-G of the guanines localized in the loops or the flanking regions were taking into account.

PDB ID	anti-anti	syn-anti	anti-syn	Diagonal + lateral loops	other	Total base steps
1EMQ	0.33	0.00	0.00	0.00	0.67	24
1EVM	0.50	0.00	0.00	0.00	0.50	16
1FQP	0.11	0.11	0.11	0.56	0.11	18
1I34	0.05	0.11	0.11	0.53	0.21	19
1LVS	0.00	0.25	0.20	0.50	0.05	20
1NP9	0.33	0.00	0.00	0.00	0.67	24
1XAV	0.38	0.00	0.00	0.00	0.62	21
139D	0.50	0.00	0.00	0.00	0.50	24
148D	0.07	0.14	0.14	0.57	0.07	14
186D	0.22	0.13	0.09	0.26	0.30	23
201D	0.00	0.22	0.22	0.56	0.00	27
156D	0.00	0.27	0.27	0.45	0.00	22
1S45	0.60	0.00	0.00	0.00	0.40	20
2AVH	0.00	0.30	0.30	0.40	0.00	20
2GKU	0.13	0.13	0.04	0.35	0.35	23
2HY9	0.12	0.12	0.04	0.32	0.40	25
2JSM	0.14	0.14	0.05	0.36	0.32	22
2JPZ	0.12	0.12	0.04	0.32	0.40	25
2JSL	0.13	0.13	0.04	0.33	0.38	24
2KF8	0.14	0.10	0.14	0.57	0.05	21
2KKA	0.14	0.09	0.09	0.55	0.14	22
2LD8	0.36	0.00	0.00	0.00	0.64	22
143D	0.00	0.19	0.19	0.57	0.05	21

Table S5. Predicted fractions of secondary structural elements, anti-anti, syn-anti, anti-syn,diagonal+lateral loops and other , by leave-one-out cross-validation constrained least-squares fitting of the CD spectra of G-quadruplex. The goodness of the prediction for each quadruplex was evaluated by structural RMSD, spectral RMSD and the spectral NRMSD .

	anti-anti	**syn-anti**	**anti-syn**	**Diagonal + lateral loops**	**Other**	**Total**	**Structural RMSD**	**Spectral NRMSD**
1EMQ	0.38	0.01	0.00	0.08	0.52	0.99	0.08	0.04
1EVM	0.40	0.04	0.06	0.10	0.39	0.99	0.08	0.03
1FQP	0.14	0.14	0.14	0.30	0.27	0.99	0.14	0.11
1I34	0.09	0.20	0.24	0.48	0.00	1.01	0.12	0.02
1LVS	0.06	0.17	0.17	0.39	0.21	0.99	0.10	0.07
1NP9	0.48	0.00	0.00	0.00	0.51	0.99	0.10	0.02
1XAV	0.40	0.00	0.02	0.02	0.55	0.99	0.04	0.04
139D	0.46	0.04	0.00	0.00	0.51	1.01	0.03	0.08
148D	0.15	0.12	0.12	0.31	0.29	0.99	0.16	0.15
186D	0.22	0.12	0.12	0.36	0.19	1.01	0.07	0.07
201D	0.08	0.16	0.15	0.36	0.24	0.99	0.15	0.06
156D	0.00	0.23	0.26	0.52	0.00	1.01	0.04	0.10
1S45	0.49	0.00	0.00	0.00	0.52	1.01	0.07	0.02
2AVH	0.00	0.21	0.17	0.64	0.00	1.01	0.13	0.06
2GKU	0.06	0.14	0.08	0.47	0.26	1.01	0.08	0.03
2HY9	0.16	0.14	0.11	0.27	0.31	0.99	0.06	0.03
2JSM	0.11	0.13	0.07	0.41	0.27	0.99	0.04	0.01
2JPZ	0.17	0.12	0.07	0.36	0.30	1.01	0.06	0.02
2JSL	0.14	0.12	0.08	0.39	0.25	0.99	0.07	0.02
2KF8	0.10	0.15	0.09	0.47	0.20	1.01	0.09	0.03
2KKA	0.11	0.16	0.11	0.45	0.18	1.01	0.06	0.02
2LD8	0.34	0.03	0.00	0.11	0.51	0.99	0.08	0.03
143D	0.09	0.17	0.18	0.32	0.23	0.99	0.14	0.10

Table S6. Overall performance of the prediction method for the secondary structural elements of G-quadruplex, anti-anti, syn-anti, anti-syn, diagonal+lateral loops and other predicted by leave-one-out cross-validation constrained least-squares fitting of the CD spectra of G-quadruplex.

	anti-anti	**syn-anti**	**anti-syn**	**Diagonal + lateral loops**	**other**
Structural std of PDB (σ_{PDB})	0.180	0.092	0.093	0.231	0.233
mean PDB	0.190	0.111	0.090	0.313	0.297
ξ parameter[10b]	2.91	2.13	1.84	1.77	1.75
Structural RMSD	0.06	0.04	0.05	0.13	0.13
R Pearson	0.94	0.88	0.83	0.82	0.82
Slope	1.06	1.14	1.02	1.00	1.13

σ_{PDB}= structural standard deviation observed in the PDB files. $\xi = \sigma_{PDB}$/RMSD structural. [10b]

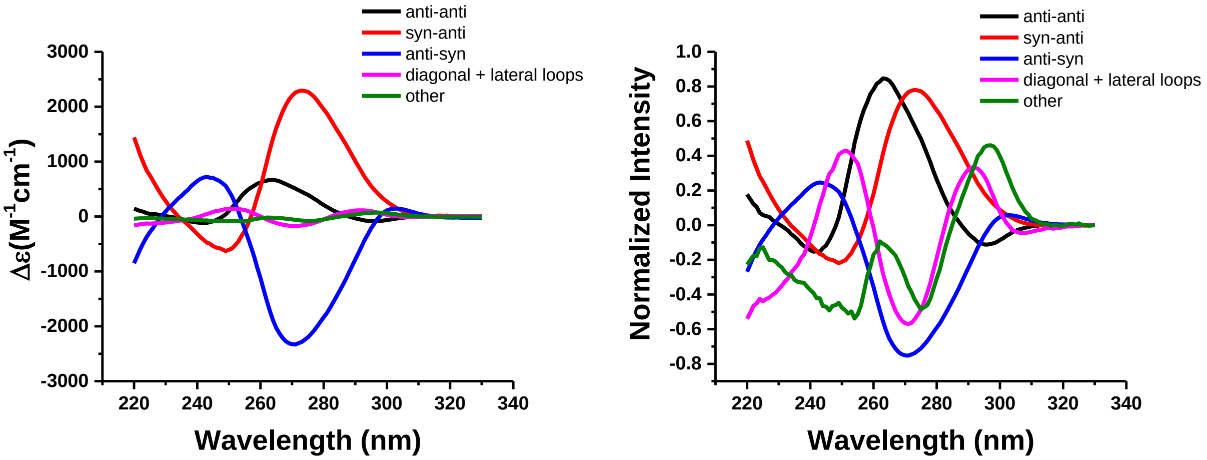


Figure S8. Secondary structure basis components CD spectra of G-quadruplex used for the fitting to obtain the secondary structural elements anti-anti, syn-anti, anti-syn, diagonal and lateral loops A) The intensity of the CD spectra was expressed as Δε (M^{-1} cm^{-1}) = θ(mdeg)/(32980*c(mol/Liter)*L(cm)) where c= concentration per quadruplex in mol/Liter. B) Normalized basis spectra.

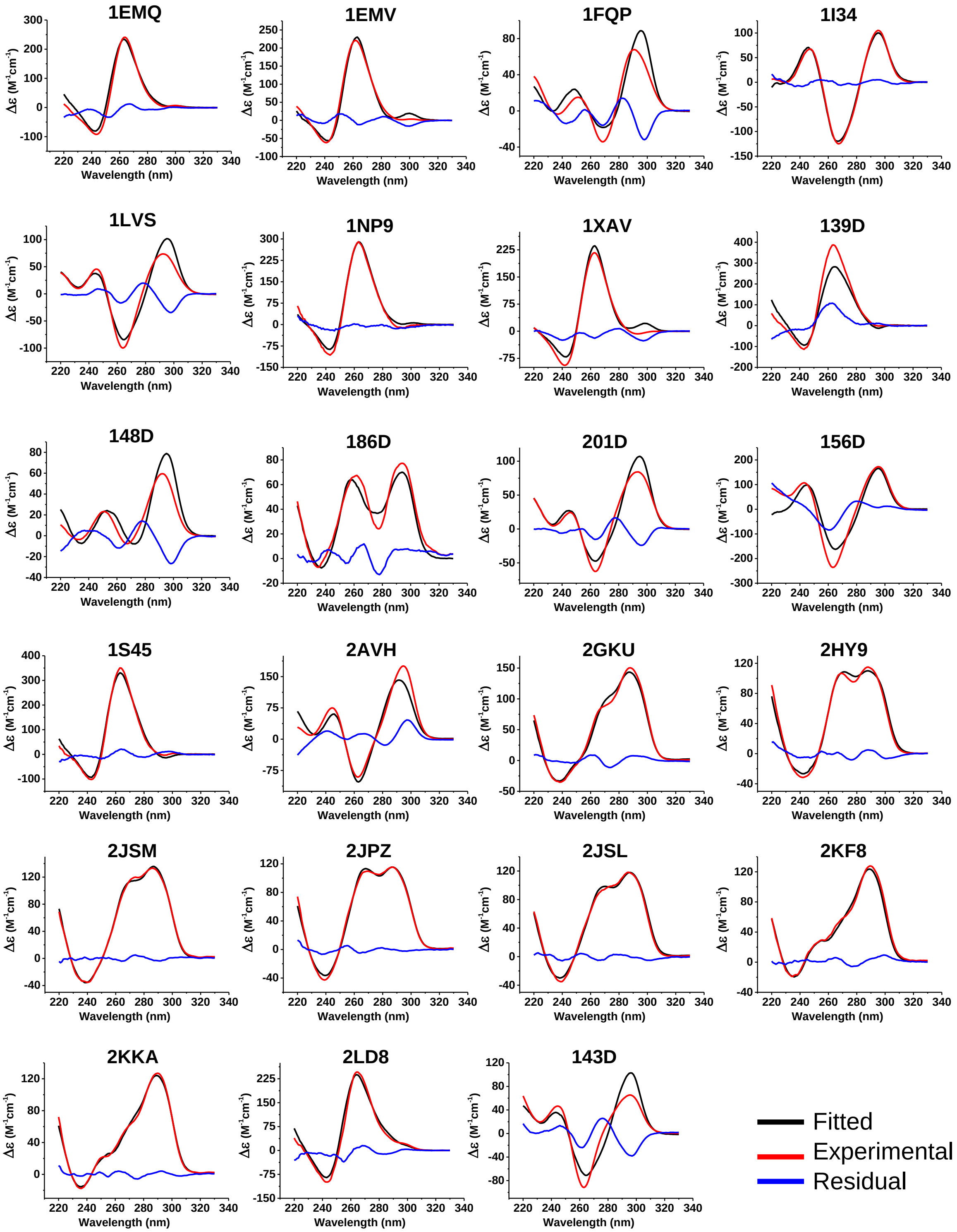


Figure S9. Experimental (red), reconstructed (black) and residual (blue) CD spectra for the 23 reference G-quadruplex obtained by leave-one-out cross-validation constrained least-squares fitting. The reconstructed CD spectra were calculated from the estimated fractions for the set of secondary structures elements 'anti-anti', 'syn-anti', 'anti-syn', 'diagonal and lateral loops' and 'other'. The intensity of the CD spectra (330-220 nm) is expressed as Δε (M^{-1} cm^{-1}) per quadruplex.

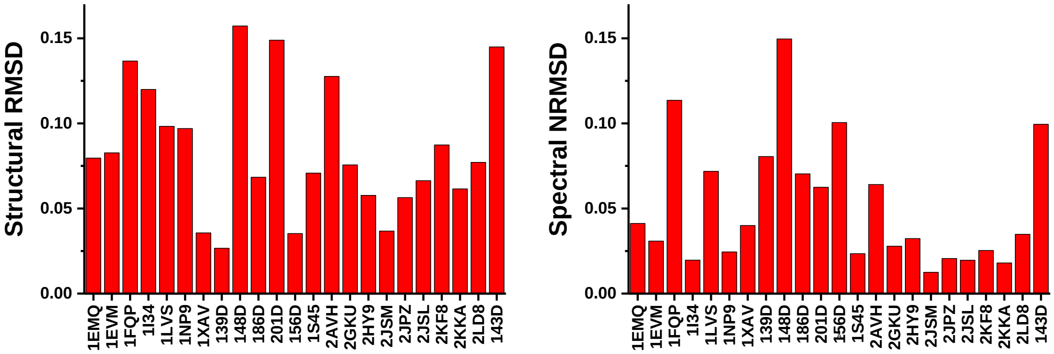


Figure S10. Quality of the prediction when the secondary structural elements 'anti-anti', 'syn-anti', 'anti-syn', 'diagonal+lateral loops' and 'other' were used for the 23 reference G-quadruplex. The structural RMSD and the spectral NRMSD were obtained by leave-one-out cross-validation using a constrained least-squares fitting.

Table S7. Fractions of tertiary structure matrix, **F,** used to predicted the tertiary structure of G-quadruplex by leave-one-out cross-validation constrained least-squares fitting of the CD spectra

PDB ID	**cluster 1**	**cluster 2**	**cluster 3**
	Parallel	**Hybrid**	**Antiparallel**
1EMQ	1	0	0
1EVM	1	0	0
1FQP	0	0	1
1I34	0	0	1
1LVS	0	0	1
1NP9	1	0	0
1XAV	1	0	0
139D	1	0	0
148D	0	0	1
186D	0	1	0
201D	0	0	1
156D	0	0	1
1S45	1	0	0
2AVH	0	0	1
2GKU	0	1	0
2HY9	0	1	0
2JSM	0	1	0
2JPZ	0	1	0
2JSL	0	1	0
2KF8	0	1	0
2KKA	0	1	0
2LD8	1	0	0
143D	0	0	1

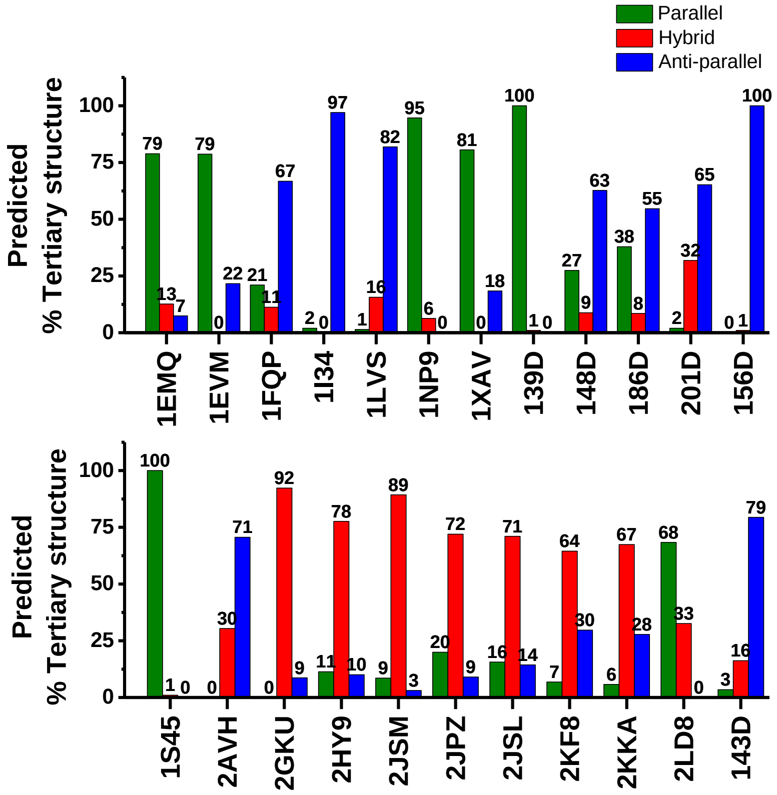


Figure S11. Predicted fractions of the tertiary structure or the reference G-quadruplex CD spectra library by leave-one-out cross-validation constrained least-squares fitting.

10. References

[1] R. del Villar-Guerra, R. D. Gray and J. B. Chaires, *Curr. Protoc. Nucleic Acid Chem. 68:17.8.1-17.8.16.* **2017**.
[2] a) N. C. Garbett, C. S. Mekmaysy and J. B. Chaires in *Sedimentation Velocity Ultracentrifugation Analysis for Hydrodynamic Characterization of G-Quadruplex Structures*, (Ed. P. Baumann), Humana Press, Totowa, NJ, **2010**, pp. 97-120; b) J. B. Chaires, W. L. Dean, H. T. Le and J. O. Trent in *Chapter Thirteen - Hydrodynamic Models of G-Quadruplex Structures*, *Vol. Volume 562* (Ed. L. C. James), Academic Press, **2015**, pp. 287-304.
[3] a) M. C. Miller, C. J. Ohrenberg, A. Kuttan and J. O. Trent, *Curr. Protoc. Nucleic Acid Chem. 61:17.7.1-17.7.18.* **2015**; b) M. C. Miller and J. O. Trent, *Curr. Protoc. Nucleic Acid Chem. 45:17.3.1-17.3.18* **2011**.
[4] a) A. I. Karsisiotis, C. O'Kane and M. Webba da Silva, *Methods* **2013**, *64*, 28-35; b) M. Webba da Silva, *Chemistry A-European Journal* **2007**, *13*, 9738-9745; c) S. Burge, G. N. Parkinson, P. Hazel, A. K. Todd and S. Neidle, *Nucleic Acids Research* **2006**, *34*, 5402-5415.
[5] *R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/*.
[6] S. Le, J. Josse and F. Husson, *Journal of Statistical Software* **2008**, *25*, 1-18.
[7] F. Husson, J. Josse and J. Pagès, *Principal component methods–hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?. Technical Report–Agrocampus (2010)* **2010**.
[8] T. Feuillet, D. Mercier, A. Decaulne and E. Cossart, *Geomorphology* **2012**, *139–140*, 577-587.
[9] a) L. A. Compton and W. C. Johnson Jr, *Analytical Biochemistry* **1986**, *155*, 155-167; b) R. D. Gray and J. B. Chaires in *Analysis of Multidimensional G-Quadruplex Melting Curves*, John Wiley & Sons, Inc.: Hoboken, NJ, **2011; Chapter 17, Unit 17 4**.
[10] a) P. Manavalan and W. C. Johnson, *Journal of Biosciences* **1985**, *8*, 141-149; b) J. G. Lees, A. J. Miles, F. Wien and B. A. Wallace, *Bioinformatics* **2006**, *22*, 1955-1962; c) T. Konno, *Protein Sci.* **1998**, *7*, 975-982.
[11] a) J. P. Hennessey and W. C. Johnson, *Biochemistry* **1981**, *20*, 1085-1094; b) N. Sreerama, S. Y. Venyaminov and R. W. Woody, *Protein Sci.* **1999**, *8*, 370-380; c) L. A. Compton, C. K. Mathews and W. C. Johnson, *Journal of Biological Chemistry* **1987**, *262*, 13039-13043.
[12] J. P. Hennessey and G. A. Scarborough, *Journal of Biological Chemistry* **1988**, *263*, 3123-3130.
[13] K. A. Oberg, J.-M. Ruysschaert and E. Goormaghtigh, *European Journal of Biochemistry* **2004**, *271*, 2937-2948.
[14] V. Hall, M. Sklepari and A. Rodger, *Chirality* **2014**, *26*, 471-482.
[15] L. Whitmore and B. A. Wallace, *Nucleic Acids Research* **2004**, *32*, W668-W673.
[16] P. K. Patel and R. V. Hosur, *Nucleic Acids Research* **1999**, *27*, 2457-2464.
[17] P. K. Patel, A. S. R. Koti and R. V. Hosur, *Nucleic Acids Research* **1999**, *27*, 3836-3843.
[18] M. A. Keniry, G. D. Strahan, E. A. Owen and R. H. Shafer, *European Journal of Biochemistry* **1995**, *233*, 631-643.
[19] V. Kuryavyi, A. Majumdar, A. Shallop, N. Chernichenko, E. Skripkin, R. Jones and D. J. Patel, *Journal of Molecular Biology* **2001**, *310*, 181-194.
[20] M. Črnugelj, N. V. Hud and J. Plavec, *Journal of Molecular Biology* **2002**, *320*, 911-924.
[21] E. Gavathiotis and M. S. Searle, *Organic & Biomolecular Chemistry* **2003**, *1*, 1650-1656.
[22] A. Ambrus, D. Chen, J. Dai, R. A. Jones and D. Yang, *Biochemistry* **2005**, *44*, 2048-2058.
[23] Y. Wang and D. J. Patel, *Journal of Molecular Biology* **1993**, *234*, 1171-1183.
[24] P. Schultze, R. F. Macaya and J. Feigon, *Journal of Molecular Biology* **1994**, *235*, 1532-1547.
[25] Y. Wang and D. J. Patel, *Structure* **1994**, *2*, 1141-1156.
[26] Y. Wang and D. J. Patel, *Journal of Molecular Biology* **1995**, *251*, 76-94.
[27] P. Schultze, F. W. Smith and J. Feigon, *Structure* **1994**, *2*, 221-233.
[28] C. Cáceres, G. Wright, C. Gouyette, G. Parkinson and J. A. Subirana, *Nucleic Acids Research* **2004**, *32*, 1097-1102.
[29] P. Hazel, G. N. Parkinson and S. Neidle, *Journal of the American Chemical Society* **2006**, *128*, 5480-5487.
[30] R. Buscaglia, R. D. Gray and J. B. Chaires, *Biopolymers* **2013**, *99*, 1006-1018.
[31] K. N. Luu, A. T. Phan, V. Kuryavyi, L. Lacroix and D. J. Patel, *Journal of the American Chemical Society* **2006**, *128*, 9963-9970.
[32] J. Dai, C. Punchihewa, A. Ambrus, D. Chen, R. A. Jones and D. Yang, *Nucleic Acids Research* **2007**, *35*, 2440-2450.
[33] A. T. Phan, V. Kuryavyi, K. N. Luu and D. J. Patel, *Nucleic Acids Research* **2007**, *35*, 6517-6525.
[34] J. Dai, M. Carver, C. Punchihewa, R. A. Jones and D. Yang, *Nucleic Acids Research* **2007**, *35*, 4927-4940.
[35] K. W. Lim, S. Amrane, S. Bouaziz, W. Xu, Y. Mu, D. J. Patel, K. N. Luu and A. T. Phan, *Journal of the American Chemical Society* **2009**, *131*, 4301-4309.
[36] Z. Zhang, J. Dai, E. Veliath, R. A. Jones and D. Yang, *Nucleic Acids Research* **2010**, *38*, 1009-1021.
[37] B. Heddi and A. T. Phan, *Journal of the American Chemical Society* **2011**, *133*, 9824-9833.
[38] Y. Wang and D. J. Patel, *Structure* **1993**, *1*, 263-282.

Author Contributions

R.V-G. designed experiments, collected data, analyzed data, wrote the analysis program code and wrote the manuscript. J.O.T. and J.B.C administered the project, designed experiments, analyzed data and wrote the manuscript.