COMPUTATIONAL IDENTIFICATION OF G-QUADRUPLEXES SECONDARY STRUCTURES

TUGAY DİREK

Master’s Thesis

Graduate School
İzmir University of Economics
İzmir
2022

COMPUTATIONAL IDENTIFICATION OF G-QUADRUPLEXES SECONDARY STRUCTURES

TUGAY DİREK

A Thesis Submitted to
The Graduate School of Izmir University of Economics
Master’ s Program in Computer Engineering

İzmir
2022

ABSTRACT

COMPUTATIONAL IDENTIFICATION OF G-QUADRUPLEXES SECONDARY STRUCTURES

Direk, Tugay

Master’ s Program Computer Engineering

Advisor: Assoc. Prof. Dr. Osman Doluca

July, 2022

Establishment of a standard for classification of G-quadruplexes has been evaded for a long time. The situation became even more complex, with discovery of bulged G-tracts and mismatched G-tetrads. For this reason there has been a very limited number of studies aiming to bring forth a standard to define G4 secondary structures. In this study, we propose a new method for the identification of secondary structures of G-quadruplexes based on three-dimensional structural data. Briefly, coordinates of guanines are processed to identify tetrads and loops. Then, we present the secondary structure in the form of a figure which shows the loop types and guanines that participate in each tetrad. Additionally, ONZ classification based on topology-based classification of tetrads and quadruplex structures is implemented and the results of our study are compared.

Keywords: G-quadruplex, Secondary structures, 3D Structures, Loop, Tetrad, Classification of G-quadruplexes

ÖZET

G DÖRTLÜ YAPILARININ İKİNCİL YAPILARININ HESAPLAMALI TANIMLANMASI

Direk, Tugay

Bilgisayar Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Doç. Dr. Osman Doluca

Temmuz, 2022

G-dörtlü yapılarının ikincil yapı sınıflandırması için bir standart eksikliği uzun bir süredir bulunmaktadır. Çıkıntı yapan G-bölgelerinin ve uyumsuz G-dörtlülerinin keşfiyle durum, hayal edilenden daha da karmaşık bir hal almıştır. Bu nedenle, günümüze kadar G4 ikincil yapılarını tanımlamak için bir standart getirmeyi amaçlayan sınırlı sayıda çalışma yapılmıştır. Bu çalışmada, üç boyutlu yapısal verileri kullanarak G-dörtlülerinin ikincil yapılarının tanımlanması için yeni bir yöntem öneriyoruz. Kısaca, tetradları ve döngüleri tanımlamak için guaninlerin koordinatları işlenir. Daha sonra, her bir tetrada katılan guaninleri ve döngüleri gösteren ikincil yapıyı bir figür şeklinde sunuyoruz. Ayrıca, dörtlü ve dörtlü yapıların topoloji tabanlı sınıflandırmasına dayalı ONZ sınıflandırması uygulanmış ve çalışmamızın sonuçları bununla karşılaştırılmıştır.

Anahtar Kelimeler: G-dörtlüleri, İkincil Yapılar, Üç Boyutlu Yapılar, Loop, Tetrad, G-dörtlülerin Sınıflandırılması

Dedicated to *all seekers*…

ACKNOWLEDGEMENTS

While I’m completing the Master of Science program, I would like to thank my advisor Assoc. Prof. Dr. Osman DOLUCA and second advisor Asst. Prof. Dr. Kaya OĞUZ now that they gave me the chance for the related project which I evaluate as a big opportunity. Additionally, I’m considering myself lucky for the relationship beyond the lecturer-student's and his high patience to me. I appreciate such facts that could belong to only a few students.

Besides, I would like to thank my other teachers who are also found in my jury members; Assoc. Prof. Dr. Gökhan KARAKÜLAH, Asst. Prof. Dr. Kutluhan EROL for their huge contributions to my academic career. In addition, Asst. Prof. Dr. Cihangir YANDIM and Asst. Prof. Dr. Oktay KAPLAN shares the same contributions. In addition, I would like to thank my valuable friends, especially Hüseyin Saygın Portakal, who helped me during the creation of the thesis process.

Besides those valuable people, I would like to thank my parents Fevzi DİREK, Habibe DİREK and my sister Başak DİREK for their great support, beliefs, and being a driving force for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1. Definition and Roles of DNA and RNA

DNA and RNA are two main components of living creatures. DNA is responsible to encode the genetic information and RNA has a role to convert this genetic info into a form that can be used later to make proteins.

DNA structure was proposed by Watson and Crick in 1953. The structure was stated as two helical chains that form a helix.(Watson and Crick, 1953) These chains contain 4 nucleobases which are adenine, thymine, guanine and cytosine. Adenine and guanine bases are named as purines, two carbon nitrogen ring bases, and thymine and cytosine are named as pyrimidine, one carbon nitrogen ring bases, as can be seen in Figure 1. DNA is made by DeoxyriboNucleic acid which contains sugar deoxyribose.

Similarly, RNA is a short term for RiboNucleic acid which contains sugar ribose. Difference between deoxyribose and ribose is that ribose has one less number of -OH groups than DNA. Uracil takes the place of thymine in RNAs. (Alberts et al., 2002)

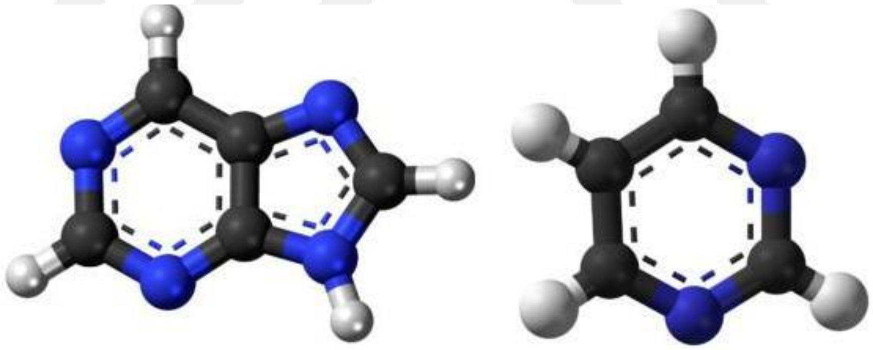


Figure 1. Structures of purine (L) and pyrimidine(R) bases. (Source: Diffen, n.d.)

The nucleobases can form a base pair between adenine-thymine and guanine-cytosine which is also called Watson-Crick pairs in DNA, as can be seen in Figure 2. Hydrogen bonds hold two bases together. These are weak bonds that can be separated easily to permit the separation of strands to copy the genetic materials as the function of DNA. The atoms of DNA are carbon, hydrogen, oxygen, nitrogen and phosphorus. When these atoms are combined, sugar-phosphate structure of DNA is formed. DNA has two main sub-structures which are Deoxyribose sugar and phosphate group. Deoxyribose sugar contains five carbons atoms which are ordered in the ring. The beginning and ending of these five carbons are named as 1’(1 prime) and 5’. The C atom at 1’ location corresponds to the right of the O atom and the C atom at the left of the O atom to 5’ location. The other sub-structure phosphate group is linked to sugar by its -OH group.

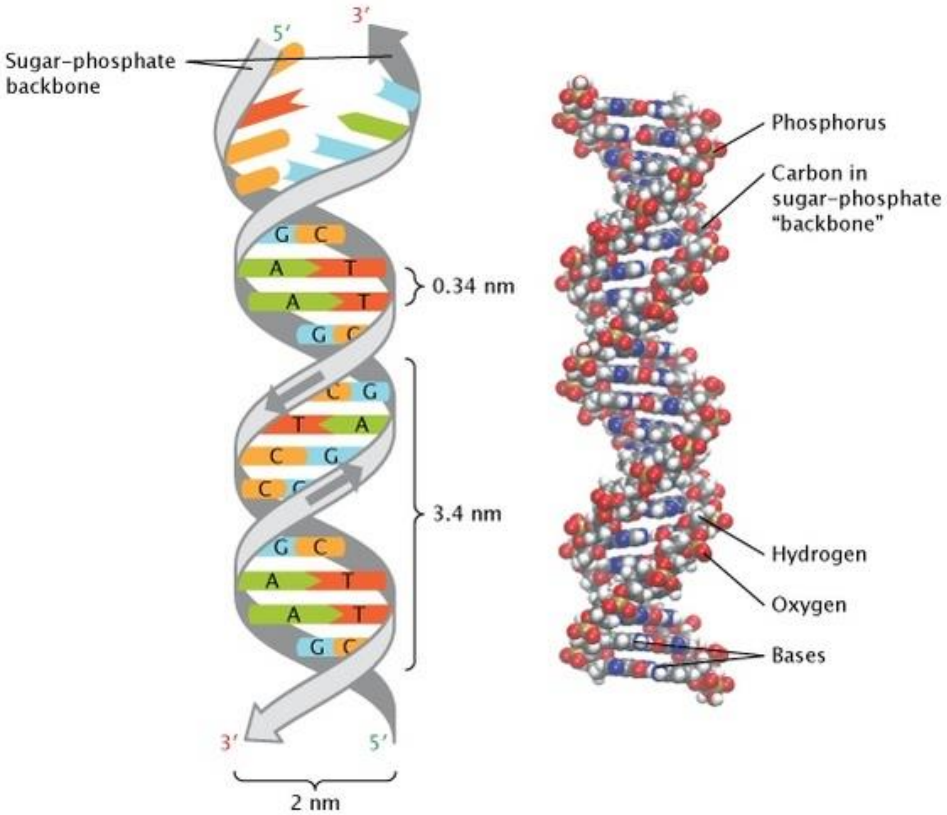


Figure 2. Structure of DNA (Source: Nature, n.d.)

Lastly, the sub-structure which makes the difference between different nucleotides are bases which are attached to the -OH group of sugar at the 1’ carbon. (adenine, guanine, thymine and cytosine) (Guerra et al., 2000) As mentioned earlier, the first discovered structure of DNA was double stranded helix. However, later it was found that there are many more possible formations. Two of these are triplex and quadruplex DNAs.

1.2. G-Quadruplex Structure

G-quadruplex (G4) is a special type of helical quadruplex which was identified in the 1960s. Researches on G-quadruplex structure is especially one of the hot topics since DNA sequences which repeats at the end of the chromosomes, called telomeres region, form G-quadruplexes (Rhodes and Lipps, 2015) and it has been discussed that they play a significant role in cancer therapy. Because of its stable structure, they regulate cellular functions. Some analyses also showed that there is a correlation between the G4 structures and intratumor heterogeneity. Different ligands, which is a compound that forms a complex by a G4, have been developed to target the G-quadruplex structure for therapies. (Kosiol et al., 2021) G-quadruplexes contain four guanines in a planar surface called G-quartets and one of these G-quartets is named as tetrad. Each guanine in the tetrad makes bonds between the two neighbouring guanines. This bond between two guanine nucleobases is called Hoogsteen hydrogen-bond. Guanines have 5 carbon, 5 nitrogen and 1 oxygen atoms. For example, nitrogen (N) atoms can be labelled as N1, N2, N3, N7, N9 consecutively according to their location in the structure.as can be seen in Figure 3. Hoogstene bond is established between N2-N7 and N1-O6 atoms of two neighbouring guanines as seen in Figure 4.

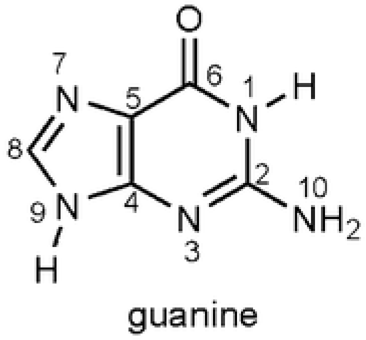


Figure 3. The structure of guanine base (Source: Madzharova et al., 2016)

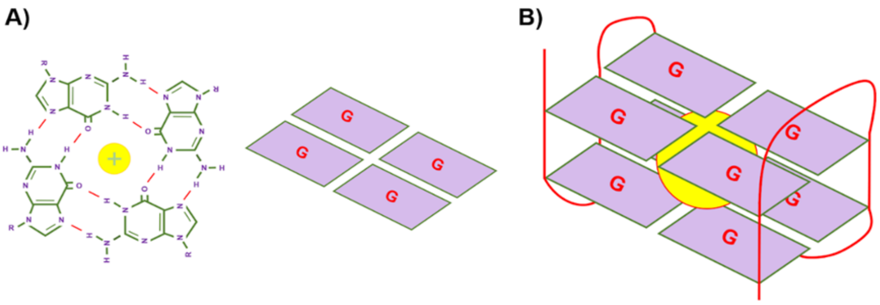


Figure 4. G-quadruplex structure A) G-tetrad formed by four guanines B) G-quadruplex generated by two stacking tetrads. (Source: Portakal et al., n.d.)

The molecular structure of G-quadruplexes can be classified based on the number of individual strands that take part in it. These are unimolecular (intramolecular), which consist of one continuous strand, and intermolecular, which consist of more than one individual strand, G4s, as can be seen in Figure 5. In unimolecular G4s, one strand has to fold on itself in order to bring the guanines to correct positions so that tetrads would be formed. The rest of the strand that does not contribute to the formation of the tetrad form loops. How the tetrads are connected to each other through the loops define the type of the loop. There can be three types of loops according to the configuration of G4s. Loops can connect guanines from the same tetrad and which are neighbours, from the same tetrad but not neighbours or from the different tetrads.

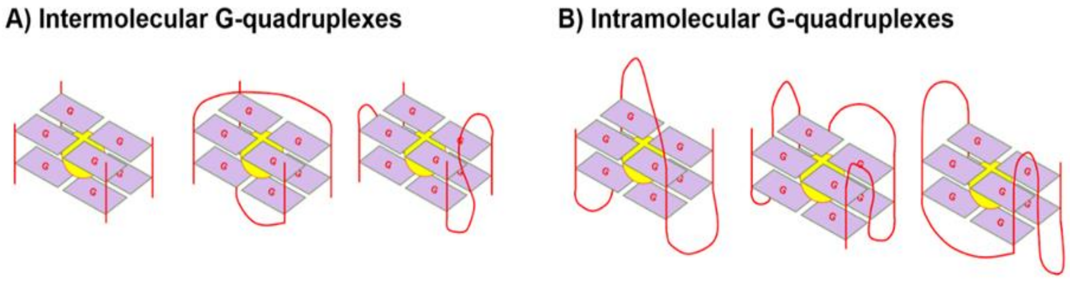

A) Intermolecular G-quadruplexes
B) Intramolecular G-quadruplexes

Figure 5. Types of G-quadruplexes A) intermolecular G-quadruplexes, B) Intramolecular G-quadruplexes. (Source: Portakal et al., n.d.)

1.3. The Aim of Study

DNA and RNAs have regulatory roles in the cell and are used as targets in therapies for different diseases. Knowing the general structure of these biomolecules can help us to find out the resultant factors of these important biomolecules. They can have a variety of shapes affected by different configurations of biological components. Therefore, extensive research has been done in this field and some specific tools have been developed. Our focus on this study is on a special structure called G-Quadruplex which contains 4 guanines in a planar surface. It is an important structure since it has been discovered that they exist in the genes of diverse species including humans and it has significant roles. Day by day, new biological roles of these structures have been discovered. Because of its unique structure and regulatory role in the cell, they are a target for cancer and neurodegenerative disorders.(Popenda et al., 2019) DNAs and RNAs have three types of structures which are primary, secondary and tertiary (3D) structures, as can be seen in Figure 6. Primary structure is the one dimensional sequence of nucleic acids. Tertiary structure is defined as the three dimensional structure which was formed after the folding of biomolecules. Secondary structure is a 2D representation to annotate the three dimensional topology of the structure. While these definitions were initially established for protein biomolecules, the nucleic acids do not have a clear definition when it comes to secondary structure, especially for non-canonical nucleic acids.

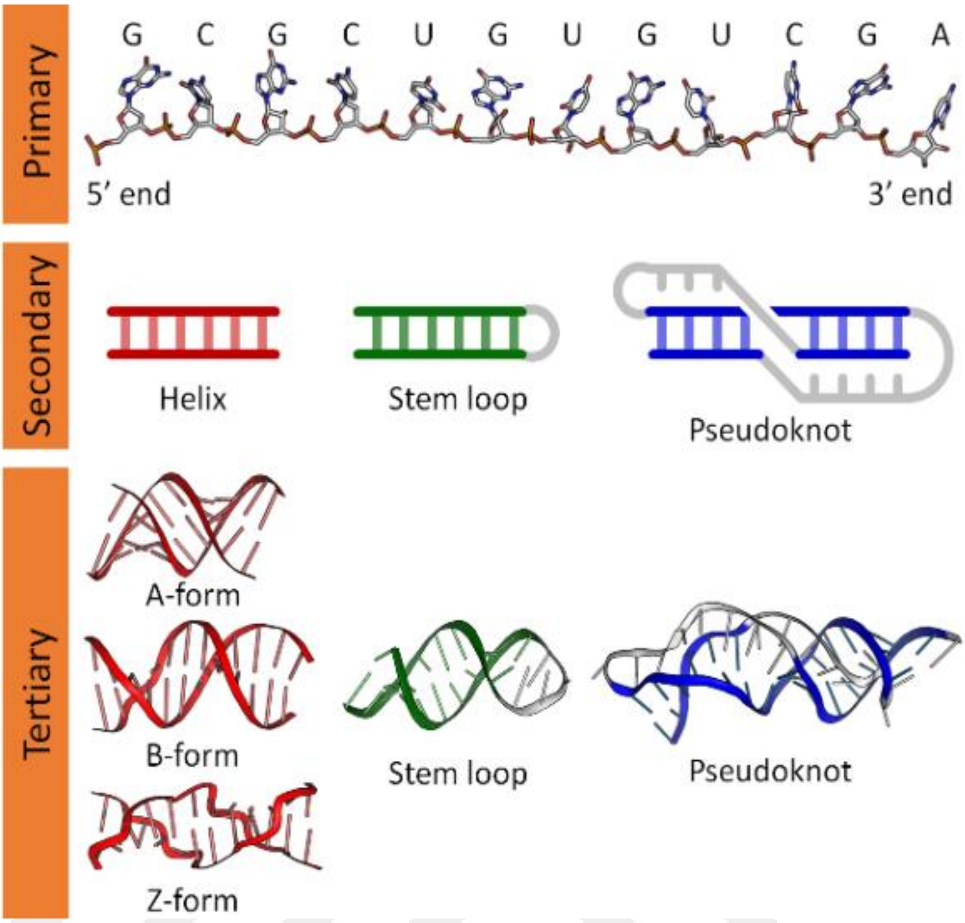


Figure 6. Different structures of DNA/RNA (Source: Nucleic Acid Structure, 2010)

Classifying the G-quadruplexes has not been standardised up to now since there are many parameters in the classification stage like bulged G-tracts, mismatched G-tetrads and different types of loops. Our focus in this study is to classify G-quadruplexes according to their secondary structures, which we define through tetrad composition, loop position and loop types. In the literature, the most common classification is based on the directionality of the strands that form the tetrads. Accordingly, a G-quadruplex has been identified as parallel, antiparallel and hybrid, as can be seen in Figure 7. However, only the three classification is not enough to represent the features of wastly variable G-quadruplex structure. In this study, with a novel approach, a G-quadruplex classification will not be restricted to any pre-decided classes. Instead it will be categorised and visualised in a way that expresses the topology clearly without omitting important details.

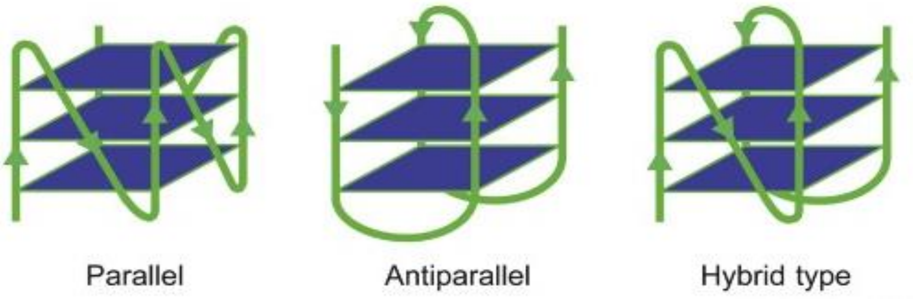


Figure 7. Different topologies based on the loop types (Source: Carvalho et al., 2020)

CHAPTER 2: LITERATURE REVIEW

2.1. DSSR (Dissecting the Spatial Structure of RNA)

The first study to be mentioned in this field is DSSR (Dissecting the Spatial Structure of RNA). DSSR emerged for the need of RNA annotation from three dimensional structure. It is a part of the 3DNA suite in which any nucleic acid structure such as DNA or RNA can be processed for annotation. RNA is a much more sophisticated structure when compared to DNA. Even RNA which contains a small number of nucleic acids can form complex structures. Therefore, DSSR was developed and added to the 3DNA suite especially for RNAs. Day by day, new RNA structures and functions of these RNAs are discovered, and the need to search and characterise these new structures should be handled. DSSR can recognize nucleotides, detect hydrogen-bonded base pairs, and high-order coplanar base associations, identify and recognize different structures including G-tetrads. DSSR uses Protein Data Bank (PDB) database to download the related data and weekly update the data sources.

To identify nucleotides, DSSR processes three dimensional data which contains x, y and z coordinates of atoms and standard names of base-ring atoms. While identifying nucleotides, they use a reference purine with nine ring atoms. The atoms in a residue are matched to the reference purine using least-square fitting, as can be seen in Figure 8. If at least three atoms are included in a residue and root mean square deviation of the fit gives the expected result, the corresponding component is identified. Again, a reference frame is used to identify the orientation and position of each base in the structure in three dimensional space. A least-square fit is used to fit the atomic coordinates for all nucleotides which are processed against a reference frame. After these procedures, three dimensional data which are positions and orientations of nucleotides in a structure are obtained. Following step is finding the hydrogen bonds. These bonds are detected using geometric calculations by taking into account nitrogen (N) and oxygen (O) atoms. All nitrogen and oxygen pairs which fall below 4 Armstrong cutoff value are found and other factors that give knowledge about bonding like donor/acceptor properties, angles with neighbouring atoms are taken into account in this process. Base pairs are found by geometric approach using base reference frames.

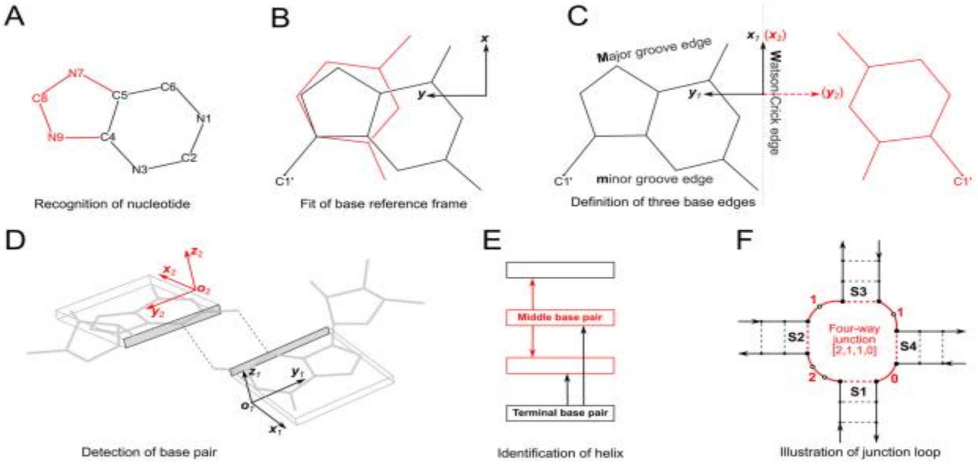


Figure 8. Base frames and some of the steps applied to identify nucleic acid structural components (Source: Lu et al., 2015)

Five different criteria are used while detecting base pairs. These are: the length between the two origins, the vertical separation between the base planes, the angle between the base normal vectors, the absence of stacking between the two bases, and the presence of at least one hydrogen bond involving a base atom, whether a base atom is involved at least in a one hydrogen bond or not. If all these criterias are satisfied when compared with the default values, these bases are labelled as pairs. DSSR also detects multiplets which contain three or more linked bases in the same planer surface geometry by hydrogen bonds. These structures are defined as interconnected base pairs and also include G-tetrads.

Lastly, DSSR output can be used to see the three dimensional structure by giving it as an input to different visualisation interfaces like VARNA. (Caruso et al., n.d.; Colasanti et al., 2013; Lu and Olson, 2003, 2008; Miskiewicz et al., 2021; Zheng et al., 2009)

2.2. 3D-NuS(A Web Server for Automated Modelling and Visualisation of Non-Canonical 3 Dimensional Nucleic Acid Structures)

One other paper which focuses on this topic is 3D-NuS(A Web Server for Automated Modelling and Visualisation of Non-Canonical 3 Dimensional Nucleic Acid Structures). The purpose of this work is to model the three dimensional structure of duplexes, triplexes, and quadruplexes. According to the paper, knowing the 3D structure of DNA and RNA is essential to predict their functions. For example, quadruplexes play a role during the replication, recombination, transcription, and chromosome stability. They also exist at the telomeres of the DNA which can be targets during drug discoveries. However, obtaining 3D structure by experimental ways has some difficulties like triplexes cannot be defined well by X-ray crystallography or NMR techniques. Therefore, 3D-NuS can be an alternative tool to view the 3D structure of different types of DNA and RNA.

3D-NuS has a web server where an input can be fed and the result can be shown visually to the user. They store atomic coordinates of all different types of structures like base-pair geometries, DNA and RNA duplexes, base triplets and G-quartets interacting by hydrogen bonds at the web server to identify the 3D structure of DNA and RNA easier. This information is from experimental results or modelled manually. To generate polymers, they place the centre of the base pair/triplet/quadrate at the centre of the Cartesian coordinate system and use them as a template. After this process, they transform the three dimensional coordinates (x,y,z) of repeating unit atoms to a cylindrical coordinate system and use helical twist and rise values to predict the polymer.

To model a duplex RNA or DNA, the user must enter a sequence for both strands. The sequence length should be the same and orientation of the first and second strands is accepted 5’->3’ and 3’->5’ sequentially. The “sequence-specific model” models the duplexes using the sequence-specific base pair step and base pair parameters. They generated 10 distinctive local base pair steps and base pair parameters by using X-ray and NMR results from the PDB database. At the beginning, they used 678 DNA and 616 RNA structures to generate the related parameters. After some filtering mechanisms like deleting extreme bending structures to obtain better results, 222 DNA and 166 RNA motifs were taken into account for parameter generation. Following these steps, mentioned 3DNA software suite were used to derive local base pairs and base pair parameters and results were kept at 3D-NuS server.

Triplex structures are modelled by using two classes and eight subclasses. These classes were identified by the orientation of the third strand at the structure as parallel and anti-parallel. Different combinations of DNA or/and RNA can be also considered to model by 3D-NuS. To get a result for triplex, right classes and subclasses must be chosen by the user and three sequences must be entered for each of them. The orientation of the first and second strand of the triplex is like in duplex, the third one is determined by the class selection either parallel or antiparallel.

Inter- and intra- G-quadruplexes can be also identified by 3D-NuS. They classified the G-quadruplex structure into 19 classes. Monomeric structures are from Q1 to Q5 and from Q14 to Q18. Dimeric structures are from Q6 to Q9, and Q19. Tetrameric structures are from Q10 to Q13, as can be seen in Figure 9. For monomers, there are two subclasses as DNA(D) or RNA(R). For intermolecular G4s, the combinations of RNA and DNA can also be part of classification, as such there are four subclasses for dimers which are DD, DR, RR, RD. To get a result from 3D-NuS for G-Quadruplex structure, class of the structure and G-quadruplex length must be given as input. To generate a monomeric q-quadruplex structure, sequences of each 3 loops must be also given as input.

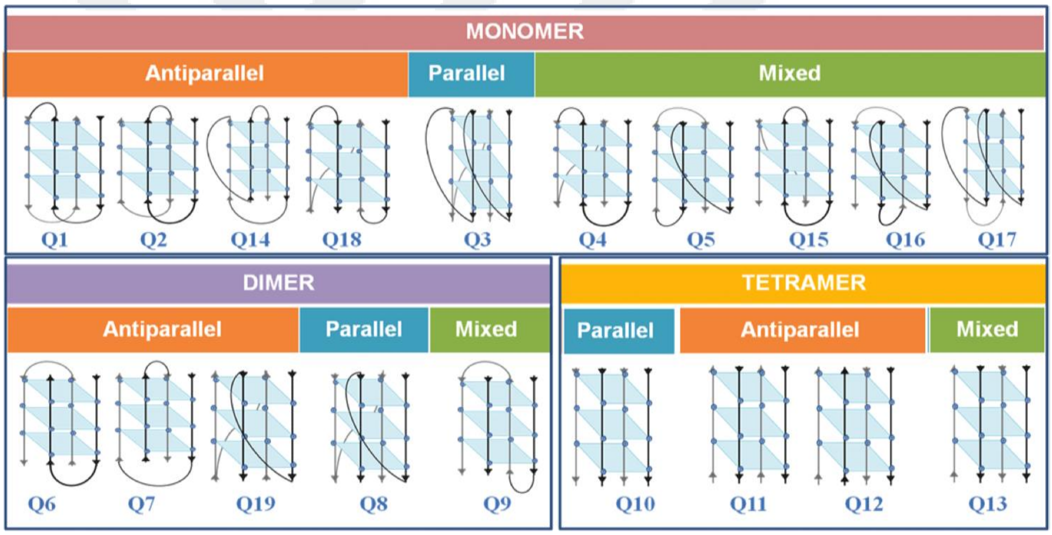


Figure 9. G-quadruplex structures with different strand directions (Source: Purnima et.al, n.d.)

After these structures are generated, the models are also optimised by Xplor-NIH. During this process, energy minimised versions of all possible structures are considered and those with minimised energy are accepted as resultant structures since the less energy the molecule has, the more stabilised it is. Since G-quadruplexes are complex structures, they constrain some of the strands sequentially, and try to minimise the energy of other strands. After this stage, a unique molecule ID is generated. And the cartesian coordinates of the energy minimised structure can be downloaded by this ID for the upcoming seven days. It can also be visualised by JSmol molecule viewer. (Patro et al. 2017)

2.3. RNApdbee

Another paper that annotates the G-quadruplex structure is RNApdbee. The main focus of RNApdbee is identifying the knotted structure in RNAs. For this reason, they firstly developed the RNApdbee web server. In this server, there are three steps which are base pairs identification, secondary structure encoding, and visualisation to reach the results. Base pair identification is found by other tools like 3DNA/DSSR. To encode the secondary structure, they developed 5 different algorithms which are Hybrid Algorithm (HYB), Dynamic Programming (DP), Elimination Min-Gain (EG), Elimination Max-Conflicts (EC), First-Come-First-Served (FCFS). For example, Dynamic Programming includes four operations iteratively. The steps are as follows. Sets of the nested base pairs are found in the input set, found base pairs are associated with the current order (initially set to 0), increasing the order by one, and removing the obtained base pair from the input set. When all of the base pairs in the input set are taken into account, the algorithm stops. After this step, visualised structures are shown. Visualisation is supported by different tools like PseudoViever and VARNA. The secondary structure can be generated by these tools which expresses the three dimensional view of the RNA with dots and brackets. By these dot-bracket notations, base pairs can be seen with different colours and topology of the structure can be examined easily. At the first version of RNApdbee server, there was no support for G-quadruplex identification. With the second version “RNApdbee 2.0: multifunctional tool for RNA structure annotation”, they also implemented this feature. As an example, the secondary structure of the G-quadruplex of human telomeric RNA is shown by the paper. And, all the tetrads can be viewed clearly as can be seen in the figure below. (Antczak et al., 2014, 2018; Zok et al., 2018)

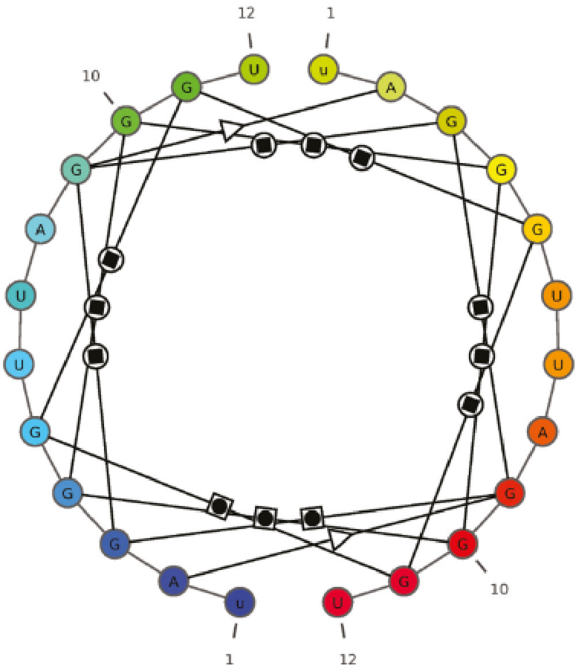


Figure 10. 2KBP RNA, secondary structure figure obtained by RNApdbee 2.0 (Source: Zok et al., 2018)

2.4. Topology-based classification of tetrads and quadruplex structures

The last work that will be mentioned in this literature is Topology-based classification of tetrads and quadruplex structures. They proposed a new method to classify the quadruplexes. This tool can classify any quadruplex which can also include different tetrads other than those containing only guanines. However, since the majority of tetrads are built by guanines, the main focus here is on the G-quadruplexes. At the first stage, 3DNA suite is used to identify the base pairs. Quadruplexes are considered as a graph data structure. Guaines correspond to vertices in the graph and edges correspond to bonds between guanines. According to this new classification method, they categorised G-quadruplexes into three main classes. These are O, N and Z classes. If we consider 1, 2, 3, 4 as the guanine numbers which are involved in a tetrad and if the bonds between these guanines are as follows (G1→G2, G2→G3,G3→4,G4→G1), this motif constructs a shape like a square. O phrase is coming from the circle that is covering this square. If the bonds between these guanins are as follows (G1→G2, G2→G4,G4→G3,G3→G1), since the shape is like a bow tie, it is called N class. Lastly, if the bonds between these guanins are as follows (G1→G3, G3→G2,G2→G4,G4→G1), since the shape looks like a glass hour, it is named as Z class, as can be seen in Figure 11.

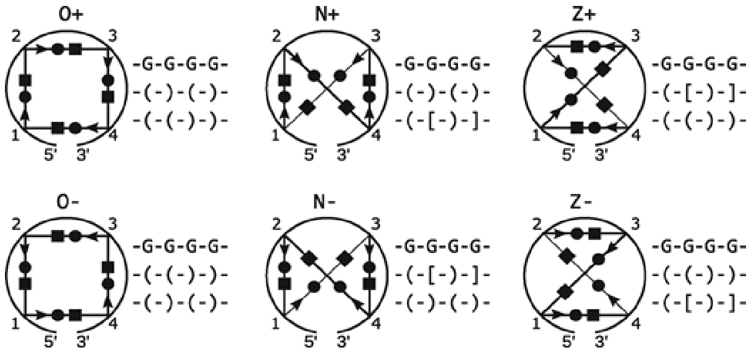


Figure 11. The order of bases in a tetrad according to the O, N and Z classes and their dot-bracket notations (Source: Popenda et al., 2019)

These are the 3 main classes in this taxonomy. According to the direction of edges in quadruplexes, these can also be categorised as + and -. Therefore, we now have 6 different subclasses also. One other subclass is also identified by the order of nucleotides which are named as parallel (p), antiparallel (a), and hybrid (h). We can also have more subclasses by this method which are Op, Oa, and Oh for the O class, and 6 more subclasses for N and Z classes. If all of the tetrads in a quadruplex belong to the same type, these are called regular quadruplexes If we have hybrid types of tetrad in a quadruplex, these are called irregular and given a class name as M. There is also one last additional class which corresponds to -bi and -tetra molecular quadruplexes. These ones are named as R class which means remaining structures. (Popenda et al., 2019)

CHAPTER 3: METHODOLOGY

3.1. Data

The data which contain the G-quadruplexes were identified by http://g4.x3dna.org/ website. All of the PDB (protein data bank) ids of DNA and RNA structures are extracted from the aforementioned website and the pdb files which contain the three dimensional data of the corresponding structures were downloaded from https://www.rcsb.org/.

After downloading pdb files, they were filtered before feeding into the algorithm. Firstly, we excluded the G-quadruplex structures which are not unimolecular since unimolecular G4s are the most abundant G-quadruplex structures in living organisms and more biologically relevant. For any PDB containing more than one model, only the first model was used. Total number of PDB structures containing ligands were counted using the “Hetatm” label and any compound with more than 5 atoms was considered a ligand. Ligands containing structures were not treated separately.

The PDB structures with modified guanine bases were also used for the study. The algorithm was designed to regard these as any guanine base during Hoogsteen base pair discovery. These modified bases are GFL, 8OG, 0G, GF2, LCG, GF0 and BGM.

3.2. Algorithm

3.2.1.Detecting H-bonds and Hoogsteen base pairs

For each guanine, we consider H-bond donor and acceptor atoms to check for potential H-bond formation. The algorithm searches for all H-bond acceptor and donor pairs (N2 for N7, N7 for N2, N1 for O6 and O6 for N1) under a distance of 3.5 Å. After detecting all putative pairs, they were considered for the correct bond angle. For each guanine, a normal vector is calculated for a plane passing through the N1, N2 and N7 atom centres. Then the vector from the H-bond-forming atom of the considered guanine to the corresponding H-bond-forming atom of the potentially paired guanine is calculated as the H-bond vector. If the angle between the H-bond vector and the normal of the considered associated guanine is in the range of 90±35 degrees, this H-bond is considered as a potential H-bond. For any two guanines, if the algorithm discovers the corresponding potential H-bonds between the same atoms in both directions (N2 to N7 and N7 to N2 or N1 to O6 and O6 to N1), then this H-bond is considered a true H-bond. If true H bond forms are detected from N2 to N7 and N1 to O6 between any guanine pair then these guanines are considered a Hoogsteen base pair.

3.2.2. Detecting tetrads

For a G-quadruplex structure, the network of Hoogsteen base pairs may be regarded as a graph where the guanines of any tetrad should form a clique of four nodes. To identify tetrads, all potential cliques are discovered.

If every node in a clique makes a Hoogsteen base pair with two others in the same clique, this indicates that the base pairs form a cycle and can be considered as a G-tetrad. However, any number of guanines may be connected consecutively and yet may not necessarily form a cycle through Hoogsteen base pairing, according to the definition above. When the algorithm finds consecutively connected guanines through Hoogsteen bonding, any two neighbouring guanines in a real tetrad may have not been identified as a Hoogsteen base pair by the algorithm. This does not necessarily indicate that the tetrad is absent, but the Hoogsteen bonding may be missed due to low stability and increased flexibility in the tetrad. To include these instances, the definition of Hoogsteen base pairing rule is relaxed as 90±70 degrees angle range to facilitate the discovery of these tetrad cycles. An example of this implementation for the 148D structure is mentioned in the results section.

Then the tetrads are sorted according to their respective positions. The distances between the centroids of the tetrads are compared to find the most distant pair of tetrads. These represent the topmost and bottommost tetrads in the structure. Starting from one of these two, chosen randomly, added to a sorted list of tetrads. Then the closest tetrad is added to the sorted list. This is repeated until the furthest tetrad is found. The sorting method is indifferent towards the orientation of the PDB structure.

3.2.3. Detecting loops

If the G-tetrad-participating guanines of the structure are visited according to the sequence order, it would be jumping between the tetrads. The direction of the jump is then used to identify the loops. A loop is registered when the direction of the jump changes from upwards to downwards along the stack or vice versa.

To determine the loops, each G-tetrad-participating guanine is visited one by one in 5′-3′ direction. During this procedure, guanines are added to a list. When a guanine is picked, the following conditions are checked to determine whether or not a loop exists. Firstly, if since the last loop or the beginning, there has been only two guanines and are participating in the same tetrad, there must be a loop between the last and current guanine. In case, the number of guanines since the last loop or the beginning is equal to the number of tetrads, this indicates that all tetrads must have been visited and there is no option but a loop to follow. In case that the aforementioned conditions were not met, and there has been more than two guanines since the last loop, the algorithm checks for the stacking features of the guanines.

Stacking occurs when two guanines are vertically aligned and their larger surfaces look towards each other. If the current and the last guanines are on the same tetrad, then these guanines are on the same planar surface and naturally can not stack. In which case, a loop is placed in between. On the other hand, if the current guanine is between the tetrads of the last and second to last guanines, then the distances between the guanines are relied on to make a decision regarding the placement of the loop. In that case, if the distance between the last and second to last guanine is less than the distance between the last and current guanine, the loop is placed between the last and last to second guanines. Otherwise it is placed between last and current guanines. An example of this procedure is shown in the discussion section.

3.2.4. Determining loop types

The loops are annotated as lateral, diagonal, and reversal as they are discovered. If two consecutive guanines are separated by a loop, the type of the loop is determined by the tetrads they took part in and existence of a Hoogsteen bond in between. If both guanines participate in the same tetrad and are also a Hoogsteen base pair, the loop is considered lateral. If they don’t share a Hoogsteen base pair, the loop is considered diagonal. Lastly, reversal loop occurs if these guanines took part in different tetrads.

3.2.5. ONZ detection

We have implemented an ONZ classification method which is Topology-based classification of tetrads and quadruplex structures (Popenda et al., 2019). To distinguish between the different types of tetrad classes, we have used the data of pairing guanines which were generated by our algorithm.

When guanines of a tetrad are sorted by their sequence order, the placement of the guanines determine the ONZ class of the tetrad. The guanines are labelled as Ga, Gb, Gc, Gd, respectively. Simply, if the bonds of the guanines in the tetrad are between Ga-Gb, Gb-Gc, Gc-Gd and Gd-Ga, it is labelled as O class. If it is Ga-Gb, Gb-Gd, Gd-Gc and Gc-Ga, then N class. If it is Ga-Gd, Gd-Gb, Gb-Gc and Gc-Ga,then Z class. Following figure shows the order of mentioned O, N and Z classes.

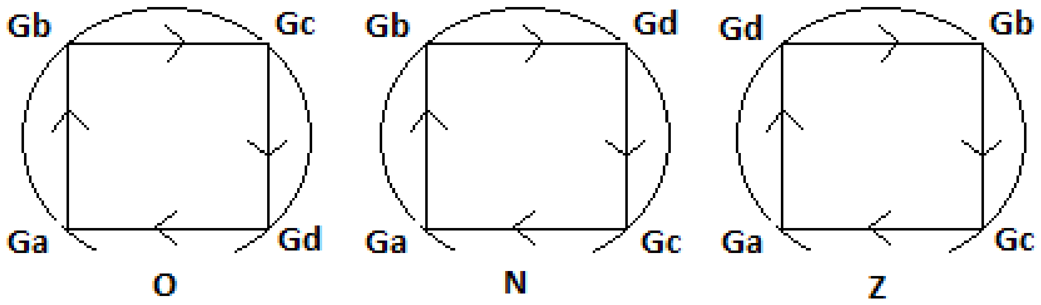


Figure 12. The arrangement of guanines according the different classes in ONZ classification

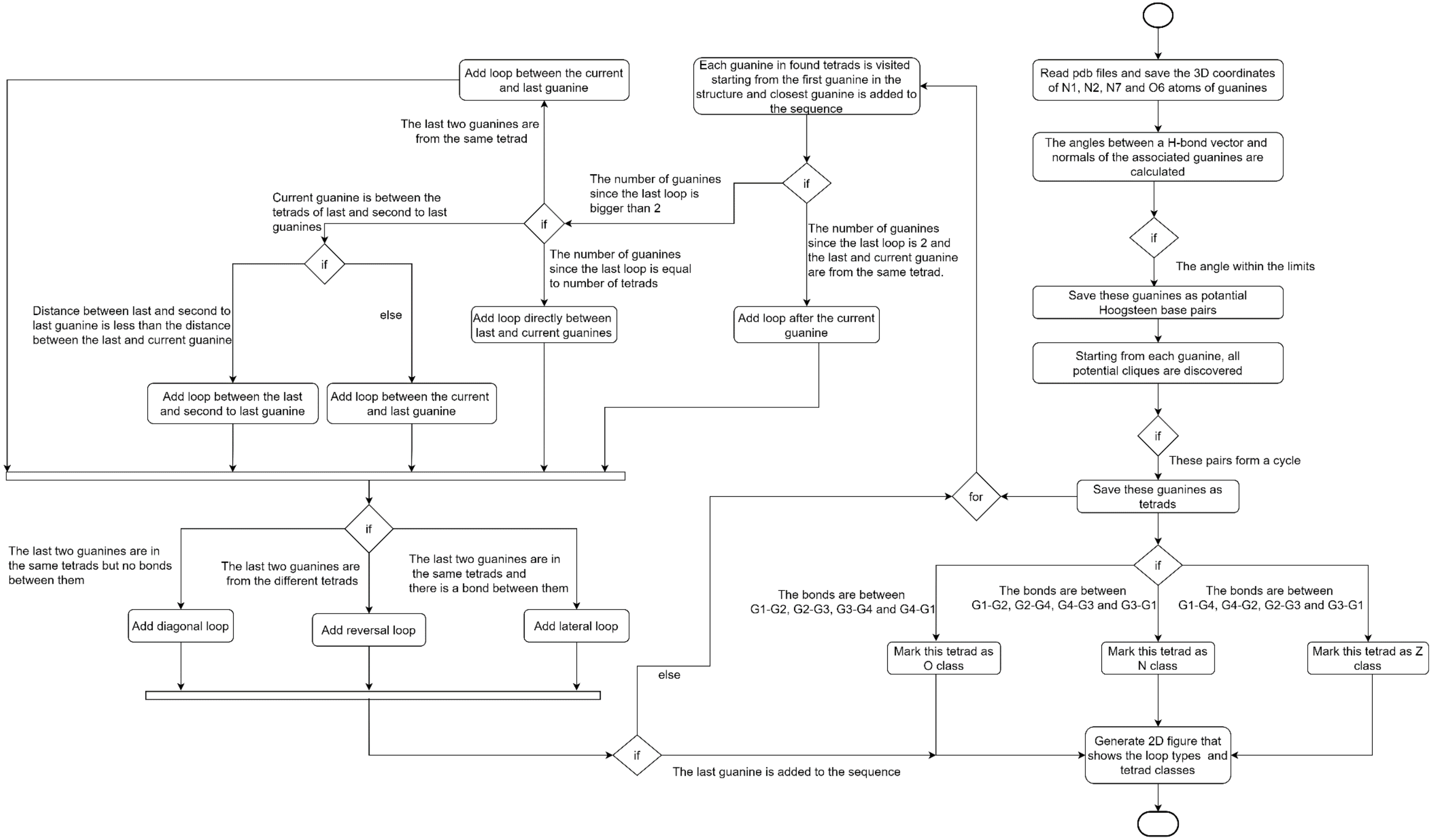

Read pdb files and save the 3D coordinates of N1, N2, N7 and O6 atoms of guanines
The angles between a H-bond vector and normals of the associated guanines are calculated
if
The angle within the limits
Save these guanines as potential Hoogsteen base pairs
Starting from each guanine, all potential cliques are discovered
if
These pairs form a cycle
Save these guanines as tetrads
for
if
The bonds are between G1-G2, G2-G3, G3-G4 and G4-G1
The bonds are between G1-G2, G2-G4, G4-G3 and G3-G1
The bonds are between G1-G4, G4-G2, G2-G3 and G3-G1
Mark this tetrad as O class
Mark this tetrad as N class
Mark this tetrad as Z class
Generate 2D figure that shows the loop types and tetrad classes
Each guanine in found tetrads is visited starting from the first guanine in the structure and closest guanine is added to the sequence
if
The number of guanines since the last loop is 2 and the last and current guanine are from the same tetrad.
Add loop after the current guanine
The number of guanines since the last loop is bigger than 2
if
The last two guanines are from the same tetrad
Add loop between the current and last guanine
The number of guanines since the last loop is equal to number of tetrads
Add loop directly between last and current guanines
Current guanine is between the tetrads of last and second to last guanines
if
Distance between last and second to last guanine is less than the distance between the last and current guanine
else
Add loop between the last and second to last guanine
Add loop between the current and last guanine
if
The last two guanines are in the same tetrads but no bonds between them
The last two guanines are from the different tetrads
The last two guanines are in the same tetrads and there is a bond between them
Add diagonal loop
Add reversal loop
Add lateral loop
if
else
The last guanine is added to the sequence

Figure 13: Flowchart of the algorithm

CHAPTER 4: Results and Discussion

The DNA and RNA structures listed in 3DNA website were identified and downloaded from Protein Data Bank. Only unimolecular structures were used for the rest of the study. Out of all 369 structures, 179 unimolecular structures remained, out of which, 162 DNA, 14 RNA and 3 of them are hybrid G-quadruplexes. Also 56 (42 DNA, 12 RNA, 2 hybrid) of all were containing non-protein ligands of at least 5 atoms, as can be seen in Figure 14. 12 PDB structures were found to include protein structures along with the nucleic acids. You can find the pdb files which were used in the data set in the supplementary section. All ligand or protein containing structures were also used for the analysis, as long as they contained a monomeric G-quadruplex.

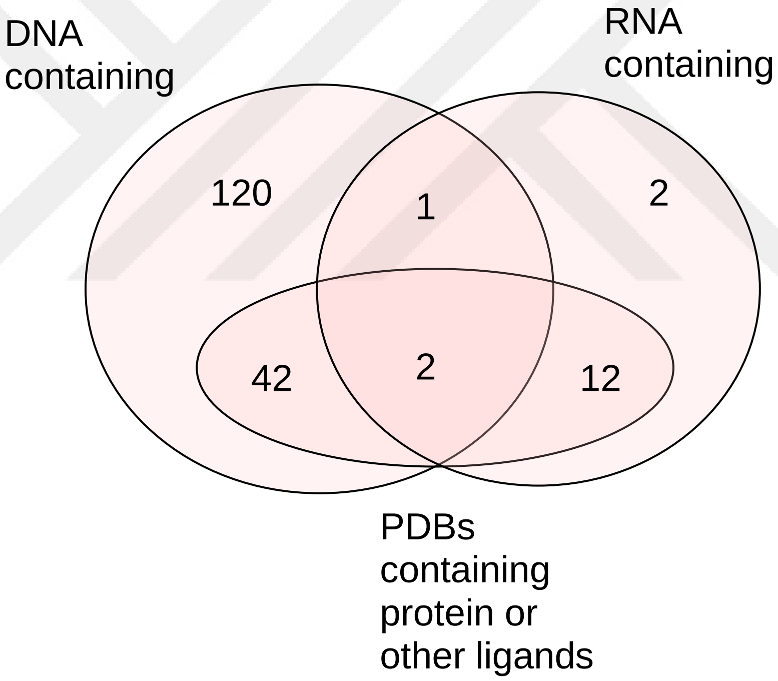


Figure 14. Number of PDB structures based on the backbone of the nucleic acid and presence of ligand.

The algorithm is run using each PDB file as the sole input, returning (1.) a collection of sets containing indexes of guanines for each tetrad, (2.) a one dimensional sequence of tetrad-forming guanines and loops, and (3.) a 2D representation of the G-quadruplex topology, as can be seen in below Figure 15.

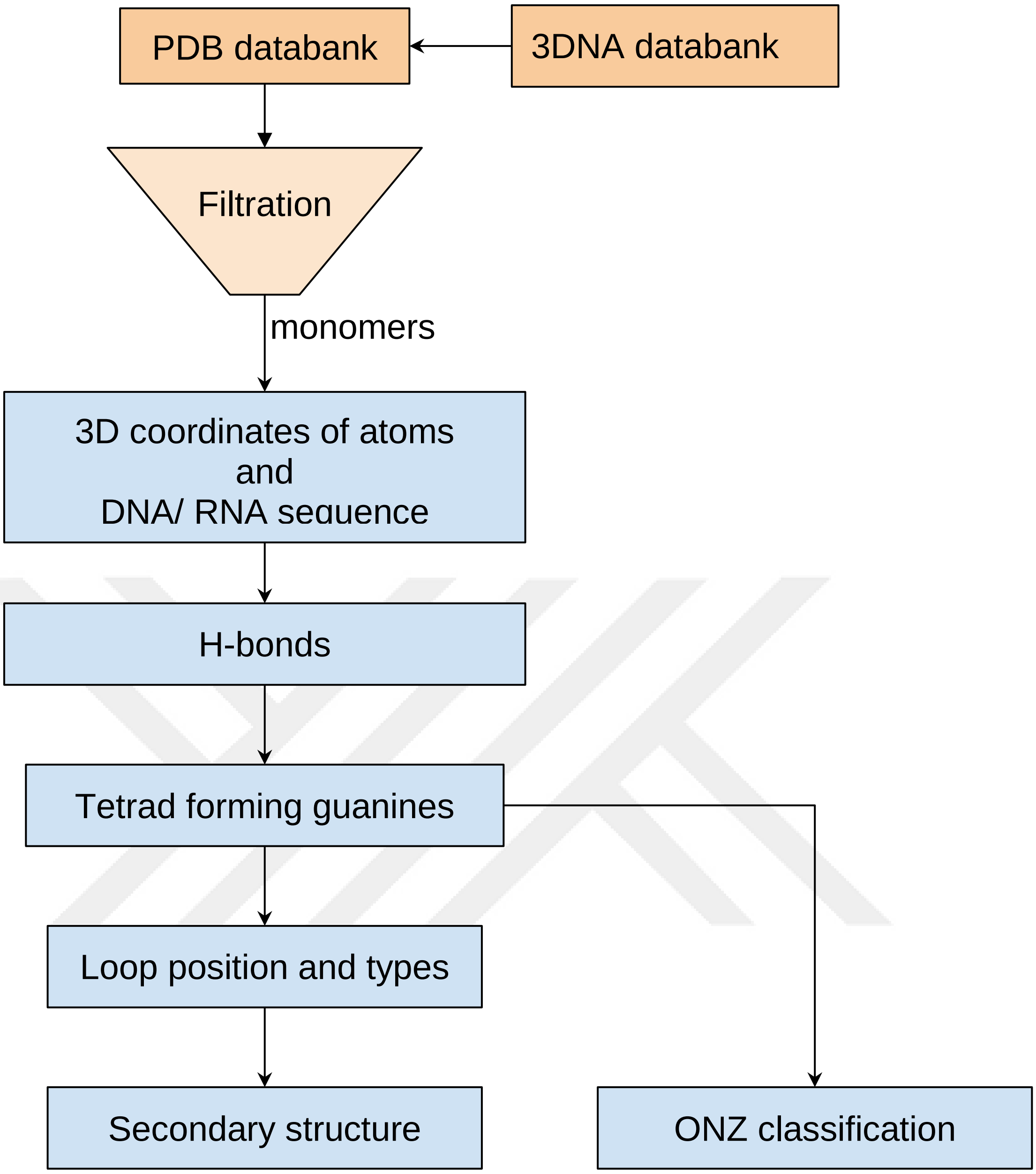


Figure 15. Brief flow chart of whole processes in the algorithm

After filtration, 3D coordinates of the structures were extracted from PDB. Should exist multiple models within a PDB file, only the first model is used. The sequence of the DNA monomer is identified using ATOM records.

All N2, N7 combinations and N1, O6 combinations are evaluated for distance to be identified as potential Hoogsteen H-bonds. Not all combinations can form H-bonds. The most important factor is the distance between the H-bond donor and acceptor. However, there is no consensus regarding the maximum distance of a H-bond. For that reason, we have decided to plot the distances for these combinations within 5 Å distance, as can be seen in Figure 16. The plot revealed a clear overlap of at least two distributions, where most of the distances are located around 2.8 Å. Such distance is expected for any H-bond. To isolate these, we have decided to use 3.5 Å as the cut-off.

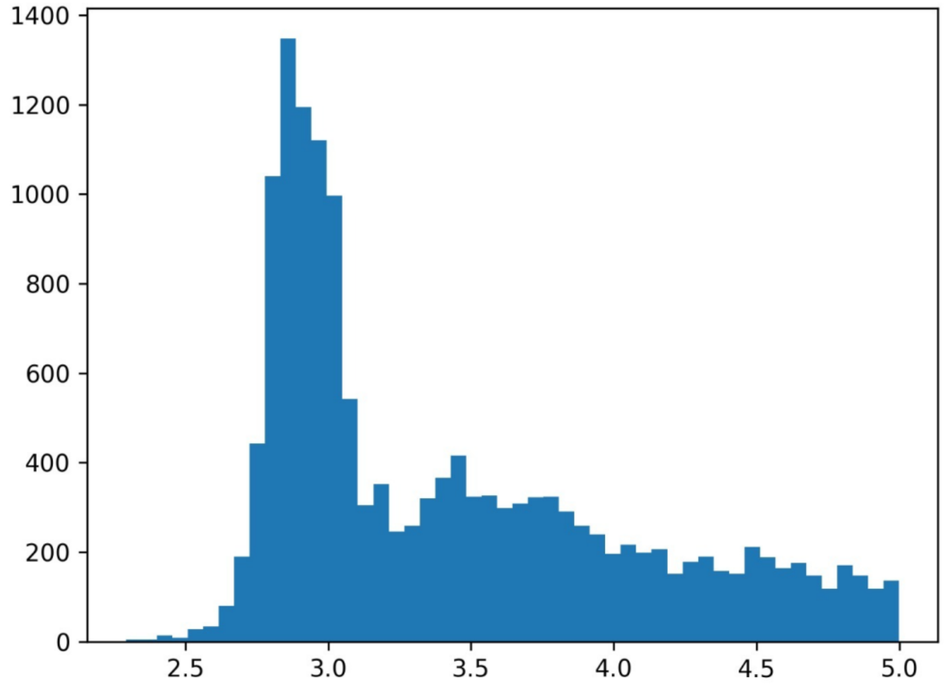


Figure 16. Histogram that shows the distribution for the distance between the H-bond donor and acceptor

The distance alone may not be enough to identify all H-bonds. Another aspect we have to consider is the respective positions of the H-bond donor and acceptor, such that these atoms need to be facing towards each other for strong attraction between the Hydrogen and electron pairs. This requires that the normals of the planes of these guanines are orthogonal to the corresponding H-bonds. However, due to molecular flexibility, a range of angles between the normals and the H-bonds are possible. To determine this range, we have plotted the distribution of the angles of possible H-bonds that can occur within 3.5 Å distance. (Figure 17, orange) It was apparent that most of these bonds had an angle above 55 degrees. Moreover, if we increase this distance limit to 5 Å, the number of putative H-bonds with sharper angles (<55 degree) dramatically increases, while the number of angles close to orthogonal (>55) shows marginal change. (Figure 17, blue) Infact, this separation is clearly observed in the plot. For that reason, we decided to employ 90±35 degrees as the threshold.

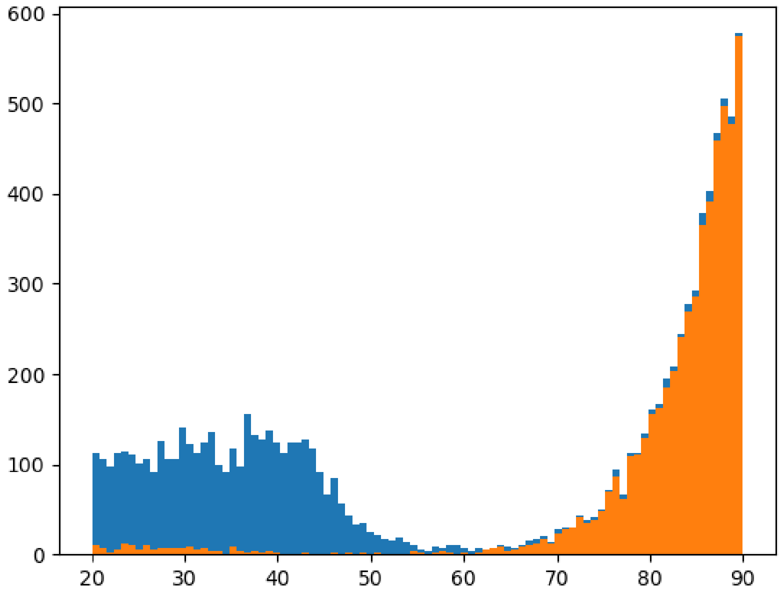


Figure 17. Histogram that shows the angles between H-bond vector of guanines and plane of corresponding guanine (x-axes) against the number of H-bonds (y-axes) for 3.5 Å threshold (orange) and 5 Å threshold (blue) in structures having 3 tetrads.

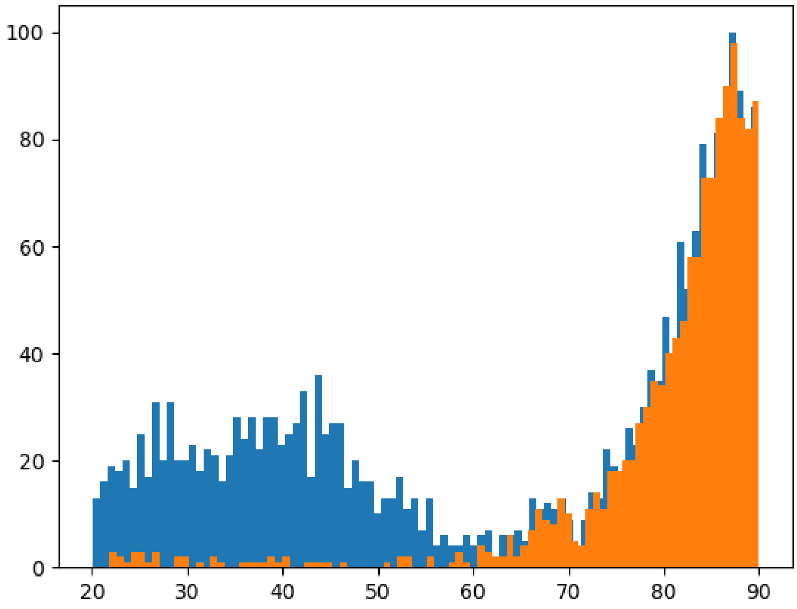


Figure 18. Distribution between the angles between H-bond and plane normals of corresponding guanines (x-axes) against the number of H-bonds (y-axes) for 3.5 Å threshold (orange) and 5 Å threshold (blue) in structures having 2 tetrads.

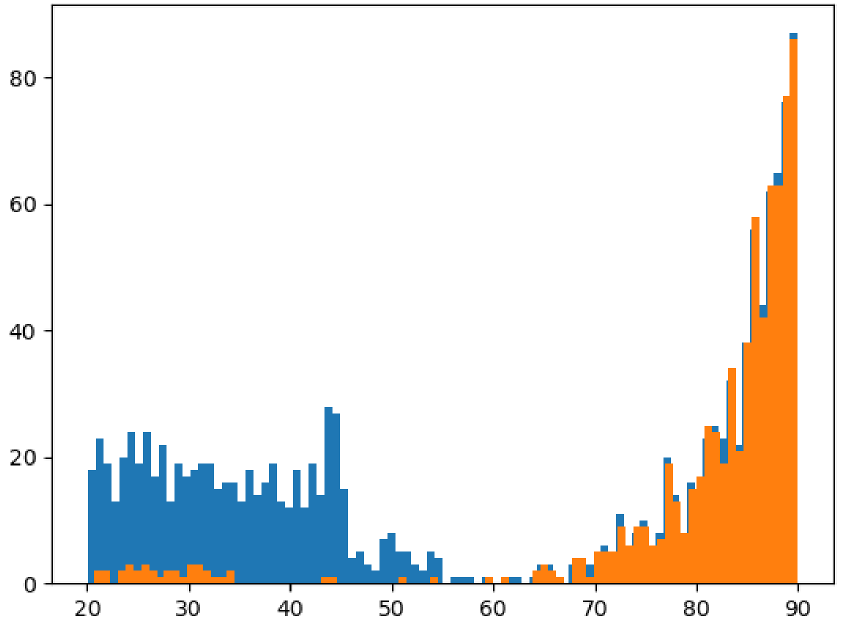


Figure 19. Histogram that shows the angles between H-bond vector of guanines and plane of corresponding guanine (x-axes) against the number of H-bonds (y-axes) for 3.5 Å threshold (orange) and 5 Å threshold (blue) in structures having 4 tetrads.

Whenever two H-bonds were identified between any two guanines, these are marked as Hoogsteen base pairs. In each structure, the hoogsteen pairs and the guanines that take part in them, forms a graph.

Any isolated community of guanines is evaluated for potential tetrads. After the algorithm detects consecutively connected four guanines within each community, the Hoogsteen bonding of the last guanine to the first guanine is necessary for the cycle to be complete. For only this final Hoogsteen bonding, the angle restriction is relaxed by expanding the 90±35 degrees range to 90±70 degrees.

This relaxation was necessary for 1OZ8 and 148D. In each of these structures, one of the tetrads do not form clear planes since some of the guanines in those tetrads have sharp slopes or are placed under or above the planar surface of corresponding tetrad. As a result, the angle range between 50 degree and 130 degree is exceeded between two particular Hoogstene base pairs and the cliques did not form a cycle. Below Figure 20 shows the angles in one of the tetrads of 148D structure. Between the 6th and 10th guanines, angle values are 147 and 154 degrees. By expanding the angle range, this pair can also be detected for cycle presence.

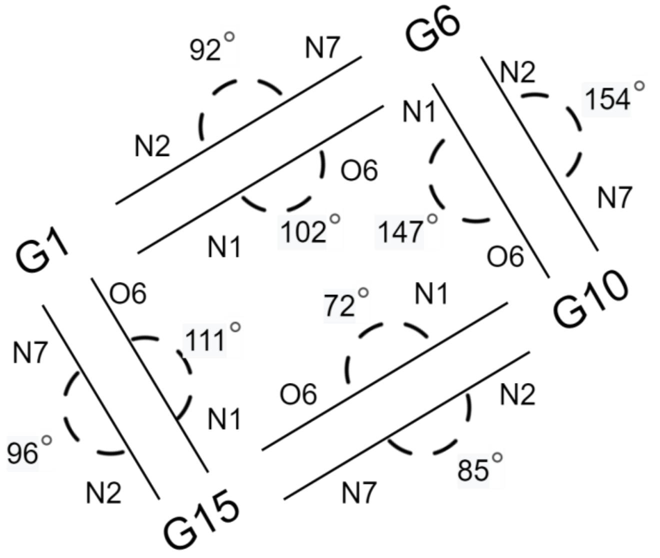


Figure 20. N2-N7 and N1-O6 bonds between Hoogstene base pairs and the angle between the H-bond vector and the normal of the considered associated guanine in one of the tetrad at 148D structure

The secondary structure figure representations are generated using tetrad and loop data and represent all conclusions drawn regarding a particular PDB structure, including sequence number of the guanines participating in each tetrad, the positions and types of the loops and ONZ classifications. (Figure 22-32) The line represents the backbone of the sequence while bases participating the tetrad are indicated with letters and aligned horizontally. The type of the loop is labelled along the line and each guanine that does not stack with already drawn guanines are added to the figure towards the right. The list of guanines participating in tetrad formation is written at the bottom and grouped by their tetrads. The column to the left of the figure indicates the ONZ classification of the tetrad.

For example 143D, is a canonical antiparallel G-quadruplex structure. This structure consists of three tetrads connected through two lateral and a diagonal loops. The algorithm has identified these three tetrads and grouped them as horizontally in the figure representation, as can be seen in Figure 21.

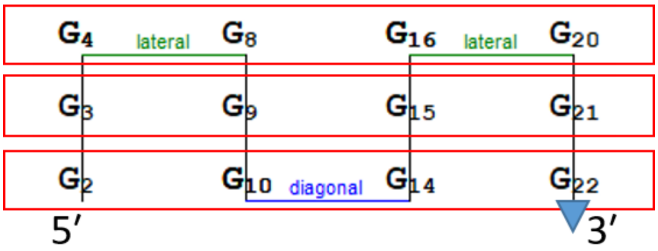


Figure 21. The figure representations of the 143D topology. The red boxes indicate tetrads and the guanines that take part in each. The arrow indicates the 5’ - 3’ direction of the nucleic acid backbone.

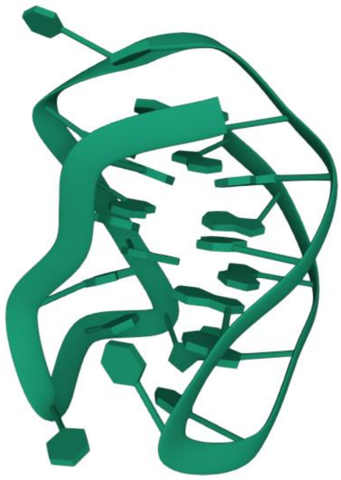
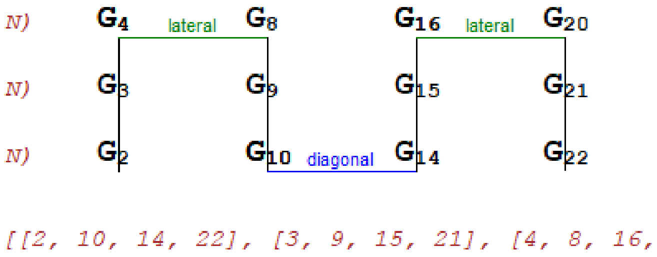
143D	
3D Structure	Secondary structure figure representation
	 N) G4 lateral G8 G16 lateral G20 N) G3 G9 G15 G21 N) G2 G10 diagonal G14 G22 [[2, 10, 14, 22], [3, 9, 15, 21], [4, 8, 16,

Figure 22. 3D Structure and Secondary structure figure representation of 143D DNA. In the figure representation, the bottom list corresponds to the tetrads. Each row in the figure representation corresponds to an individual tetrad with the ONZ classification on the left.

179 PDB structures were available for evaluation of G-quadruplex detection but not classification. The algorithm would fail to identify any tetrad for 1OZ8, 1HAO, 1HAP, 1HUT, 6T2G and 6E84 depending on the parameter choice. Increasing the distance threshold gradually yielded detection of all but 6E84. However, for the detection of 6E84, increasing angle range was also necessary.

47 PDBs were curated to identify their tetrads and loops, and consecutively,

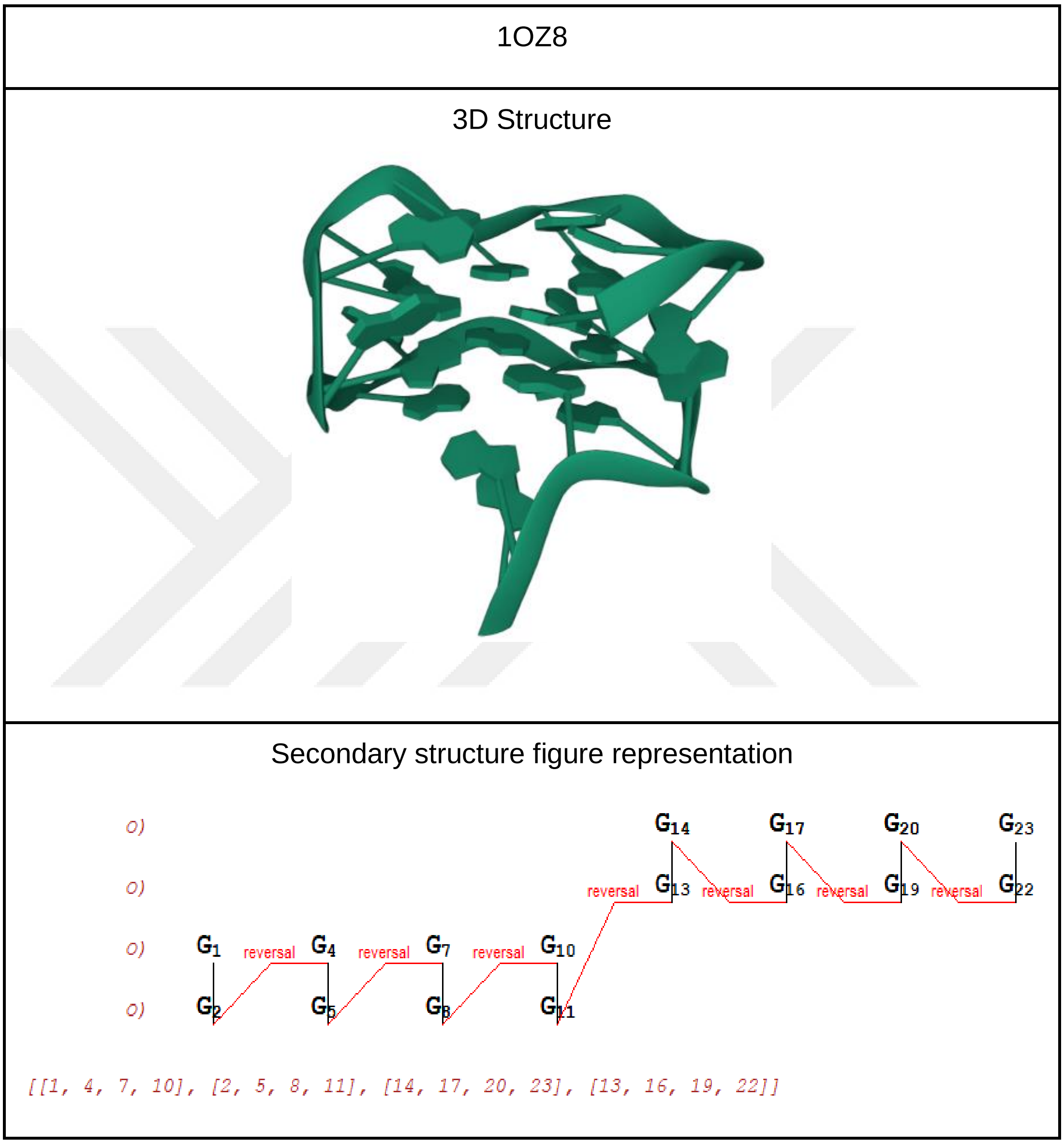
classify the G-quadruplexes. These structures were selected to contain diverse G-quadruplex topology, including all that yielded false negatives with strict parameters and used to determine the accuracy of the algorithm. (see Supplementary Table 1) Out of all, the curated determination had complete agreement with the algorithm findings for a number of parameter combinations. (see table 1) In particular, the algorithm would fail to identify 1OZ8 structure's tetrads with 3.5 Å threshold and 90±35 degree angle range. In this structure 3 out of 4 tetrads are missed due to increased flexibility. For that reason, relaxing thresholds from 3.5 Å to 5 Å and angle range from 90±35 to 90±40 degrees was necessary to identify these tetrads. Similarly, 6T2G was either unidentified as G-quadruplexes with the initial parameters or its tetrads were misidentified until parameters were relaxed.

Table 1. Misidentified and misclassified results according to the different parameters applied

Distance threshold	Angle degree ranges (narrow/wide)	Misclassified count, n= 47 (PDBs missing tetrads)	Misidentified, n= 179 (PDBs missing all)
3.5*	90±35 / 90±70*	0	6 (1hao, 1hap, 1hut,1oz8, 6t2g, 6e84)
3.5	90±40 / 90±70	2 (1oz8, 6t2g)	3 (1hao, 1hap, 1hut)
4	90±35 / 90±60	1 (6t2g)	3 (1oz8, 6e84, 148d)
4	90±35 / 90±70	1 (6t2g)	2 (1oz8, 6e84)
5	90±35 / 90±70	2 (1oz8, 6t2g)	1 (6e84)
5	90±40 / 90±70	0	0

* stringent rules identified according to distribution of H-bond properties.

1OZ8 is a good example where figure representation simplifies the recognition of the structure. This structure has a unique topology with four tetrads and peculiarly the sequence may be split into two at N12, and yet the resulting structures would still form two individual G-quadruplexes. It was enthralling to see that this phenomenon was reflected in secondary structure figure representation with guanines of the two substructures being listed separately, connected through a reversal loop at N12, as can be seen in Figure 23.



```
[[1, 4, 7, 10], [2, 5, 8, 11], [14, 17, 20, 23], [13, 16, 19, 22]]
```

Figure 23. 3D Structure and Secondary structure figure representation of 1OZ8 DNA. In the figure representation, the bottom list corresponds to the tetrads, left part to the ONZ tetrad classes which shows the corresponding tetrad specifically by aligning each class according to the tetrad.

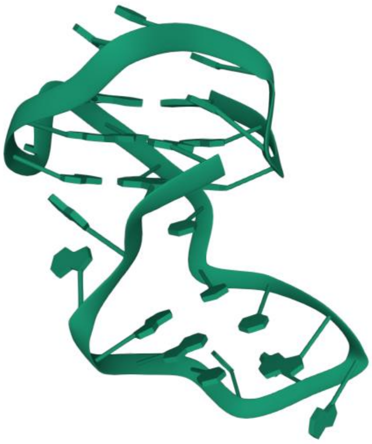
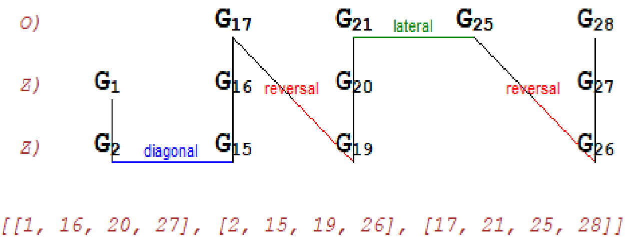
<table>
<tr><th colspan="2">6H1K</th></tr>
<tr><td>3D Structure
</td><td>Secondary structure figure representation

 </td></tr>
</table>

Figure 24. 3D Structure and Secondary structure figure representation of 6H1K DNA. In the figure representation, the bottom list corresponds to the tetrads. Each row in the figure representation corresponds to an individual tetrad with the ONZ classification on the left.

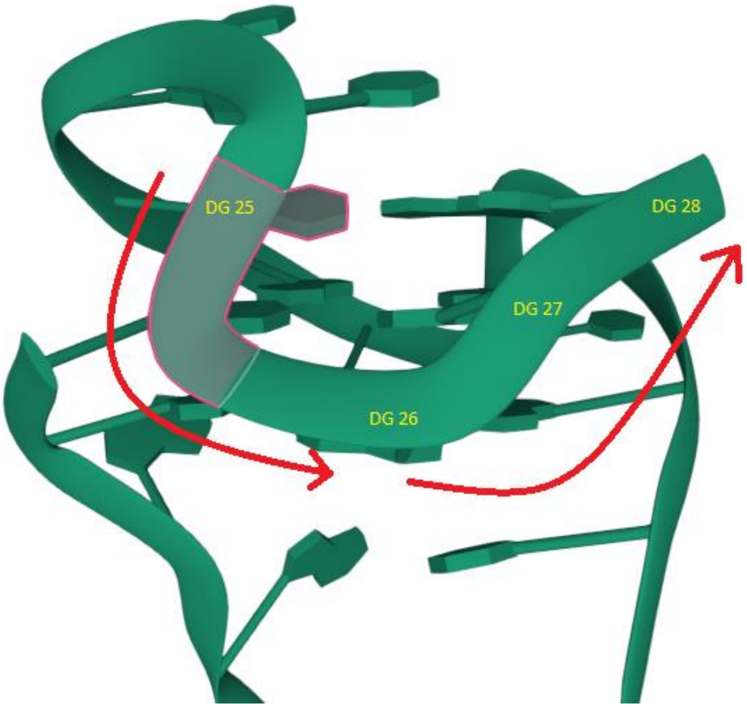
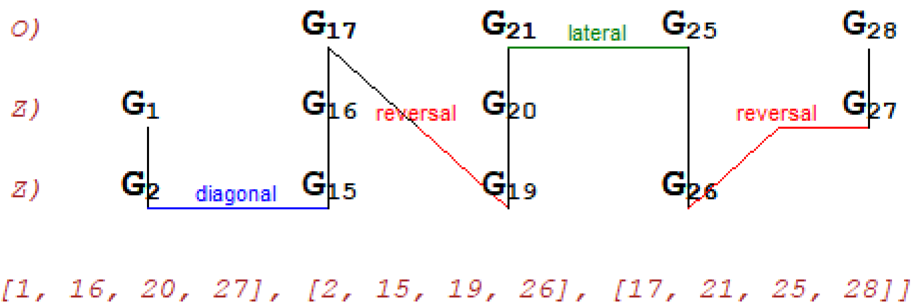


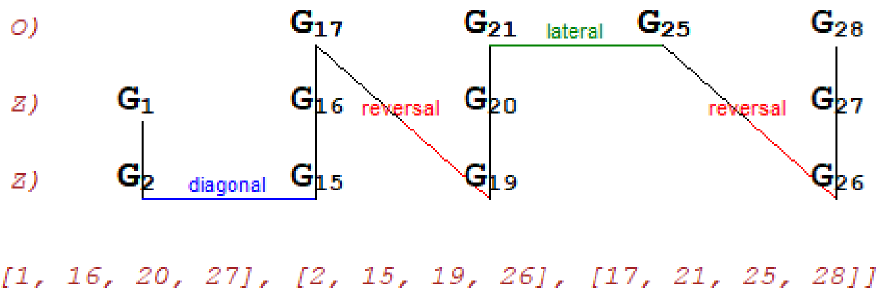
Figure 25. Some part of 6H1K’ s 3D structure

6H1K is one of the structures with unique topology, as can be seen in Figure 24 and 25. Here the structure has 3 tetrads. After the 25th base (G25) the following guanine skips a tetrad and stacks with G27 and G28. The curation concluded that it would be suitable if the loop is placed between G25 and G26, rather than G26 and G27 because of this stacking. The algorithm detects such uncommon motifs by checking the stacks when the strand direction has changed. Without such consideration the loop would have been placed between G26 and G27, showing the importance of stacks for making a conclusive prediction. Figure 26 and 27 show the results of loop placing with or without stacking consideration.



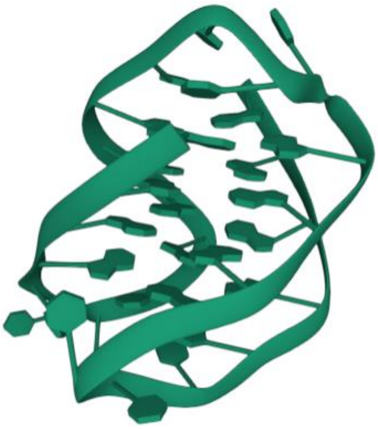
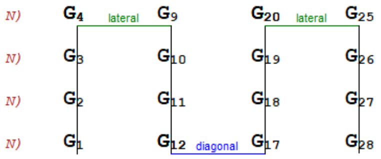
```
[[1, 16, 20, 27], [2, 15, 19, 26], [17, 21, 25, 28]]
```

Figure 26. Loop is placed without considering stacking guanines



```
[[1, 16, 20, 27], [2, 15, 19, 26], [17, 21, 25, 28]]
```

Figure 27. Loop placed according to the stacking guanines

<table>
<tr><th colspan="2">201D</th></tr>
<tr><td>3D Structure
</td><td>Secondary structure figure representation


[[1, 12, 17, 28], [2, 11, 18, 27], [3, 10, 19, 26], [4, 9, 20, 25]]</td></tr>
</table>

Figure 28. 3D Structure and Secondary structure figure representation of 201D DNA. In the figure representation, the bottom list corresponds to the tetrads, left part to the ONZ tetrad classes which shows the corresponding tetrad specifically by aligning each class according to the tetrad

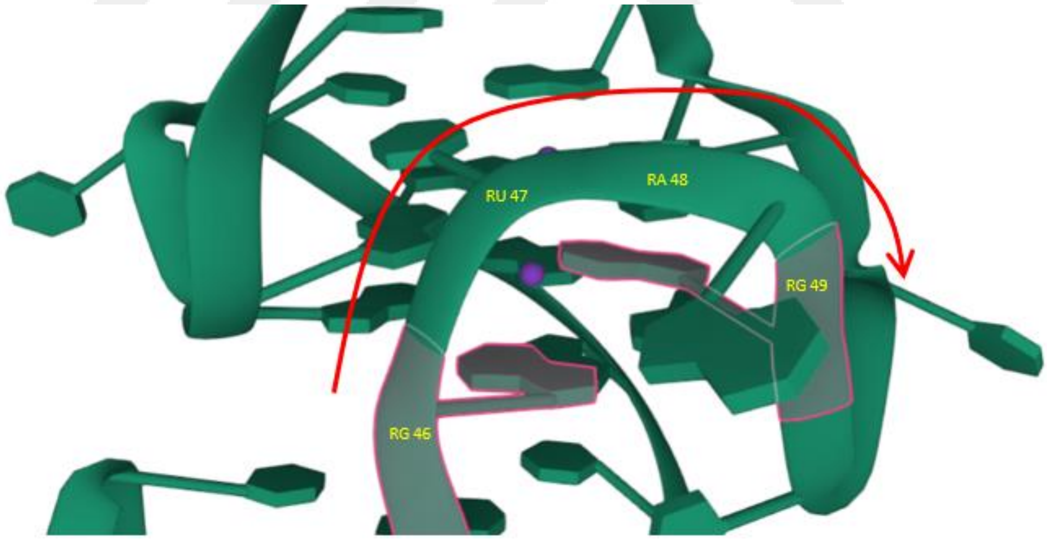


Figure 29. Some part of 5OB3’ s 3D structure

Another example where stacking was required for a decision is 5OB3 structure. This RNA structure consists of two tetrads. Here, G46 and G49 participate in different tetrads, yet between G46 and G49 exist two nucleotides. Curiously, G49 is followed by another base before G51, which shares formation of a tetrad with G46. In such circumstances, the loop may be placed before or after G49, or even in both by curation.

Defining one as the loop would render the other side as bulge rather than loop. Since our algorithm is capable of working with bulges, we decided that stacking should lead to the final decision over which bases would form the loop. Our algorithm concluded that G49 stacks with G46 and thus, N47 and N48 are labelled as the bulge, while N50 is the loop.

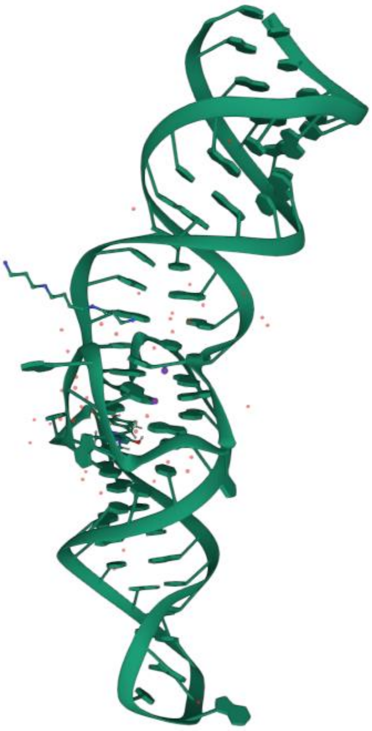
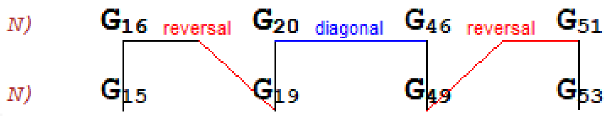
5OB3	
3D Structure	Secondary structure figure representation
	 N) N) G16 reversal G20 diagonal G46 reversal G51 G15 G19 G49 G53 [[15, 19, 49, 53], [16, 20, 46, 51]]

Figure 30. 3D Structure and Secondary structure figure representation of 5OB3 RNA. In the figure representation, the bottom list corresponds to the tetrads, left part to the ONZ tetrad classes which shows the corresponding tetrad specifically by aligning each class according to the tetrad.

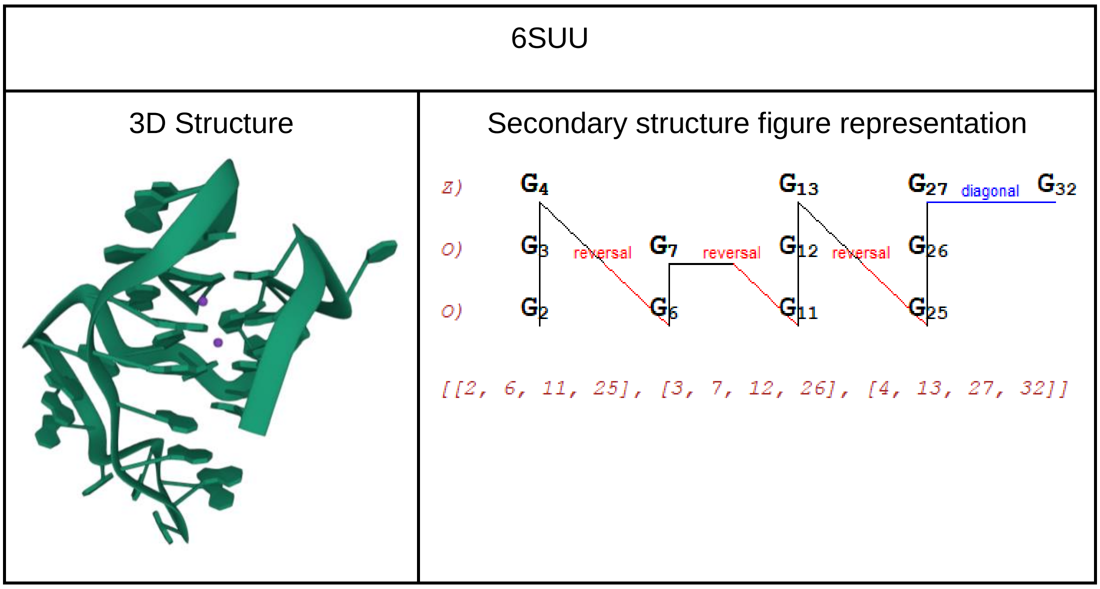


Figure 31. 3D Structure and Secondary structure figure representation of 6SUU DNA. In the figure representation, the bottom list corresponds to the tetrads, left part to the ONZ tetrad classes which shows the corresponding tetrad specifically by aligning each class according to the tetrad.

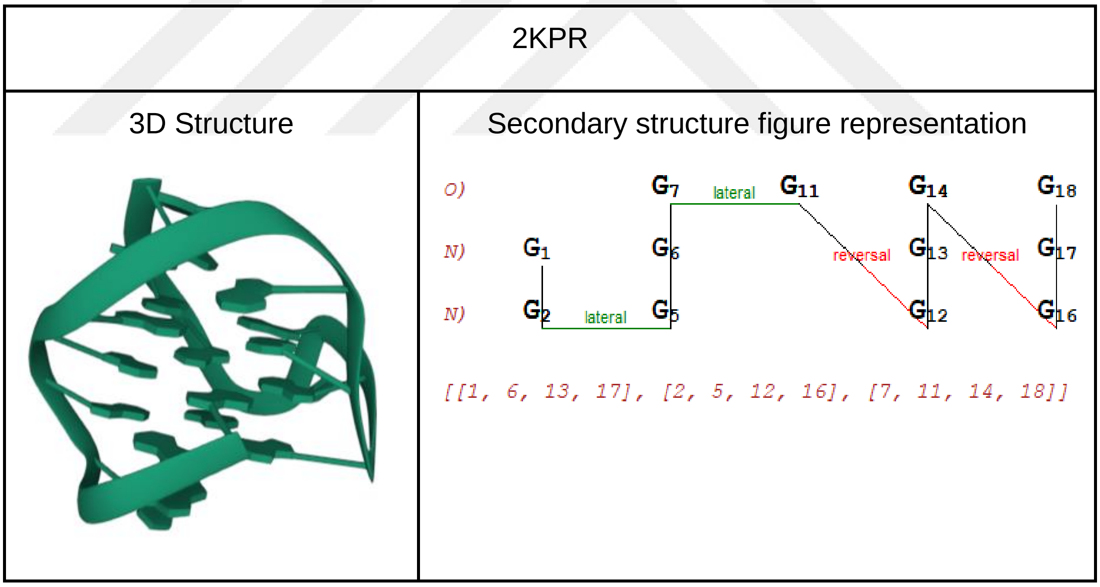


Figure 32. 3D Structure and Secondary structure figure representation of 2KPR DNA. In the figure representation, the bottom list corresponds to the tetrads, left part to the ONZ tetrad classes which shows the corresponding tetrad specifically by aligning each class according to the tetrad.

The structures of 2KPR and 6SUU points to a potential misinterpretation. In 6SUU, the algorithm places G32 to the rightmost of the figure, however, since it stacks with G6 and G7, one might expect that it would be placed on the same column with these bases. Similarly in 2KPR G1, G2 and G11 are stacked but not placed along the same column. However, the algorithm does not consider stacking except when locating the loop. Briefly, one should note that while each row represents a tetrad, each column does not necessarily represent stacked bases, as can be seen in Figure 31 and 32.

4.1. Comparison with ONZ classification

We have compared our results with the ONZ classification method. Among all DNA structures, we have 13 NNN, 29 OO, 79 OOO, 1 ZZ, 9 OOOO, 3 NNNN, 5 OOZ, 11 NN, 7 NNO, 1 NOO, 2 ONN, 1 ZZZ, 2 ZZO, 1 ZO and 1 ZOO classes were identified. Analysis of these structures also showed consistent patterns between the types of the loops and types of ONZ classifications. Firstly, it was observed that if a loop occurs after all tetrads are visited since the last loop or the sequence beginning, there is always a single ONZ class identified for all tetrads. This is actually an expected result. When all of the tetrads have the same ONZ class, the order of guanines in sequence goes from minimum to maximum sequentially from the bottom tetrad through the top one for each region between two consecutive loops. Therefore, loop occurs after each guanine from all tetrads is visited.

Among 13 NNN classes, all structures have 3 loops, where the first and the third are lateral or reversal, and the second one is diagonal. Among 3 NNNN and 11 NN classes, all contain lateral, diagonal and lateral loops in that order. It was apparent that for any G-quadruplex containing only N type tetrad, a diagonal loop is consistently present.

Among 29 OO classes, 27 structures have 3 loops and 2 of them which contain proteins have 4 loops. 22 of them have only lateral loops, while 5 of them have only reversal loops, and 2 contain both reversal and lateral ones. Among 79 OOO classes, 50 of them are structures which contain 3 reversal loops. 2 of them contain 3 lateral loops and the rest contain a mixture of lateral and reversal. Additionally, 78 of these structures contain 3 loops while only 1 of them contains 4 loops. Among 9 OOOO classes, 7 of them contain only reversal loops which have a number at least 6. And 2 of them contain 3 lateral loops. It was surprising to see that with the all-O type G-quadruplexes, we observed that the diagonal loop was absent and the reversal loop was dominant.

There were only 2 all-Z type G-quadruplexes. One being ZZZ and the other

ZZ, both consist of diagonal, reversal and diagonal loops. For all-Z and all-N type structures, it is expected that two guanines that lay across to each other in a tetrad to be consecutive in the sequence order. For Z-types this is possible if there are two diagonal loops, while in N-type, only one diagonal is adequate. This is in accordance with our observations where N-types have exactly one diagonal.

On the other hand, mixed type G-quadruplexes showed characteristics independent from their uniform type counterparts. 7 NNO G-quadruplexes, where all of them followed a lateral, lateral, reversal and reversal loop pattern. In the meanwhile, among 5 OOZ classes, all structures have 4 loops which are reversal, reversal, reversal and diagonal. Other classes which contain Z types also contain at least one diagonal loop. It is apparent that containing N or Z-type tetrad alone did not determine the existence of a diagonal loop. However, it can be said that when N and Z classes are in the structure, the probability of observing the diagonal loop is high.

The comparison of secondary structures determined by the algorithm with ONZ classification reveals that there is a strong correlation between the loop orders with ONZ classification, as seen in Z and N-type containing classes. However, this could be due to the unavailability of a larger sample size. As such, there are a large number of structures with all-O type tetrads, which also does not show a single loop combination. This indicates ONZ classification alone would not be adequate for annotation for secondary structures of the G-quadruplexes.

4.2. Comparison with DSSR

DSSR is a stand alone tool which does not require any other software to find base pairs, tetrads and loops. However, it utilises other tools for visualisation purposes, such as PyMol. In comparison our code is capable of generating a figure representation using python’s turtle library. DSSR has a detailed explanation of strands, tetrads and loops. However, the comprehensive output of DSSR is often hard to understand and grasp the details of the structure. The plain 2D illustration generated by our algorithm presents an advantage, where tetrads can be identified horizontally, and stacking guanines can be distinguished vertically.

Additionally, DSSR requires licensing to get annotation results for G-quadruplex structures. Fortunately, the annotation results for a number of G-quadruplexes were already published at http://g4.x3dna.org/ and we were able to compare. DSSR has three types of loops which are lateral, diagonal and propeller, which corresponds to lateral, diagonal and reversal in our algorithm, respectively.

Another advantage of our method is that it requires only two thresholds, the thresholds of distance and angle parameters that can be modified to detect loosely connected tetrads. Due to this advantage, the identification of the tetrads were possible in at least two structures. In case of 1oz8, DSSR found three tetrads (G1-G4-G7-G10, G13-G16-G19-G22 and G14-G17-G20-G23) as shown at: http://skmatic.x3dna.org/pdb/1oz8/1oz8.out, while our algorithm have found one more tetrad which is G2-G5-G8-G11. In comparison DSSR highlighted G5-G8-G11 as a multiplet, omitting the G2. Based on this difference, loop classification differs with our method. DSSR has identified 3 stem reversal loops and 3 non-stem lateral loops while we have identified 7 reversal loops. Stem loop is defined as any loop that also forms a duplex within itself.

A similar difference exists in 6t2g. DSSR could find 2 tetrads in this structure (G2-G6-G11-G26 and G4-G9-G13-G28) as shown at http://skmatic.x3dna.org/pdb/6t2g/6t2g.out while our algorithm found one more tetrad, G3-G7-G12-G27. DSSR is able to show these four guanines as a multiplet in the list of multiplets section, however does not present it as a tetrad like the other two. As a result, our loop types and placements are also different. DSSR has found six lateral loops while our algorithm has found three reversal loops.

Another advantage of our study is that it is an open source, does not require licensing and is freely available to the scientific community.

CHAPTER 5: CONCLUSION

G4 sub-structures are one of the biomolecules which take important roles and knowing the secondary structure of G4 quadruplexes lets us to predict these important roles. Some different methods were developed to annotate these structures as mentioned in the literature section. 3DNA suite is a stand-alone environment which can identify and annotate different topologies of DNA and RNA. It seems to have many functionalities when compared with the other bioinformatic tools currently existing to annotate any nucleic acids. DSSR is a specialised extension of 3DNA when the RNA structure is analysed. In another study, 3D-NuS models duplex, triplex and quadruplex structures. For quadruplexes, they define 19 different classes and with a user given parameters, they can be categorised into one of these classes. Lastly, they optimise the structures to find the energy minimised version since it is the most stable state of the corresponding structure. RNApdbee is another tool which can identify the G-quadruplexes in PDB format. Lastly, ONZ classification is the most interesting and novel method according to the extent of our knowledge. They categorise tetrads according to the orientation of participating guanines. Therefore, each structure is labelled according to the classes of all tetrads in a topology.

All mentioned works, other than DSSR rely on secondary applications for characterisation of a given G-quadruplex structure while DSSR requires secondary tools like VARNA for visualisation purposes. Our method discovers Hoogsten base pairs, tetrads and loops represented in any G-quadruplex as a stand-alone application. Moreover, a self-explanatory figure which helps comprehension of the structure through a simple visualisation generated without trusting on a secondary tool.

While ONZ classification did have a correlation with our results, some details were lost in the ONZ classification. It is not enough to make a classification for G4s since it looks only for the type of the tetrads. Loops also reflect the characteristic of G4 structures. By considering stacking bases and different loop types, G4s can be classified in a way that expresses the topology by its details. This shows that one classification alone is not enough to distinctly classify G-quadruplexes.

Stacking is crucial to decide the correct placement of loops. 6h1k and 5ob3 are two of the structures, illustrated in the results and discussion section, which show the importance of choosing the right stacking bases. By calculating the closeness of guanine centres, our algorithm is able to catch this important criteria to place the loops correctly.

Thresholds are important to detect loosely connected tetrads. As shown in Table 1, when different values are assigned to angle and distance parameters, different results are obtained by the algorithm. 1oz8 and 6t2g structures are two examples of this implementation which allows the algorithm to identify the loosely connected tetrads. This flexibility of code is one of the properties that makes the biggest difference when compared with other studies.

REFERENCES

3D-NuS: A Web Server for Automated Modelling and Visualisation of Non-Canonical 3-Dimensional Nucleic Acid Structures. (2017). Journal of Molecular Biology, Vol. 429(16), pp. 2438–2448.

Agrawal, P., Hatzakis, E., Guo, K., Carver, M., and Yang, D. (2013). *Major G-quadruplex structure formed in human VEGF promoter, a monomeric parallel-stranded quadruplex*. Worldwide Protein Data Bank.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *The Structure and Function of DNA*. In Molecular Biology of the Cell. 4th edition. Garland Science.

Ambrus, A., Chen, D., Dai, J., Jones, R. A., and Yang, D. (2005). *Major G-quadruplex structure formed in human c-MYC promoter, a monomeric parallel-stranded quadruplex.* Worldwide Protein Data Bank.

Amrane, S., Adrian, M., Heddi, B., Serero, A., Nicolas, A., Mergny, J. L., and Phan, A. T. (2012). *human CEB25 minisatellite G-quadruplex*. Worldwide Protein Data Bank.

Antczak, M., Popenda, M., Zok, T., Zurkowski, M., Adamiak, R. W., and Szachniuk, M. (2018). *New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation*. In Bioinformatics (Vol. 34, Issue 8, pp. 1304–1312).

Antczak, M., Zok, T., Popenda, M., Lukasiak, P., Adamiak, R. W., Blazewicz, J., and Szachniuk, M. (2014). *RNApdbee—a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs*. In Nucleic Acids Research (Vol. 42, Issue W1, pp. W368–W372).

Bakalar, B., Heddi, B., Schmitt, E., Mechulam, Y., and Phan, A. T. (2019). *Structure of a left-handed G-quadruplex.* Worldwide Protein Data Bank.

Bazzicalupi, C., Gratteri, P., and Papi, F. (2019). *Structure of the complex of a human telomeric DNA with bis(1-butyl-3-methyl-imidazole-2-ylidene) gold(I).* Worldwide Protein Data Bank.

Bielskute, S., Plavec, J., and Podbevsek, P. (2019a). *Human telomeric G-quadruplex with 8-oxo-G substitution in the central G-quartet.* Worldwide Protein Data Bank.

Bielskute, S., Plavec, J., and Podbevsek, P. (2019b). *Human telomeric G-quadruplex with 8-oxo-G substitution in the outer G-quartet*. Worldwide Protein Data Bank.

Bolton, P. H., Marathias, V. M., and Wang, K. (1999). *Solution structure of a* *quadruplex forming DNA and its intermediate.* Worldwide Protein Data Bank.

Brcic, J., and Plavec, J. (2015). *Structure of DNA G-quadruplex adopted by ALS and FTD related GGGGCC repeat with G21 to Br-G21 substitution*. Worldwide Protein Data Bank.

Brcic, J., and Plavec, J. (2018). *G-quadruplex structure of DNA oligonucleotide containing GGGGCC repeats linked to ALS and FTD*. Worldwide Protein Data Bank.

Caruso, Í. P., Sanches, K., Da Poian, A. T., Pinheiro, A. S., and Almeida, F. C. L. (n.d.). *Dynamics of the N-terminal domain of SARS-CoV-2 nucleocapsid protein drives dsRNA melting in a counterintuitive tweezer-like mechanism.*

Chen, X., and Walters, K. J. (2018). *Solution structure of the MYC G-quadruplex bound to small molecule DC-34*. Worldwide Protein Data Bank.

Cheong, V. V., Heddi, B., Lech, C. J., and Phan, A. T. (2015). *A structure of G-quadruplex.* Worldwide Protein Data Bank.

Chung, W. J., Heddi, B., Hamon, F., Teulade-Fichou, M. P., and Phan, A. T. (2014). *Solution structure of a G-quadruplex bound to the bisquinolinium compound Phen-DC3.* Worldwide Protein Data Bank.

Chung, W. J., Heddi, B., Schmitt, E., Lim, K. W., Mechulam, Y., and Phan, A. T. (2015). *Solution structure of a G-quadruplex.* Worldwide Protein Data Bank.

Chung, W. J., Heddi, B., Tera, M., Iida, K., Nagasawa, K., and Phan, A. T. (2013). *Solution structure of an intramolecular (3+1) human telomeric G-quadruplex bound to a telomestatin derivative*. Worldwide Protein Data Bank.

Colasanti, A. V., Lu, X.-J., and Olson, W. K. (2013). *Analysing and building nucleic acid structures with 3DNA*. Journal of Visualised Experiments: JoVE, 74, e4401.

Collie, G. W., Haider, S. M., Neidle, S., and Parkinson, G. N. (2010). *A crystallographic and modelling study of a human telomeric RNA (TERRA) quadruplex.* Nucleic Acids Research, Vol. 38(16), pp. 5569–5580.

Collie, G. W., and Neidle, S. (2013a). *Crystal structure of an intramolecular human telomeric DNA G-quadruplex bound by the naphthalene diimide compound, MM41.* Worldwide Protein Data Bank.

Collie, G. W., and Neidle, S. (2013b). *Crystal structure of an intramolecular human telomeric DNA G-quadruplex 21-mer bound by the naphthalene diimide compound BMSG-SH-3.* Worldwide Protein Data Bank.

Collie, G. W., and Neidle, S. (2013c). *Crystal structure of an intramolecular human telomeric DNA G-quadruplex 21-mer bound by the naphthalene diimide compound*

MM41. Worldwide Protein Data Bank.

Collie, G. W., Promontorio, R., and Parkinson, G. N. (2012a). *Crystal structure of an intramolecular human telomeric DNA G-quadruplex bound by the naphthalene diimide BMSG-SH-3*. Worldwide Protein Data Bank.

Collie, G. W., Promontorio, R., and Parkinson, G. N. (2012b). *Crystal structure of an intramolecular human telomeric DNA G-quadruplex bound by the naphthalene diimide BMSG-SH-4*. Worldwide Protein Data Bank.

Dai, J., Carver, M., Mathad, R., and Yang, D. (2011). *Quindoline/G-quadruplex complex.* Worldwide Protein Data Bank.

Dai, J., Carver, M., Punchihewa, C., Jones, R., and Yang, D. (2007*). Human telomere DNA quadruplex structure in K+ solution hybrid-2 form.* Worldwide Protein Data Bank.

Dai, J., Chen, D., Carver, M., and Yang, D. (2006). *G-quadruplex structure formed in human Bcl-2 promoter, hybrid form*. Worldwide Protein Data Bank.

Dai, J., Punchihewa, C., Jones, R. A., Hurley, L., and Yang, D. (2007*). Human telomere DNA quadruplex structure in K+ solution hybrid-1 form*. Worldwide Protein Data Bank.

DeNicola, B., Lech, C. J., Heddi, B., Regmi, S., Frasson, I., Perrone, R., Richter, S. N., and Phan, A. T. (2016). *Structure and possible function of a G-quadruplex in the long terminal repeat of the proviral HIV-1 genome*. Worldwide Protein Data Bank.

Dickerhoff, J., Haase, L., Langel, W., and Weisz, K. (2017a). *Quadruplex with flipped tetrad formed by a human telomeric sequence*. Worldwide Protein Data Bank.

Dickerhoff, J., Haase, L., Langel, W., and Weisz, K. (2017b). *Quadruplex with flipped tetrad formed by an artificial sequence*. Worldwide Protein Data Bank.

Dickerhoff, J., Onel, B., Chen, L., Chen, Y., and Yang, D. (2019). *MYC Promoter G-Quadruplex with 1:6:1 loop length*. Worldwide Protein Data Bank.

Dickerhoff, J., and Weisz, K. (2017). 2’F-ANA-G modified quadruplex with a flipped tetrad. Worldwide Protein Data Bank. https://doi.org/10.2210/pdb5ov2/pdb

Dickerhoff, J., and Weisz, K. (2018). *2’F-araG modified quadruplex with flipped G-tract and central tetrad*. Worldwide Protein Data Bank.

Do, N. Q., Chung, W. J., Truong, T. H. A., Heddi, B., and Phan, A. T. (2016). *G-quadruplex structure of an anti-proliferative DNA sequence*. Worldwide Protein Data Bank.

Do, N. Q., and Phan, A. T. (2012). *Monomer-dimer equilibrium for 5’-5' stacking of* *propeller-type parallel-stranded G-quadruplexes: NMR structural study*. Worldwide Protein Data Bank.

Dvorkin, S. A., Karsisiotis, A. I., and Webba da Silva, M. (2017a). *DIY G-Quadruplexes: Solution structure of d(GGGTTTGGGTTTTGGGAGGG) in sodium*. Worldwide Protein Data Bank.

Dvorkin, S. A., Karsisiotis, A. I., and Webba da Silva, M. (2017b). *DIY G-Quadruplexes: Solution Structure of d(GGGGTTTGGGGTTTTGGGGAAGGGG) in sodium.* Worldwide Protein Data Bank.

Dvorkin, S. A., and Webba da Silva, M. (2017a). *DIY G-Quadruplexes: Solution structure of d(GGTTTGGTTTTGGTTGG) in sodium*. Worldwide Protein Data Bank.

Dvorkin, S. A., and Webba da Silva, M. (2017b). *DIY G-Quadruplexes: Solution structure of d(GGTTTGGTTTTGGTTTGG) in sodium*. Worldwide Protein Data Bank

Fernandez-Millan, P., Autour, A., Ennifar, E., Westhof, E., and Ryckelynck, M. (2017*). Crystal structure and fluorescence properties of the iSpinach aptamer in complex with DFHBI*. RNA, Vol. 23(12), pp. 1788–1795.

Ferraroni, M., Bazzicalupi, C., Gratteri, P., and Bilia, A. R. (2012). *Structure of the complex of an intramolecular human telomeric DNA with Berberine formed in K+ solution.* Worldwide Protein Data Bank.

Ferraroni, M., Bazzicalupi, C., Gratteri, P., Messori, L., and Papi, F. (2016). *Structure of the complex of a human telomeric DNA with Au(caffein-2-ylidene)2*. Worldwide Protein Data Bank.

Galer, P., Wang, B., Sket, P., and Plavec, J. (2018a). *A two-quartet G-quadruplex formed by human telomere in KCl solution at pH 5.0*. Worldwide Protein Data Bank.

Galer, P., Wang, B., Sket, P., and Plavec, J. (2018b). *A two-quartet G-quadruplex formed by human telomere in KCl solution at neutral pH*. Worldwide Protein Data Bank.

Geng, Y., Cai, Q., Liu, C., and Zhu, G. (2019). *Crystal structure of G-quadruplex formed by bromo-substituted human telomeric DNA*. Worldwide Protein Data Bank.

Gomez-Pinto, I., Vengut-Climent, E., Lucas, R., Avio, A., Eritja, R., Gonzalez-Ibaez, C., and Morales, J. (2014). *Fuc_TBA*. Worldwide Protein Data Bank.

Guerra, C. F., Matthias Bickelhaupt, F., Snijders, J. G., and Baerends, E. J. (2000). *Hydrogen Bonding in DNA Base Pairs: Reconciliation of Theory and Experiment.* In Journal of the American Chemical Society (Vol. 122, Issue 17, pp. 4117–4128).

Haase, L., Dickerhoff, J., and Weisz, K. (2018a). *DNA-RNA hybrid quadruplex with* *flipped tetrad.* Worldwide Protein Data Bank.

Haase, L., Dickerhoff, J., and Weisz, K. (2018b). *DNA-RNA Hybrid Quadruplexes Reveal Interactions that Favor RNA Parallel Topologies*. Chemistry, Vol. 24(57), pp. 15365–15371.

Haase, L., and Weisz, K. (2019). *2'-F-riboguanosine modified G-quadruplex with V-loop.* Worldwide Protein Data Bank.

Haase, L., and Weisz, K. (2020a). *2'-F-arabinoguanosine and 2'-F-riboguanosine modified hybrid type G-quadruplex with V-loop*. Worldwide Protein Data Bank.

Haase, L., and Weisz, K. (2020b). *2'-F-riboguanosine and 2'-F-arabinoguanosine modified G-quadruplex with V-loop and all-syn G-tract*. Worldwide Protein Data Bank.

Heddi, B., Martin-Pintado, N., Serimbetov, Z., Kari, T. M., and Phan, A. T. (2016). *G-quadruplexes with (4n-1) guanines in the G-tetrad core: formation of a G-triad water complex and implication for small-molecule binding*. Worldwide Protein Data Bank.

Heddi, B., and Phan, A. T. (2011). *Structure of human telomeric DNA in crowded solution.* Worldwide Protein Data Bank.

Hsu, S.-T. D., Varnai, P., Bugaut, A., Reszka, A. P., Neidle, S., and Balasubramanian, S. (2009). *A G-rich sequence within the c-kit oncogene promoter forms a parallel G-quadruplex having asymmetric G-tetrad dynamics*. Worldwide Protein Data Bank.

Hu, L., Lim, K. W., Bouaziz, S., and Phan, A. T. (2009). *Structure of a two-G-tetrad basket-type intramolecular G-quadruplex formed by Giardia telomeric repeat d(TAGGG)4 in K+ solution (with G18-to-INO substitution).* Worldwide Protein Data Bank.

Juribasic Kulcsar, M., and Plavec, J. (2018a). *G-quadruplex formed within promoters of Plasmodium falciparum B var genes*. Worldwide Protein Data Bank.

Juribasic Kulcsar, M., and Plavec, J. (2018b). *G-quadruplex formed within promoters of Plasmodium falciparum B var genes - form I.* Worldwide Protein Data Bank.

Karg, B., and Weisz, K. (2018). *Quadruplex with flipped tetrad formed by the c-myc promoter sequence.* Worldwide Protein Data Bank.

Karg, B., and Weisz, K. (2019a). *A quadruplex hybrid structure with lpp loop orientation and 3 syn residues.* Worldwide Protein Data Bank.

Karg, B., and Weisz, K. (2019b). *A quadruplex hybrid structure with lpp loop orientation and 5 syn residues.* Worldwide Protein Data Bank.

Karsisiotis, A., Dillon, P., and Webba da Silva, M. (2014). *Solution NMR structure of* *quadruplex d(TGGGTTTGGGTTGGGTTTGGG) in sodium conditions*. Worldwide Protein Data Bank.

Karsisiotis, A. I., and Webba da Silva, M. (2014a). *Solution NMR structure of the d(GGGTTGGGTTTTGGGTGGG) quadruplex in sodium conditions.* Worldwide Protein Data Bank.

Karsisiotis, A. I., and Webba da Silva, M. (2014b). *Solution NMR structure of the d(GGGGTTGGGGTTTTGGGGAAGGGG) quadruplex in sodium conditions*. Worldwide Protein Data Bank.

Karsisiotis, A. I., and Webba da Silva, M. (2014c). *Solution NMR structure of the d(GGGTTTTGGGTGGGTTTTGGG) quadruplex in sodium conditions*. Worldwide Protein Data Bank.

Kosiol, N., Juranek, S., Brossart, P., Heine, A., and Paeschke, K. (2021). *G-quadruplexes: a promising target for cancer therapy.* Molecular Cancer, Vol. 20(1), pp. 1–18.

Kotar, A., Rigo, R., Sissi, C., and Plavec, J. (2019*). Two-quartet kit* G-quadruplex is formed via double-stranded pre-folded structure*. Worldwide Protein Data Bank.

Kotar, A., Wang, B., Shivalingam, A., Gonzalez-Garcia, J., Vilar, R., and Plavec, J. (2016). *G-Quadruplex formed at the 5'-end of NHEIII_1 Element in human c-MYC promoter bound to triangulenium based fluorescence probe DAOTA-M2*. Worldwide Protein Data Bank.

Kumar, A., and Tawani, A. (2016*). Solution structure for quercetin complexed with c-myc G-quadruplex DNA.* Worldwide Protein Data Bank.

Kuryavyi, V., Majumdar, A., Shallop, A., Chernichenko, N., Skripkin, E., Jones, R., and Patel, D. J. (2001). *Solution DNA quadruplex with double chain reversal loop and two diagonal loops connecting gggg tetrads flanked by g-(t-t) triad and t-t-t triple*. Worldwide Protein Data Bank.

Kuryavyi, V., and Patel, D. J. (2010). *Monomeric intronic human chl1 gene quadruplex DNA NMR, 17 structures.* Worldwide Protein Data Bank.

Kuryavyi, V., Phan, A. T., and Patel, D. J. (2010*). Monomeric Human CKIT-2 proto-oncogene promoter quadruplex DNA NMR, 12 structures*. Worldwide Protein Data Bank.

Kuryavyi, V. V., and Patel, D. J. (2012). *Monomeric PilE G-Quadruplex DNA from Neisseria Gonorrhoeae*. Worldwide Protein Data Bank.

Kuryavyi, V. V., Phan, A. T., Luu, K. N., and Patel, D. J. (2008). *Monomeric human* *telomere DNA tetraplex with 3+1 strand fold topology, two edgewise loops and double-chain reversal loop, form 2 15BrG, NMR, 10 structures*. Worldwide Protein Data Bank.

Lech, C., Heddi, B., Adrian, M., Li, Z., and Phan, A. T. (2015). *Structure of a G-quadruplex containing a single LNA modification.* Worldwide Protein Data Bank.

Lenarcic Zivkovic, M., Rozman, J., and Plavec, J. (2018). *Adenine-driven structural switch from two- to three-quartet DNA G-quadruplex*. Worldwide Protein Data Bank.

Lietard, J., Abou Assi, H., Gomez-Pinto, I., Gonzalez, C., Somoza, M. M., and Damha, M. J. (2017). *2'F-ANA/DNA Chimeric TBA Quadruplex structure*. Worldwide Protein Data Bank.

Lim, K. W., Alberti, P., Guedin, A., Lacroix, L., Riou, J. F., Royle, N. J., Mergny, J. L., and Phan, A. T. (2009). *Structure of an intramolecular G-quadruplex containing a G.C.G.C tetrad formed by human telomeric variant CTAGGG repeats.* Worldwide Protein Data Bank.

Lim, K. W., Amrane, S., Bouaziz, S., Xu, W., Mu, Y., Patel, D. J., Luu, K. N., and Phan, A. T. (2009a). *Structure of a two-G-tetrad basket-type intramolecular G-quadruplex formed by human telomeric repeats in K+ solution.* Worldwide Protein Data Bank.

Lim, K. W., Amrane, S., Bouaziz, S., Xu, W., Mu, Y., Patel, D. J., Luu, K. N., and Phan, A. T. (2009b). *Structure of a two-G-tetrad basket-type intramolecular G-quadruplex formed by human telomeric repeats in K+ solution (with G7-to-BRG substitution).* Worldwide Protein Data Bank.

Lim, K. W., Lacroix, L., Yue, D. J. E., Lim, J. K. C., Lim, J. M. W., and Phan, A. T. (2010*). Structure of a (3+1) G-quadruplex formed by hTERT promoter sequence.* Worldwide Protein Data Bank.

Lim, K. W., Ng, V. C. M., Martin-Pintado, N., Heddi, B., and Phan, A. T. (2013). *Structure of an antiparallel (2+2) G-quadruplex formed by human telomeric repeats in Na+ solution (with G22-to-BrG substitution).* Worldwide Protein Data Bank.

Lim, K. W., and Phan, A. T. (2013a). *Structure of d[AGGGTGGGTGCTGGGGCGCGAAGCATTCGCGAGG] quadruplex-duplex hybrid.* Worldwide Protein Data Bank.

Lim, K. W., and Phan, A. T. (2013b). *Structure of d[GCGCGAAGCATTCGCGGGGAGGTGGGGAAGGG] quadruplex-duplex hybrid*. Worldwide Protein Data Bank.

Lim, K. W., and Phan, A. T. (2013c). *Structure of d[GGGAAGGGCGCGAAGCATTCGCGAGGTAGG] quadruplex-duplex hybrid.* Worldwide Protein Data Bank.

Lim, K. W., and Phan, A. T. (2013d). *Structure of d[GGTTGGCGCGAAGCATTCGCGGGTTGG] quadruplex-duplex hybrid.* Worldwide Protein Data Bank.

Lim, K. W., and Phan, A. T. (2013e). *Structure of d[TTGGGTGGGCGCGAAGCATTCGCGGGGTGGGT] quadruplex-duplex hybrid.* Worldwide Protein Data Bank.

Lin, C., Liu, W., and Yang, D. (2019). *NMR structure of the 2:1 complex of a carbazole derivative BMVC bound to c-MYC G-quadruplex*. Worldwide Protein Data Bank.

Lin, C., and Yang, D. Z. (2018). *Hybrid-2 form Human Telomeric G-quadruplex in Complex with Epiberberine.* Worldwide Protein Data Bank.

Liu, C., Zhou, B., Kuryavyi, V. V., and Zhu, G. (2018). *The structure of a chair-type G-quadruplex of the human telomeric variant in K+ solution*. Worldwide Protein Data Bank.

Liu, W., Lin, C., and Yang, D. (2019*). NMR structure of the 1:1 complex of a carbazole derivative BMVC bound to c-MYC G-quadruplex*. Worldwide Protein Data Bank.

Liu, W. T., Zhong, Y. F., Liu, L. Y., Zeng, W. J., Wang, F. Y., Yang, D. Z., and Mao, Z. W. (2018). *Solution structure for the 1:1 complex of a platinum(II)-based tripod bound to a hybrid-1 human telomeric G-quadruplex*. Worldwide Protein Data Bank.

Liu, Y., and Lan, W. X. (2018). *Solution structure of G-quadruplex formed in vegfr-2 proximal promoter sequence*. Worldwide Protein Data Bank.

Luu, K. N., Phan, A. T., Kuryavyi, V. V., Lacroix, L., and Patel, D. J. (2006). *Monomeric human telomere DNA tetraplex with 3+1 strand fold topology, two edgewise loops and double-chain reversal loop, NMR, 12 structures*. Worldwide Protein Data Bank.

Lu, X.-J., and Olson, W. K. (2003). *3DNA: a software package for the analysis, rebuilding and visualisation of three-dimensional nucleic acid structures.* Nucleic Acids Research, Vol. 31(17), pp. 5108–5121.

Lu, X.-J., and Olson, W. K. (2008*). 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualisation of three-dimensional nucleic-acid structures.* Nature Protocols, Vol. 3(7), pp. 1213–1227.

Maity, A., Winnerdy, F. R., Chang, W. D., Chen, G., and Phan, A. T. (2020*). Structure* *of an intra-locked G-quadruplex*. Worldwide Protein Data Bank.

Mao, X., Marky, L. A., and Gmeiner, W. H. (2003). *NMR structure of the thrombin-binding DNA aptamer stabilised by Sr2+*. Worldwide Protein Data Bank.

Marathias, V. M., Beger, R. D., and Bolton, P. H. (1998). *Determination of internuclear angles of DNA using paramagnetic assisted magnetic alignment*. Worldwide Protein Data Bank.

Marathias, V. M., Wang, K. Y., Kumar, S., Swaminathan, S., and Bolton, P. H. (1996). *THE NMR STUDY OF DNA QUADRUPLEX STRUCTURE, APTAMER (15MER) DNA*. Worldwide Protein Data Bank.

Marquevielle, J., and Salgado, G. (2019). *NMR structure of KRAS22RT G-quadruplex forming within KRAS promoter region at physiological temperature*. Worldwide Protein Data Bank.

Marusic, M., and Plavec, J. (2018). *G-quadruplex of Human papillomavirus type 52*. Worldwide Protein Data Bank.

Marusic, M., Sket, P., Bauer, L., Viglasky, V., and Plavec, J. (2012). *Solution-state structure of an intramolecular G-quadruplex with propeller, diagonal and edgewise loops*. Worldwide Protein Data Bank.

Mashima, T., Lee, J.-H., Kamatari, Y. O., Hayashi, T., Nagata, T., Nishikawa, F., Nishikawa, S., Kinoshita, M., Kuwata, K., and Katahira, M. (2020). *Development and structural determination of an anti-PrP aptamer that blocks pathological conformational conversion of prion protein*. Scientific Reports, Vol. 10(1), pp. 4934.

Mathad, R. I., Hatzakis, E., Dai, J., and Yang, D. (2011). *c-MYC promoter G-quadruplex formed at the 5'-end of NHE III1 element: insights into biological relevance and parallel-stranded G-quadruplex stability*. Nucleic Acids Research, Vol. 39(20), pp. 9023–9033.

Matsugami, A., Okuizumi, T., Uesugi, S., and Katahira, M. (2004). *Intramolecular higher-order packing of parallel quadruplexes comprising a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad of GGA triplet repeat DNA*. Worldwide Protein Data Bank.

Matsugami, A., Xu, Y., Noguchi, Y., Sugiyama, H., and Katahira, M. (2007). *Human Telomeric DNA mixed-parallel/antiparallel quadruplex under Physiological Ionic Conditions Stabilised by Proper Incorporation of 8-Bromoguanosines*. Worldwide Protein Data Bank.

Miskiewicz, J., Sarzynska, J., and Szachniuk, M. (2021). *How bioinformatics* *resources work with G4 RNAs*. Briefings in Bioinformatics, Vol. 22(3).

Mukundan, V. T., and Phan, A. T. (2013). *Solution structure of an intramolecular propeller-type G-quadruplex containing a single bulge*. Worldwide Protein Data Bank.

Nguyen, T. Q. N., Lim, K. W., and Phan, A. T. (2020). *Solution structure of an intramolecular G-quadruplex containing a duplex bulge*. Worldwide Protein Data Bank.

Nicoludis, J. M., Miller, S. T., Jeffrey, P., Lawton, T. J., Rosenzweig, A. C., and Yatsunyk, L. A. (2012). *Crystal structure of the complex of a human telomeric repeat G-quadruplex and N-methyl mesoporphyrin IX (P21212).* Worldwide Protein Data Bank.

Parkinson, G. N., Lee, M. P. H., and Neidle, S. (2002a). *Structure and packing of human telomeric DNA.* Worldwide Protein Data Bank.

Parkinson, G. N., Lee, M. P. H., and Neidle, S. (2002b). *Structure of the Human G-quadruplex reveals a novel topology*. Worldwide Protein Data Bank.

Phan, A. T., Heddi, B., Butovskaya, E., Bakalar, B., and Richter, S. N. (2018). *The major G-quadruplex form of HIV-1 LTR.* Worldwide Protein Data Bank.

Phan, A. T., Kuryavyi, V. V., Burge, S., Neidle, S., and Patel, D. J. (2007). *Monomeric G-DNA tetraplex from human C-kit promoter.* Worldwide Protein Data Bank.

Phan, A. T., Kuryavyi, V. V., Gaw, H. Y., and Patel, D. J. (2005a). *Complex of tetra-(4-n-methylpyridyl) porphin with monomeric parallel-stranded DNA tetraplex, snap-back 3+1 3' G-tetrad, single-residue chain reversal loops, GAG triad in the context of GAAG diagonal loop, C-MYC promoter, NMR, 6 struct*. Worldwide Protein Data Bank.

Phan, A. T., Kuryavyi, V. V., Gaw, H. Y., and Patel, D. J. (2005b). *Monomeric parallel-stranded DNA tetraplex with snap-back 3+1 3' G-tetrad, single-residue chain reversal loops, GAG triad in the context of GAAG diagonal loop, NMR, 8 struct.* Worldwide Protein Data Bank.

Popenda, M., Miskiewicz, J., Sarzynska, J., Zok, T., and Szachniuk, M. (2019). *Topology-based classification of tetrads and quadruplex structures.* Bioinformatics, Vol. 36(4), pp. 1129–1134.

Randazzo, A., Esposito, V., Ohlenschläger, O., Ramachandran, R., and Mayola, L. (2004). *NMR solution structure of a parallel LNA quadruplex*. Nucleic Acids Research, Vol. 32(10), pp. 3083–3092.

Randazzo, A., Martino, L., Virno, A., Mayol, L., and Giancola, C. (2007). *NMR structure of a new modified Thrombin Binding Aptamer containing a 5'-5' inversion of polarity site.* Worldwide Protein Data Bank.

Rhodes, D., and Lipps, H. J. (2015). *G-quadruplexes and their regulatory roles in biology*. Nucleic Acids Research, Vol. 43(18), pp. 8627–8637.

Saikrishnan, K., Nuthanakanti, A., Srivatsan, S. G., and Ahmad, I. (2019a). *Structure of human telomeric DNA at 1.4 Angstroms resolution.* Worldwide Protein Data Bank.

Saikrishnan, K., Nuthanakanti, A., Srivatsan, S. G., and Ahmad, I. (2019b). *Structure of human telomeric DNA with 5-Selenophene-modified deoxyuridine at residue 11*. Worldwide Protein Data Bank.

Saikrishnan, K., Nuthanakanti, A., Srivatsan, S. G., and Ahmad, I. (2019c). *Structure of human telomeric DNA with 5-selenophene-modified deoxyuridine at residue 12.* Worldwide Protein Data Bank.

Salgado, G. F., Kerkour, A., and Mergny, J.-L. (2016). *NMR structure of a new G-quadruplex forming sequence within the KRAS proto-oncogene promoter region*. Worldwide Protein Data Bank.

Santana, A., Serrano, I., Montalvillo-Jimenez, L., Corzana, F., Bastida, A., Jimenez-Barbero, J., Gonzalez, C., and Asensio, J. L. (2019). *The 1,8-bis(aminomethyl)anthracene and Quadruplex-duplex junction complex.* Worldwide Protein Data Bank.

Schmitt, E., Mechulam, Y., Phan, A. T., Brahim, H., Chung, W. J., and Lim, K. W. (2015). *Structure of a left-handed DNA G-quadruplex.* Worldwide Protein Data Bank.

Schultze, P., Macaya, R. F., and Feigon, J. (1994). *Three-dimensional solution structure of the thrombin binding DNA aptamer d(ggttggtgtggttgg).* Worldwide Protein Data Bank.

Sengar, A., Heddi, B., and Phan, A. T. (2014). *parallel-stranded G-quadruplex in DNA poly-G stretches.* Worldwide Protein Data Bank.

Sengar, A., Vandana, J. J., Chambers, V. S., Di Antonio, M., Winnerdy, F. R., Balasubramanian, S., and Phan, A. T. (2019). *Structure of a (3+1) hybrid G-quadruplex in the PARP1 promoter*. Worldwide Protein Data Bank.

Sjekloća, L., and Ferré-D'Amaré, A. R. (2019). *Binding between G-quadruplexes at the Homodimer Interface of the Corn RNA Aptamer Strongly Activates Thioflavin T Fluorescence.* Cell Chemical Biology, Vol. 26(8), pp. 1159–1168.e4.

Tong, X., and Cao, C. (2011). *Solution structure of all parallel G-quadruplex formed* *by the oncogene RET promoter sequence.* Worldwide Protein Data Bank.

Trachman, R. J., 3rd, Autour, A., Jeng, S. C. Y., Abdolahzadeh, A., Andreoni, A., Cojocaru, R., Garipov, R., Dolgushina, E. V., Knutson, J. R., Ryckelynck, M., Unrau, P. J., and Ferré-D'Amaré, A. R. (2019). *Structure and functional reselection of the Mango-III fluorogenic RNA aptamer*. Nature Chemical Biology, Vol. 15(5), pp. 472–479.

Trachman, R. J., 3rd, Stagno, J. R., Conrad, C., Jones, C. P., Fischer, P., Meents, A., Wang, Y. X., and Ferré-D'Amaré, A. R. (2019). *Co-crystal structure of the iMango-III fluorescent RNA aptamer using an X-ray free-electron laser*. Acta Crystallographica. Section F, Structural Biology and Crystallisation Communications, Vol. 75(Pt 8), pp. 547–551.

Trajkovski, M., da Silva, M. W., and Plavec, J. (2012). *Unique structural features of interconverting monomeric and dimeric G-quadruplexes adopted by a sequence from the intron of the N-myc gene*. Journal of the American Chemical Society, Vol. 134(9), pp. 4132–4141.

Trajkovski, M., Plavec, J., Endoh, T., Tateishi-Karimata, H., and Sugimoto, N. (2018a*). M2 G-quadruplex 10 wt% PEG8000.* Worldwide Protein Data Bank.

Trajkovski, M., Plavec, J., Endoh, T., Tateishi-Karimata, H., and Sugimoto, N. (2018b). *M2 G-quadruplex 20 wt% ethylene glycol*. Worldwide Protein Data Bank.

Trajkovski, M., Plavec, J., Endoh, T., Tateishi-Karimata, H., and Sugimoto, N. (2018c). *M2 G-quadruplex dilute solution.* Worldwide Protein Data Bank.

Wang, Y., and Patel, D. J. (1994a). *Solution structure of the human telomeric repeat d(ag3[t2ag3]3) of the G-quadruplex*. Worldwide Protein Data Bank.

Wang, Y., and Patel, D. J. (1994b). *Solution structure of the tetrahymena telomeric repeat d(t2g4)4 g-tetraplex*. Worldwide Protein Data Bank.

Wang, Y., and Patel, D. J. (1995). *Solution structure of the oxytricha telomeric repeat d[g4(t4g4)3] g-tetraplex*. Worldwide Protein Data Bank.

Wang, Z. F., Li, M. H., Chu, I. T., Winnerdy, F. R., Phan, A. T., and Chang, T. C. (2019a). *A basket type G-quadruplex in WNT DNA promoter.* Worldwide Protein Data Bank.

Wang, Z. F., Li, M. H., Chu, I. T., Winnerdy, F. R., Phan, A. T., and Chang, T. C. (2019b). *WNT DNA promoter mutant G-quadruplex.* Worldwide Protein Data Bank.

Watson, J. D., and Crick, F. H. C. (1953). *Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.* In Nature (Vol. 171, Issue 4356, pp. 737–

738).

Weisz, K., and Haase, L. (2020a). *2'-F-ribo guanosine and LNA modified hybrid type G-quadruplex with V-loop.* Worldwide Protein Data Bank.

Weisz, K., and Haase, L. (2020b). *LNA modified G-quadruplex with flipped G-tract and central tetrad*. Worldwide Protein Data Bank.

Williamson, M. P., Wilson, T., Thomas, J. A., Felix, V., and Costa, P. J. (2013). *Structural studies on dinuclear ruthenium(II) complexes that bind diastereoselectivity to an anti-parallel folded human telomere sequence*. Worldwide Protein Data Bank.

Winnerdy, F. R., Bakalar, B., Maity, A., Vandana, J. J., Mechulam, Y., Schmitt, E., and Phan, A. T. (2019). *NMR solution and X-ray crystal structures of a DNA containing both right-and left-handed parallel-stranded G-quadruplexes.* Worldwide Protein Data Bank.

Winnerdy, F. R., Heddi, B., and Phan, A. T. (2020). *G-quadruplex complex with cyclic dinucleotide 3'-3' cGAMP*. Worldwide Protein Data Bank.

Winnerdy, F. R., Truong, T. H. A., and Phan, A. T. (2019). *G-quadruplex peripheral knot.* Worldwide Protein Data Bank.

Wirmer-Bartoschek, J., Jonker, H. R. A., Bendel, L. E., Gruen, T., Bazzicalupi, C., Messori, L., Gratteri, P., and Schwalbe, H. (2017). *Solution structure of a human G-Quadruplex hybrid-2 form in complex with a Gold-ligand*. Worldwide Protein Data Bank.

Zhang, Z., Dai, J., and Yang, D. (2009). *Human telomere DNA two-tetrad quadruplex structure in K+ solution.* Worldwide Protein Data Bank.

Zheng, G., Lu, X.-J., and Olson, W. K. (2009). *Web 3DNA--a web server for the analysis, reconstruction, and visualisation of three-dimensional nucleic-acid structures.* In Nucleic Acids Research (Vol. 37, Issue Web Server, pp. W240–W246).

Zok, T., Antczak, M., Zurkowski, M., Popenda, M., Blazewicz, J., Adamiak, R. W., and Szachniuk, M. (2018). *RNApdbee 2.0: multifunctional tool for RNA structure annotation. In Nucleic Acids Research* (Vol. 46, Issue W1, pp. W30–W35).

[Nature]. (n.d.) Discovery of DNA Double Helix: Watson and Crick. [Web-based visual]. Available at:. https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/.

Carvalho J, Mergny JL, Salgado GF, Queiroz JA, Cruz C. (2020). *G-Quadruplex, Friend or Foe: The Role of the G-Quartet in Anticancer Strategies*. Trends in Molecular Medicine Vol. 26 (9): pp. 848–61.

Madzharova, Fani, Zsuzsanna Heiner, Marina Gühlke, and Janina Kneipp. 2016. *“Surface-Enhanced Hyper-Raman Spectra of Adenine, Guanine, Cytosine, Thymine, and Uracil*.” The Journal of Physical Chemistry. C, Nanomaterials and Interfaces, Vol. 120 (28): pp. 15415–23.

[Wikipedia]. (2010, April 6). Nucleic Acid Structure. [Web-based visual]. Available at. https://en.wikipedia.org/wiki/Nucleic_acid_structure.

[Diffen]. (n.d.). Purines vs Pyrimidines. [Web-based visual]. Available at: https://www.diffen.com/difference/Purines_vs_Pyrimidines#:~:text=Purines%20and%20Pyrimidines%20are%20nitrogenous,thymine%20and%20cytosine)%20are%20pyrimidines.

[Nature]. (n.d.). Discovery of DNA Double Helix: Watson and Crick. [Web-based visual]. Available at. https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/.

APPENDIX

Table 1: Validation set

Pdb Name	Tetrad	Loops
6h1k	[1,20,16,27], [2,19,15,26], [25,28,17,21]	["1,2,d,15,16,17,r,19,20,21,l,25,r,26,27,28"]
2m91	[2,6,26,29], [1,25,30,7]	["1,2,d,15,16,17,r,19,20,21,l,25,r,26,27,28"]
2f8u	[1,9,17,21], [2,8,18,22], [3,7,19,23]	["1,2,3,l,7,8,9,l,17,18,19,r,21,22,23"]
7cv3	[3,7,22,25], [4,8,21,26], [5,9,20,27]	["3,4,5,r,7,8,9,l,20,21,22,l,25,26,27"]
7cv4	[2,25,18,7], [6,3,24,19]	["2,3,l,6,7,l,18,19,l,24,25"]
5zev	[1,15,10,22], [2,13,9,21], [20,16,11,23]	["1,2,d,9,10,11,r,13,15,16,l,20,r,21,22,23"]
6kvb	[2,12,9,6], [15,11,5,8], [17,3,23,20],[18,27,24,21]	["2,3,r,5,6,r,8,9,r,11,12,r,15,17,18,r,20,21,r,23,24,l,27"]
2lpw	[5,9,21,25], [4,8,20,24], [3,7,19,23]	["3,4,5,r,7,8,9,r,19,20,21,r,23,24,25"]
2n60	[3,7,15,11], [5,8,16,12]	["3,5,r,7,8,r,11,12,r,15,16"]
1hao	[401, 406, 410, 415], [402, 405, 411, 414]	["401,402,l,405,406,l,410,411,l,414,415"]
1hap	[401, 406, 410, 415], [402, 405, 411, 414]	["401,402,l,405,406,l,410,411,l,414,415"]

1hut	[401, 406, 410, 415], [402, 405, 411, 414]	["401,402,l,405,406,l,410,411,l,414,415"]
143d	[2,10,22,14], [3,9,21,15], [4,8,20,16]	["2,3,4,l,8,9,10,d,14,15,16,l,20,21,22"]
148d	[1,6,10,15], [2,5,11,14]	["1,2,l,5,6,l,10,11,l,14,15"]
186d	[3,12,16,21], [4,11,17,22], [5,10,18,23]	["3,4,5,l,10,11,12,l,16,17,18,r,21,22,23"]
1bub	[1,6,10,15], [2,5,11,14]	["1,2,l,5,6,l,10,11,l,14,15"]
1c32	[1,6,10,15], [2,5,11,14]	["1,2,l,5,6,l,10,11,l,14,15"]
1c34	[1,6,10,15], [2,5,11,14]	["1,2,l,5,6,l,10,11,l,14,15"]
1c35	[1,6,10,15], [2,5,11,14]	["1,2,l,5,6,l,10,11,l,14,15"]
1c38	[1,6,10,15], [2,5,11,14]	["1,2,l,5,6,l,10,11,l,14,15"]
1i34	[1,12,8,18], [2,11,7,19]	["1,2,d,7,8,r,11,12,d,18,19"]
1k8p	[3,9,15,21], [4,10,16,22], [5,11,17,23]	["3,4,5,r,9,10,11,r,15,16,17,r,21,22,23"]
1kf1	[2,8,14,20], [3,9,15,21], [4,10,16,22]	["2,3,4,r,8,9,10,r,14,15,16,r,20,21,22"]
1oz8	[2,5,8,11], [1,4,7,10], [13,16,19,22], [14,17,20,23]	["1,2,r,4,5,r,7,8,r,10,11,r,13,14,r,16,17,r,19,20,r,22,23"]
1qdf	[1,6,10,15], [2,5,11,14]	["1,2,l,5,6,l,10,11,l,14,15"]
1qdh	[1,6,10,15], [2,5,11,14]	["1,2,l,5,6,l,10,11,l,14,15"]
1rde	[1,6,10,15], [2,5,11,14]	["1,2,l,5,6,l,10,11,l,14,15"]

1xav	[4,8,13,17], [5,9,14,18], [6,10,15,19]	["4,5,6,r,8,9,10,r,13,14,15,r,17,18,19"]
201d	[1,12,28,17], [2,11,27,18], [3,10,26,19], [4,9,25,20]	["1,2,3,4,l,9,10,11,12,d,17,18,19,20,l,25,26,27,28"]
2a5p	[4,8,13,17], [5,9,14,18], [6,24,15,19]	["4,5,6,r,8,9,r,13,14,15,r,17,18,19,d,24"]
2e4i	[2,8,16,20], [3,9,15,21], [4,10,14,22]	["2,3,4,r,8,9,10,l,14,15,16,l,20,21,22"]
2kpr	[1,6,17,13], [2,5,16,12], [7,11,14,18]	["1,2,l,5,6,7,l,11,r,12,13,14,r,16,17,18"]
2m53	[4,11,14,19], [5,7,15,20], [6,8,16,21]	["4,5,6,r,7,8,r,11,l,14,15,16,r,19,20,21"]
2m90	[1,17,30,24], [18,21,25,31], [19,22,26,32]	["1,l,17,18,19,r,21,22,r,24,25,26,r,30,31,32"]
2m92	[2, 6, 10, 13], [3, 7, 14, 33], [4, 8, 15, 34]	["2,3,4,r,6,7,8,r,10,l,13,14,15,r,33,34"]
2m6v	[1, 8, 13, 18], [2, 7, 14, 17]	["1,2,l,7,8,d,13,14,l,17,18"]
2mft	[1, 10, 14, 19], [2, 9, 13, 20], [3, 8, 12, 21]	["1,2,3,d,8,9,10,r,12,13,14,d,19,20,21"]
2n6c	[4, 8, 13, 17], [5, 9, 14, 18], [6, 15, 19, 24]	["4,5,6,r,8,9,r,13,14,15,r,17,18,19,d,24"]
5j6u	[1, 11, 16, 25], [2, 10, 17, 24], [3, 9, 18, 23], [4, 8, 19, 22]	["1,2,3,4,l,8,9,10,11,d,16,17,18,19,l,22,23,24,25"]
5mtg	[3, 10, 25, 30], [4, 11, 24, 31], [5, 12, 23, 32]	["3,4,5,r,10,11,12,d,23,24,25,l,30,31,32"]
5ov2	[3, 8, 14, 22], [2, 7, 15, 21], [1, 6, 16,	["1,2,3,r,6,7,8,d,14,15,16,l,2

	20]	0,21,22"]
6suu	[2, 6, 11, 25], [3, 7, 12, 26], [4, 13, 27, 32]	["2,3,4,r,6,7,r,11,12,13,r,25,26,27,d,32"]
6zte	[18, 22, 25, 29], [19, 23, 26, 30], [20, 27, 31, 36]	["18,19,20,r,22,23,r,25,26,27,r,29,30,31,d,36"]
2mgn	[4, 8, 13, 17], [5, 9, 14, 18], [6, 15, 19, 24]	["4,5,6,r,8,9,r,13,14,15,r,17,18,19,d,24"]
6e84	[12, 15, 22, 25], [13, 16, 23, 26]	["12,13,r,15,16,r,22,23,r,25,26"]
6t2g	[2, 6, 11, 26], [3, 7, 12, 27], [4, 9, 13, 28]	["2,3,4,r,6,7,9,r,11,12,13,r,26,27,28"]
5ob3	[15, 19, 49, 53], [16, 20, 46, 51]	["15,16,r,19,20,d,46,49,r,51,53"]

Pdb Files Classified According to the ONZ Classification

DNAs

NNN 13 ['pdb143d', 'pdb2lod', 'pdb2mcc', 'pdb2mco', 'pdb5j05', 'pdb5mcr', 'pdb5mta', 'pdb5mtg', 'pdb5ov2', 'pdb6f4z', 'pdb6ffr', 'pdb6l92', 'pdb6yep']

(Dickerhoff et al., 2017b; Dickerhoff and Weisz, 2017, 2018; Dvorkin et al., 2017a; Haase et al., 2018a; Juribasic Kulcsar and Plavec, 2018a, 2018b; Marusic et al., 2012; Y. Wang and Patel, 1994a; Z. F. Wang et al., 2019a; Weisz and Haase, 2020b; Williamson et al., 2013)

OO 22 ['pdb148d', 'pdb1bub', 'pdb1c32', 'pdb1c34', 'pdb1c35', 'pdb1c38', 'pdb1qdf', 'pdb1qdh', 'pdb1rde', 'pdb2idn', 'pdb2km3', 'pdb2lyg', 'pdb2m8z', 'pdb2mfu', 'pdb2mwz', 'pdb2n60', 'pdb5mjx', 'pdb6fc9', 'pdb6gh0', 'pdb6k3y', 'pdb6v0l', 'pdb7cv4']

OO 29 ['pdb148d', 'pdb1bub', 'pdb1c32', 'pdb1c34', 'pdb1c35', 'pdb1c38', 'pdb1hao', 'pdb1hap', 'pdb1hut', 'pdb1qdf', 'pdb1qdh', 'pdb1rde', 'pdb2idn', 'pdb2km3', 'pdb2lyg', 'pdb2m8z', 'pdb2mfu', 'pdb2mwz', 'pdb2n60', 'pdb4wb2', 'pdb4wb3', 'pdb5mjx', 'pdb6eo6', 'pdb6eo7', 'pdb6fc9', 'pdb6gh0', 'pdb6k3y', 'pdb6v0l', 'pdb7cv4']

(Bolton et al., 1999; Cheong et al., 2015; Gomez-Pinto et al., 2014; Heddi et al., 2016; A. Karsisiotis et al., 2014; Kotar et al., 2019; Lietard et al., 2017; Lim, Alberti, et al., 2009; Lim and Phan, 2013d; Mao et al., 2003; Marathias et al., 1996, 1998; Randazzo et al., 2007; Santana et al., 2019; Schultze et al., 1994; 2019a, 2020a; Winnerdy et al., 2020)

OOO 79 ['pdb186d', 'pdb1k8p', 'pdb1kf1', 'pdb1xav', 'pdb2e4i', 'pdb2f8u', 'pdb2gku', 'pdb2hy9', 'pdb2jpz', 'pdb2jsk', 'pdb2jsl', 'pdb2jsm', 'pdb2jsq', 'pdb2kqg', 'pdb2kqh', 'pdb2kyp', 'pdb2kzd', 'pdb2kze', 'pdb2l7v', 'pdb2l88', 'pdb2lby', 'pdb2ld8', 'pdb2lee', 'pdb2lk7', 'pdb2lpw', 'pdb2lxq', 'pdb2m27', 'pdb2m4p', 'pdb2m53', 'pdb2m93', 'pdb2may', 'pdb2mb2', 'pdb2mb3', 'pdb2mbj', 'pdb2n4y', 'pdb3r6r', 'pdb3sc8', 'pdb3t5e', 'pdb3uyh', 'pdb4da3', 'pdb4daq', 'pdb4fxm', 'pdb4g0f', 'pdb5ccw', 'pdb5i2v', 'pdb5lig', 'pdb5mbr', 'pdb5mvb', 'pdb5nys', 'pdb5nyt', 'pdb5nyu', 'pdb5w77', 'pdb5yey', 'pdb5z80', 'pdb6ac7', 'pdb6ccw', 'pdb6erl', 'pdb6h5r', 'pdb6ia0', 'pdb6ia4', 'pdb6ip3', 'pdb6ip7', 'pdb6isw', 'pdb6jj0', 'pdb6jkn', 'pdb6jwd', 'pdb6jwe', 'pdb6kfi', 'pdb6kfj', 'pdb6neb', 'pdb6o2l', 'pdb6r9k', 'pdb6r9l', 'pdb6t2g', 'pdb6t51', 'pdb6zl2', 'pdb6zl9', 'pdb7cls', 'pdb7cv3']

(Agrawal et al., 2013; Ambrus et al., 2005; Amrane et al., 2012; Bazzicalupi et al., 2019; Bielskute et al., 2019a, 2019b; Chen and Walters, 2018; Chung et al., 2013; Collie et al., 2012a, 2012b; Collie and Neidle, 2013a, 2013b, 2013c; Dai, Carver, et al., 2007; Dai et al., 2006, 2011; Dai, Punchihewa, et al., 2007; DeNicola et al., 2016; Dickerhoff et al., 2017a, 2019; Do and Phan, 2012; Ferraroni et al., 2012, 2016; Geng et al., 2019; Heddi and Phan, 2011; Hsu et al., 2009; Karg and Weisz, 2018, 2019a, 2019b; Kotar et al., 2016; V. Kuryavyi et al., 2010; V. V. Kuryavyi et al., 2008; V. V. Kuryavyi and Patel, 2012; Lech et al., 2015; Lim et al., 2010, 2013; Lim and Phan, 2013e; Lin et al., 2019; Lin and Yang, 2018; C. Liu et al., 2018; W. Liu et al., 2019; W. T. Liu et al., 2018; Luu et al., 2006; Marquevielle and Salgado, 2019; Mathad et al., 2011; Matsugami et al., 2007; Mukundan and Phan, 2013; Nguyen et al., 2020;

Nicoludis et al., 2012; Parkinson et al., 2002a, 2002b; Saikrishnan et al., 2019a, 2019b, 2019c; Salgado et al., 2016; Sengar et al., 2014, 2019; Tong and Cao, 2011; Trajkovski et al., 2012, 2018a, 2018b, 2018c; 2019b, 2019c, 2019d, 2020b, 2020c, 2020d; Y. Wang and Patel, 1994b; Wirmer-Bartoschek et al., 2017)

ZZ 1 ['pdb1i34']

(V. Kuryavyi et al., 2001)

OOOO 9 ['pdb1oz8', 'pdb2ms9', 'pdb2n2d', 'pdb2n3m', 'pdb4u5m', 'pdb5oph', 'pdb6gz6', 'pdb6jce', 'pdb6kvb']

(Bakalar et al., 2019; Brcic and Plavec, 2015, 2018; Chung et al., 2015; Do et al., 2016; Maity et al., 2020; Matsugami et al., 2004; Schmitt et al., 2015; Winnerdy, Bakalar, et al., 2019)

NNNN 3 ['pdb201d', 'pdb2m6w', 'pdb5j6u']

(Dvorkin et al., 2017b; A. I. Karsisiotis and Webba da Silva, 2014b; Y. Wang and Patel, 1995)

OOZ 6 ['pdb2a5p', 'pdb2a5r', 'pdb2mgn', 'pdb2n6c', 'pdb6zte']

(Chung et al., 2014; Kumar and Tawani, 2016; Phan et al., 2005a, 2005b; 2019e, 2020e)

NN 11 ['pdb2kf7', 'pdb2kf8', 'pdb2kka', 'pdb2kow', 'pdb2m6v', 'pdb2m91', 'pdb5j4p', 'pdb5j4w', 'pdb5lqg', 'pdb5lqh', 'pdb6gzn']

(Dvorkin and Webba da Silva, 2017a, 2017b; Galer et al., 2018a, 2018b; Hu et al., 2009; A. I. Karsisiotis and Webba da Silva, 2014a; Lenarcic Zivkovic et al., 2018; Lim, Amrane, et al., 2009a, 2009b; Lim and Phan, 2013c; Zhang et al., 2009)

NNO 7 ['pdb2kpr', 'pdb5o4d', 'pdb6l8m', 'pdb6rs3', 'pdb6tc8', 'pdb6tcg', 'pdb6ycv']

(Haase and Weisz, 2019, 2020a, 2020b; V. Kuryavyi and Patel, 2010; Marusic and Plavec, 2018; Z. F. Wang et al., 2019b; Weisz and Haase, 2020a)

NOO 1 ['pdb2m90']

(Lim and Phan, 2013b)

ONN 2 ['pdb2m92', 'pdb2o3m']
(Lim and Phan, 2013a; Phan et al., 2007)

ZZZ 1 ['pdb2mft']
(A. I. Karsisiotis and Webba da Silva, 2014c)

ZZO 2 ['pdb5zev', 'pdb6h1k']
(Y. Liu and Lan, 2018; Phan et al., 2018)

ZO 1 ['pdb6jcd']
(Winnerdy, Truong, et al., 2019)

ZOO 1 ['pdb6suu']
(Marquevielle et al. 2020)

RNAs

[pdb1s9l, pdb2la5, pdb3ibk, pdb4kzd, pdb4kze, pdb4q9q, pdb4q9r, pdb4wb2, pdb4wb3, pdb5ob3, pdb6e80, pdb6e81, pdb6e84, pdb6e8u, pdb6ffr, pdb6k84, pdb6pq7]

(Collie et al., 2010; Fernandez-Millan et al., 2017; Haase et al., 2018b; Mashima et al., 2020; Randazzo et al., 2004; Sjekloća and Ferré-D’Amaré, 2019; Trachman, Autour, et al., 2019; Trachman, Stagno, et al., 2019)

ONZ Classes and Loops

NNN 12 ['pdb143d.ent', 'pdb2lod.ent', 'pdb2mcc.ent', 'pdb5j05.ent', 'pdb5mcr.ent', 'pdb5mta.ent', 'pdb5mtg.ent', 'pdb5ov2.ent', 'pdb6f4z.ent', 'pdb6ffr__.ent', 'pdb6l92.ent', 'pdb6yep.ent']

pdb143d.ent ['2', '3', '4', 'lateral', '8', '9', '10', 'diagonal', '14', '15', '16', 'lateral', '20', '21', '22']

pdb2lod.ent ['1', '2', '3', 'reversal', '6', '7', '8', 'diagonal', '14', '15', '16', 'lateral', '20', '21', '22']

pdb2mcc.ent ['2', '3', '4', 'lateral', '8', '9', '10', 'diagonal', '14', '15', '16', 'lateral', '20', '21', '22']

pdb5j05.ent ['1', '2', '3', 'lateral', '7', '8', '9', 'diagonal', '14', '15', '16', 'lateral', '18', '19', '20']

pdb5mcr.ent ['1', '2', '3', 'reversal', '6', '7', '8', 'diagonal', '14', '15', '16', 'lateral', '20', '21', '22']

pdb5mta.ent ['3', '4', '5', 'reversal', '10', '11', '12', 'diagonal', '23', '24', '25', 'lateral', '30', '31', '32']

pdb5mtg.ent ['3', '4', '5', 'reversal', '10', '11', '12', 'diagonal', '23', '24', '25', 'lateral', '30', '31', '32']

pdb5ov2.ent ['1', '2', '3', 'reversal', '6', '7', '8', 'diagonal', '14', '15', '16', 'lateral', '20', '21', '22']

pdb6f4z.ent ['1', '2', '3', 'lateral', '6', '7', '8', 'diagonal', '14', '15', '16', 'lateral', '20', '21', '22']

pdb6ffr__.ent ['1', '2', '3', 'reversal', '6', '7', '8', 'diagonal', '14', '15', '16', 'lateral', '20', '21', '22']

pdb6l92.ent ['1', '2', '3', 'lateral', '9', '10', '11', 'diagonal', '16', '17', '18', 'reversal', '20', '21', '22']

pdb6yep.ent ['1', '2', '3', 'lateral', '6', '7', '8', 'diagonal', '14', '15', '16', 'lateral', '20', '21', '22']

OO 27 ['pdb148d.ent', 'pdb1bub.ent', 'pdb1c32.ent', 'pdb1c34.ent', 'pdb1c35.ent', 'pdb1c38.ent', 'pdb1qdf.ent', 'pdb1qdh.ent', 'pdb1rde.ent', 'pdb2idn.ent', 'pdb2km3.ent', 'pdb2lyg.ent', 'pdb2m8z.ent', 'pdb2mfu.ent', 'pdb2mwz.ent', 'pdb2n60.ent', 'pdb4wb2.ent', 'pdb4wb3.ent', 'pdb5mjx.ent', 'pdb6eo6.ent', 'pdb6eo7.ent', 'pdb6fc9.ent', 'pdb6gh0.ent', 'pdb6ia0.ent', 'pdb6k3y.ent', 'pdb6v0l.ent', 'pdb7cv4.ent']

pdb148d.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb1bub.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb1c32.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15'] pdb1c34.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb1c35.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb1c38.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb1qdf.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb1qdh.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb1rde.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb2idn.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb2km3.ent ['3', '4', 'lateral', '9', '10', 'lateral', '15', '16', 'lateral', '21', '22']

pdb2lyg.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb2m8z.ent ['1', '2', 'lateral', '5', '6', 'lateral', '22', '23', 'lateral', '26', '27']

pdb2mfu.ent ['3', '4', 'lateral', '9', '10', 'lateral', '14', '15', 'reversal', '19', '20']

pdb2mwz.ent ['4', '5', 'reversal', '10', '11', 'lateral', '15', '16', 'lateral', '22', '23']

pdb2n60.ent ['3', '5', 'reversal', '7', '8', 'reversal', '11', '12', 'reversal', '15', '16']

pdb4wb2.ent ['17', 'reversal', '18', '19', 'reversal', '22', '25', 'reversal', '26', '27', 'reversal', '32']

pdb4wb3.ent ['17', 'reversal', '18', '19', 'reversal', '22', '25', 'reversal', '26', '27', 'reversal', '32']

pdb5mjx.ent ['1', '2', 'lateral', '5', '6', 'lateral', '10', '11', 'lateral', '14', '15']

pdb6eo6.ent ['401', '402', 'lateral', '405', '406', 'lateral', '410', '411', 'lateral', '414', '415']

pdb6eo7.ent ['401', '402', 'lateral', '405', '406', 'lateral', '410', '411', 'lateral', '414', '415']

pdb6fc9.ent ['1', '2', 'lateral', '5', '6', 'lateral', '22', '23', 'lateral', '26', '27']

pdb6gh0.ent ['1', '2', 'lateral', '6', '7', 'lateral', '11', '12', 'lateral', '16', '17']

pdb6ia0.ent ['3', '5', 'lateral', '9', '11', 'lateral', '15', 'reversal', '17', 'reversal', '21', '23']

pdb6k3y.ent ['3', '4', 'reversal', '6', '7', 'reversal', '10', '11', 'reversal', '14', '15']

pdb6v0l.ent ['4', '5', 'reversal', '8', '9', 'reversal', '11', '12', 'reversal', '15', '16']

pdb7cv4.ent ['2', '3', 'lateral', '6', '7', 'lateral', '18', '19', 'lateral', '24', '25']

OOO 77 ['pdb186d.ent', 'pdb1k8p.ent', 'pdb1kf1.ent', 'pdb1xav.ent', 'pdb2e4i.ent', 'pdb2f8u.ent', 'pdb2gku.ent', 'pdb2hy9.ent', 'pdb2jpz.ent', 'pdb2jsk.ent', 'pdb2jsl.ent', 'pdb2jsm.ent', 'pdb2jsq.ent', 'pdb2kqg.ent', 'pdb2kqh.ent', 'pdb2kyp.ent', 'pdb2kzd.ent', 'pdb2kze.ent', 'pdb2l7v.ent', 'pdb2l88.ent', 'pdb2lby.ent', 'pdb2ld8.ent', 'pdb2lee.ent', 'pdb2lk7.ent', 'pdb2lpw.ent', 'pdb2lxq.ent', 'pdb2m27.ent', 'pdb2m4p.ent', 'pdb2m53.ent', 'pdb2m93.ent', 'pdb2may.ent', 'pdb2mb2.ent', 'pdb2mb3.ent', 'pdb2mbj.ent', 'pdb2n4y.ent', 'pdb3r6r.ent', 'pdb3sc8.ent', 'pdb3t5e.ent', 'pdb3uyh.ent', 'pdb4da3.ent', 'pdb4daq.ent', 'pdb4fxm.ent', 'pdb4g0f.ent', 'pdb5ccw.ent', 'pdb5i2v.ent', 'pdb5lig.ent', 'pdb5mbr.ent', 'pdb5mvb.ent', 'pdb5nys.ent', 'pdb5nyt.ent', 'pdb5nyu.ent', 'pdb5w77.ent', 'pdb5yey.ent', 'pdb5z80.ent', 'pdb6ac7.ent', 'pdb6ccw.ent', 'pdb6erl.ent', 'pdb6h5r.ent', 'pdb6ia4.ent', 'pdb6ip3.ent', 'pdb6ip7.ent', 'pdb6isw.ent', 'pdb6jj0.ent', 'pdb6jkn.ent', 'pdb6jwd.ent', 'pdb6jwe.ent', 'pdb6kfi.ent', 'pdb6kfj.ent', 'pdb6neb.ent', 'pdb6o2l.ent', 'pdb6r9k.ent', 'pdb6r9l.ent', 'pdb6t51.ent', 'pdb6zl2.ent', 'pdb6zl9.ent', 'pdb7cls.ent', 'pdb7cv3.ent']

pdb186d.ent ['3', '4', '5', 'lateral', '10', '11', '12', 'lateral', '16', '17', '18', 'reversal', '21', '22', '23']

pdb1k8p.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'reversal', '15', '16', '17', 'reversal', '21', '22', '23']

pdb1kf1.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'reversal', '14', '15', '16', 'reversal', '20', '21', '22']

pdb1xav.ent ['4', '5', '6', 'reversal', '8', '9', '10', 'reversal', '13', '14', '15', 'reversal', '17', '18', '19']

pdb2e4i.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'lateral', '14', '15', '16', 'lateral', '20', '21', '22']

pdb2f8u.ent ['1', '2', '3', 'lateral', '7', '8', '9', 'lateral', '17', '18', '19', 'reversal', '21', '22', '23']

pdb2gku.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'lateral', '15', '16', '17', 'lateral', '21', '22', '23']

pdb2hy9.ent ['4', '5', '6', 'reversal', '10', '11', '12', 'lateral', '16', '17', '18', 'lateral', '22', '23', '24']

pdb2jpz.ent ['4', '5', '6', 'lateral', '10', '11', '12', 'lateral', '16', '17', '18', 'reversal', '22', '23', '24']

pdb2jsk.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'lateral', '15', '16', '17', 'lateral', '21', '22', '23']

pdb2jsl.ent ['3', '4', '5', 'lateral', '9', '10', '11', 'lateral', '15', '16', '17', 'reversal', '21', '22', '23']

pdb2jsm.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'lateral', '15', '16', '17', 'lateral', '21', '22', '23']

pdb2jsq.ent ['3', '4', '5', 'lateral', '9', '10', '11', 'lateral', '15', '16', '17', 'reversal',

'21', '22', '23']

pdb2kqg.ent ['2', '3', '4', 'reversal', '6', '7', '8', 'reversal', '14', '15', '16', 'reversal', '18', '19', '20']

pdb2kqh.ent ['2', '3', '4', 'reversal', '6', '7', '8', 'reversal', '14', '15', '16', 'reversal', '18', '19', '20']

pdb2kyp.ent ['2', '3', '4', 'reversal', '6', '7', '8', 'reversal', '14', '15', '16', 'reversal', '18', '19', '20']

pdb2kzd.ent ['2', '3', '4', 'lateral', '8', '9', '10', 'lateral', '13', '14', '15', 'reversal', '17', '18', '19']

pdb2kze.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '13', '14', '15', 'reversal', '17', '18', '19']

pdb2l7v.ent ['7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '16', '17', '18', 'reversal', '20', '21', '22']

pdb2l88.ent ['1', '2', '3', 'reversal', '7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '17', '18', '19']

pdb2lby.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '12', '13', '14', 'reversal', '16', '17', '18']

pdb2ld8.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'reversal', '15', '16', '17', 'reversal', '21', '22', '23']

pdb2lee.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '15', '16', '17']

pdb2lk7.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '15', '16', '17']

pdb2lpw.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '19', '20', '21', 'reversal', '23', '24', '25']

pdb2lxq.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '12', '13', '14', 'reversal', '16', '17', '18']

pdb2m27.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '14', '15', '16', 'reversal', '18', '19', '20']

pdb2m4p.ent ['3', '5', '6', 'reversal', '8', '9', '10', 'reversal', '12', '13', '14', 'reversal', '16', '17', '18']

pdb2m53.ent ['4', '5', '6', 'reversal', '7', '8', 'reversal', '11', 'lateral', '14', '15', '16', 'reversal', '19', '20', '21']

pdb2m93.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '25', '26', '27', 'reversal',

'29', '30', '31']

pdb2may.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'lateral', '15', '16', '17', 'lateral', '21', '22', '23']

pdb2mb2.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '15', '16', '17']

pdb2mb3.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'lateral', '15', '16', '17', 'lateral', '21', '22', '23']

pdb2mbj.ent ['4', '5', '6', 'lateral', '10', '11', '12', 'reversal', '16', '17', '18', 'lateral', '22', '23', '24']

pdb2n4y.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '14', '15', '16', 'reversal', '18', '20', '21']

pdb3r6r.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'reversal', '15', '16', '17', 'reversal', '21', '22', '23']

pdb3sc8.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'reversal', '14', '15', '16', 'reversal', '20', '21', '22']

pdb3t5e.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'reversal', '14', '15', '16', 'reversal', '20', '21', '22']

pdb3uyh.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'reversal', '14', '15', '16', 'reversal', '20', '21', '22']

pdb4da3.ent ['1', '2', '3', 'reversal', '7', '8', '9', 'reversal', '13', '14', '15', 'reversal', '19', '20', '21']

pdb4daq.ent ['1', '2', '3', 'reversal', '7', '8', '9', 'reversal', '13', '14', '15', 'reversal', '19', '20', '21']

pdb4fxm.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'reversal', '14', '15', '16', 'reversal', '20', '21', '22']

pdb4g0f.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'reversal', '14', '15', '16', 'reversal', '20', '21', '22']

pdb5ccw.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'reversal', '15', '16', '17', 'reversal', '21', '22', '23']

pdb5i2v.ent ['2', '3', '4', 'reversal', '6', '7', '9', 'reversal', '11', '12', '13', 'reversal', '18', '19', '20']

pdb5lig.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '12', '13', '14', 'reversal', '16', '17', '18']

pdb5mbr.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'lateral', '15', '16', '17', 'lateral',

'21', '22', '23']

pdb5mvb.ent ['4', '5', '6', 'lateral', '10', '11', '12', 'lateral', '16', '17', '18', 'reversal', '22', '23', '24']

pdb5nys.ent ['3', '4', '5', 'reversal', '8', '9', '10', 'reversal', '12', '13', '14', 'reversal', '17', '18', '19']

pdb5nyt.ent ['3', '4', '5', 'reversal', '8', '9', '10', 'reversal', '12', '13', '14', 'reversal', '17', '18', '19']

pdb5nyu.ent ['3', '4', '5', 'reversal', '8', '9', '10', 'reversal', '12', '13', '14', 'reversal', '17', '18', '19']

pdb5w77.ent ['7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '16', '17', '18', 'reversal', '20', '21', '22']

pdb5yey.ent ['1', '2', '3', 'lateral', '7', '8', '9', 'lateral', '13', '14', '15', 'lateral', '19', '20', '21']

pdb5z80.ent ['4', '5', '6', 'reversal', '10', '11', '12', 'lateral', '16', '17', '18', 'lateral', '22', '23', '24']

pdb6ac7.ent ['3', '4', '5', 'lateral', '9', '11', '12', 'lateral', '15', '16', '17', 'reversal', '21', '22', '23']

pdb6ccw.ent ['4', '5', '6', 'lateral', '10', '11', '12', 'lateral', '16', '17', '18', 'reversal', '22', '23', '24']

pdb6erl.ent ['4', '5', '6', 'reversal', '8', '9', '10', 'reversal', '13', '14', '15', 'reversal', '17', '18', '19']

pdb6h5r.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'reversal', '15', '16', '17', 'reversal', '21', '22', '23']

pdb6ia4.ent ['3', '4', '5', 'reversal', '9', '10', '11', 'lateral', '15', '16', '17', 'lateral', '21', '22', '23']

pdb6ip3.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'reversal', '14', '15', '16', 'reversal', '20', '21', '22']

pdb6ip7.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'reversal', '14', '15', '16', 'reversal', '20', '21', '22']

pdb6isw.ent ['2', '3', '4', 'reversal', '8', '9', '10', 'reversal', '14', '15', '16', 'reversal', '20', '21', '22']

pdb6jj0.ent ['7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '16', '17', '18', 'reversal', '20', '21', '22']

pdb6jkn.ent ['1', '2', '3', 'lateral', '7', '8', '9', 'lateral', '13', '14', '15', 'lateral', '19', '20', '21']

pdb6jwd.ent ['1', '2', '3', 'reversal', '7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '17', '18', '19']

pdb6jwe.ent ['1', '2', '3', 'reversal', '7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '17', '18', '19']

pdb6kfi.ent ['4', '5', '6', 'reversal', '10', '11', '12', 'lateral', '16', '17', '18', 'lateral', '22', '23', '24']

pdb6kfj.ent ['4', '5', '6', 'lateral', '10', '11', '12', 'lateral', '16', '17', '18', 'reversal', '22', '23', '24']

pdb6neb.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'reversal', '16', '17', '18', 'reversal', '20', '21', '22']

pdb6o2l.ent ['7', '8', '9', 'reversal', '11', '12', '13', 'reversal', '16', '17', '18', 'reversal', '20', '21', '22']

pdb6r9k.ent ['5', '6', '7', 'lateral', '11', '12', '13', 'reversal', '15', '16', '17', 'reversal', '19', '20', '21']

pdb6r9l.ent ['5', '6', '7', 'lateral', '11', '12', '13', 'reversal', '15', '16', '17', 'reversal', '19', '20', '21']

pdb6t51.ent ['2', '3', '4', 'reversal', '6', '7', '9', 'reversal', '11', '12', '13', 'reversal', '18', '19', '20']

pdb6zl2.ent ['4', '5', '6', 'reversal', '8', '9', '10', 'reversal', '13', '14', '15', 'reversal', '17', '18', '19']

pdb6zl9.ent ['18', '19', '20', 'reversal', '22', '23', '24', 'reversal', '27', '28', '29', 'reversal', '31', '32', '33']

pdb7cls.ent ['3', '4', '20', 'reversal', '22', '23', '24', 'reversal', '26', '27', '28', 'reversal', '30', '31', '32']

pdb7cv3.ent ['3', '4', '5', 'reversal', '7', '8', '9', 'lateral', '20', '21', '22', 'lateral', '25', '26', '27']

ZZ 1 ['pdb1i34.ent']

pdb1i34.ent ['1', '2', 'diagonal', '7', '8', 'reversal', '11', '12', 'diagonal', '18', '19']

NNNN 3 ['pdb201d.ent', 'pdb2m6w.ent', 'pdb5j6u.ent']

pdb201d.ent ['1', '2', '3', '4', 'lateral', '9', '10', '11', '12', 'diagonal', '17', '18', '19', '20', 'lateral', '25', '26', '27', '28']

pdb2m6w.ent ['1', '2', '3', '4', 'lateral', '7', '8', '9', '10', 'diagonal', '15', '16', '17', '18', 'lateral', '21', '22', '23', '24'] pdb5j6u.ent ['1', '2', '3', '4', 'lateral', '8', '9', '10', '11', 'diagonal', '16', '17', '18', '19', 'lateral', '22', '23', '24', '25']

OOZ 5 ['pdb2a5p.ent', 'pdb2a5r.ent', 'pdb2mgn.ent', 'pdb2n6c.ent', 'pdb6zte.ent']

pdb2a5p.ent ['4', '5', '6', 'reversal', '8', '9', 'reversal', '13', '14', '15', 'reversal', '17', '18', '19', 'diagonal', '24']

pdb2a5r.ent ['4', '5', '6', 'reversal', '8', '9', 'reversal', '13', '14', '15', 'reversal', '17', '18', '19', 'diagonal', '24']

pdb2mgn.ent ['4', '5', '6', 'reversal', '8', '9', 'reversal', '13', '14', '15', 'reversal', '17', '18', '19', 'diagonal', '24']

pdb2n6c.ent ['4', '5', '6', 'reversal', '8', '9', 'reversal', '13', '14', '15', 'reversal', '17', '18', '19', 'diagonal', '24']

pdb6zte.ent ['18', '19', '20', 'reversal', '22', '23', 'reversal', '25', '26', '27', 'reversal', '29', '30', '31', 'diagonal', '36']

NN 12 ['pdb2kf7.ent', 'pdb2kf8.ent', 'pdb2kka.ent', 'pdb2kow.ent', 'pdb2m6v.ent', 'pdb2m91.ent', 'pdb2mco.ent', 'pdb5j4p.ent', 'pdb5j4w.ent', 'pdb5lqg.ent', 'pdb5lqh.ent', 'pdb6gzn.ent']

pdb2kf7.ent ['1', '2', 'lateral', '7', '8', 'diagonal', '14', '15', 'lateral', '19', '20']

pdb2kf8.ent ['1', '2', 'lateral', '7', '8', 'diagonal', '14', '15', 'lateral', '19', '20']

pdb2kka.ent ['2', '3', 'lateral', '8', '9', 'diagonal', '15', '16', 'lateral', '20', '21']

pdb2kow.ent ['3', '4', 'lateral', '8', '9', 'diagonal', '14', '15', 'lateral', '19', '20']

pdb2m6v.ent ['1', '2', 'lateral', '7', '8', 'diagonal', '13', '14', 'lateral', '17', '18']

pdb2m91.ent ['1', '2', 'lateral', '6', '7', 'diagonal', '25', '26', 'lateral', '29', '30']

pdb2mco.ent ['2', '4', 'lateral', '8', '10', 'diagonal', '14', '16', 'lateral', '20', '22']

pdb5j4p.ent ['1', '2', 'lateral', '6', '7', 'diagonal', '12', '13', 'lateral', '17', '18']

pdb5j4w.ent ['1', '2', 'lateral', '6', '7', 'diagonal', '12', '13', 'lateral', '16', '17']

pdb5lqg.ent ['3', '4', 'lateral', '9', '10', 'diagonal', '16', '17', 'lateral', '21', '22']

pdb5lqh.ent ['3', '4', 'lateral', '9', '10', 'diagonal', '16', '17', 'lateral', '21', '22']

pdb6gzn.ent ['1', '2', 'lateral', '6', '7', 'diagonal', '13', '14', 'lateral', '18', '19']

NNO 7 ['pdb2kpr.ent', 'pdb5o4d.ent', 'pdb6l8m.ent', 'pdb6rs3.ent', 'pdb6tc8.ent', 'pdb6tcg.ent', 'pdb6ycv.ent']

pdb2kpr.ent ['1', '2', 'lateral', '5', '6', '7', 'lateral', '11', 'reversal', '12', '13', '14', 'reversal', '16', '17', '18']

pdb5o4d.ent ['1', '2', 'lateral', '6', '7', '8', 'lateral', '11', 'reversal', '12', '13', '14',

```
'reversal', '20', '21', '22']
        pdb6l8m.ent ['1', '2', 'lateral', '9', '10', '11', 'lateral', '14', 'reversal', '16', '17', '18',
'reversal', '20', '21', '22']
        pdb6rs3.ent ['1', '2', 'lateral', '6', '7', '8', 'lateral', '14', 'reversal', '15', '16', '17',
'reversal', '20', '21', '22']
        pdb6tc8.ent ['1', '2', 'lateral', '6', '7', '8', 'lateral', '14', 'reversal', '15', '16', '17',
'reversal', '20', '21', '22']
        pdb6tcg.ent ['1', '2', 'lateral', '6', '7', '8', 'lateral', '14', 'reversal', '15', '16', '17',
'reversal', '20', '21', '22']
        pdb6ycv.ent ['1', '2', 'lateral', '6', '7', '8', 'lateral', '14', 'reversal', '15', '16', '17',
'reversal', '20', '21', '22']
        NOO 1 ['pdb2m90.ent']
        pdb2m90.ent ['1', 'lateral', '17', '18', '19', 'reversal', '21', '22', 'reversal', '24', '25',
'26', 'reversal', '30', '31', '32']
        ONN 2 ['pdb2m92.ent', 'pdb2o3m.ent']
        pdb2m92.ent ['2', '3', '4', 'reversal', '6', '7', '8', 'reversal', '10', 'lateral', '13', '14',
'15', 'reversal', '33', '34']
        pdb2o3m.ent ['2', '3', '4', 'reversal', '6', '7', '8', 'reversal', '10', 'lateral', '13', '14',
'15', 'reversal', '21', '22']
        ZZZ 1 ['pdb2mft.ent']
        pdb2mft.ent ['1', '2', '3', 'diagonal', '8', '9', '10', 'reversal', '12', '13', '14',
'diagonal', '19', '20', '21']
        OOOO 8 ['pdb2ms9.ent', 'pdb2n2d.ent', 'pdb2n3m.ent', 'pdb4u5m.ent',
'pdb5oph.ent', 'pdb6gz6.ent', 'pdb6jce.ent', 'pdb6kvb.ent']
        pdb2ms9.ent ['2', '3', 'reversal', '5', '6', 'reversal', '8', '9', 'reversal', '11', '12', '15',
'reversal', '17', '18', 'reversal', '20', '21', 'reversal', '23', '24', 'reversal', '26']
        pdb2n2d.ent ['1', '2', '3', '4', 'lateral', '7', '8', '9', '10', 'lateral', '13', '14', '15', '16',
'lateral', '19', '20', '21', '22']
        pdb2n3m.ent ['2', '3', 'reversal', '5', '6', 'reversal', '8', '9', 'reversal', '12', '15', '17',
'18', 'reversal', '20', '21', 'reversal', '23', '24', 'reversal', '26', '27']
        pdb4u5m.ent ['2', '3', 'reversal', '5', '6', 'reversal', '8', '9', 'reversal', '11', '12', '15',
'reversal', '17', '18', 'reversal', '20', '21', 'reversal', '23', '24', 'reversal', '26']
        pdb5oph.ent ['1', '2', '3', '4', 'lateral', '7', '8', '9', '10', 'lateral', '13', '14', '15', '16',
'lateral', '19', '20', '21', '22']
```

pdb6gz6.ent ['1', 'reversal', '3', '4', 'reversal', '6', '7', 'reversal', '9', '10', 'reversal', '12', '15', 'reversal', '17', '18', 'reversal', '20', '21', 'reversal', '23', '24', 'reversal', '26']

pdb6jce.ent ['1', '2', 'reversal', '5', '6', 'reversal', '10', '11', 'reversal', '14', '15', 'reversal', '18', 'reversal', '20', '21', 'reversal', '23', '24', 'reversal', '26', '27', 'reversal', '29']

pdb6kvb.ent ['2', 'reversal', '3', '5', '6', 'reversal', '8', '9', 'reversal', '11', '12', 'reversal', '15', '17', '18', 'reversal', '20', '21', 'reversal', '23', '24', 'lateral', '27']

ZZO 2 ['pdb5zev.ent', 'pdb6h1k.ent']

pdb5zev.ent ['1', '2', 'diagonal', '9', '10', '11', 'reversal', '13', '15', '16', 'lateral', '20', 'reversal', '21', '22', '23']

pdb6h1k.ent ['1', '2', 'diagonal', '15', '16', '17', 'reversal', '19', '20', '21', 'lateral', '25', 'reversal', '26', '27', '28']

ZO 1 ['pdb6jcd.ent']

pdb6jcd.ent ['2', '3', 'lateral', '6', 'reversal', '8', '9', 'lateral', '13', '14', 'diagonal', '19']

ZOO 1 ['pdb6suu.ent']

pdb6suu.ent ['2', '3', '4', 'reversal', '6', '7', 'reversal', '11', '12', '13', 'reversal', '25', '26', '27', 'diagonal', '32']