chemistry proceedings

MDPI

Proceeding Paper

Advanced Computational Frameworks for Characterizing Abnormal DNA Architectures and Their Implications in Genome Dynamics †

Sameen Masroor [1], Chhavi Dudeja [1], Richa Sanka [2], Yukti Sabikhi [1], Anshika Singh [1], Amish Mishra [1] and Richa Gupta [1,*]

[1] School of Engineering Science and Technology, Jamia Hamdard University, New Delhi 110062, India; masroorsameen@gmail.com (S.M.); chhavidudeja2002@gmail.com (C.D.); yukti257@gmail.com (Y.S.); anshikasingh090304@gmail.com (A.S.); iamamishmishra111121@gmail.com (A.M.)
[2] College of Innovation and Technology, University of Michigan Flint, Flint, MI 48502, USA; sankariricha@gmail.com
* Correspondence: richagupta@jamiahamdard.ac.in; Tel.: +91-9958535142
† Presented at the 29th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-29); Available online: https://sciforum.net/event/ecsoc-29.

Abstract

Computational and machine learning approaches are playing a pivotal role in the identification, characterization, and targeting of noncanonical DNA structures, including G-quadruplexes, Z-DNA, hairpins, and triplexes. These configurations play critical roles in maintaining genomic stability, facilitating DNA repair, and regulating chromatin organization. Although the human genome predominantly adopts the B DNA conformation, evidence indicates that non-B DNA forms exert significant influence on gene expression and disease development. This highlights the need for dedicated computational frameworks to systematically investigate these alternative structures. Machine learning models encompassing supervised and unsupervised algorithms such as K Nearest Neighbours, Support Vector Machines, and deep learning architectures including Convolutional Neural Networks have shown considerable potential in predicting sequence motifs predisposed to forming non-B DNA conformations. These predictive tools contribute to identifying genomic regions associated with disease susceptibility. Complementary bioinformatics platforms and molecular docking tools, notably Auto Dock, along with chemical libraries like ZINC, facilitate the virtual screening of small molecules targeting specific DNA structures. Stabilizers of G quadruplexes, exemplified by CX 5461, have demonstrated therapeutic promise in BRCA deficient cancers, highlighting the translational impact of computational methods on drug discovery. Anticipating DNA structural shifts opens new avenues in personalized medicine for complex diseases, with computational chemistry and machine learning deepening our understanding of DNA topology and guiding smarter ligand design. The integrated approach proposed in this review addresses the previous studies done in this field and highlights the current limitations in structural genomics and advances the development of precision therapeutics aligned with individual genomic profiles.

Keywords: machine learning; drug discovery; personalized medicine; bioinformatics tools; non-canonical DNA

Academic Editor(s): Name

Published: date

Citation: Masroor, S.; Dudeja, C.; Sanka, R.; Sabikhi, Y.; Singh, A.; Mishra, A.; Gupta, R. Advanced Computational Frameworks for Characterizing Abnormal DNA Architectures and Their Implications in Genome Dynamics. *Chem. Proc.* **2025**, *volume number*, x. https://doi.org/10.3390/xxxxx

1. Introduction

The human genome primarily exists as right-handed B-DNA, featuring 10 base pairs per turn and distinct major and minor grooves that facilitate essential protein-DNA interactions during transcription and replication. However, DNA polymorphism allows the formation of non-canonical structures such as triplex DNA, Z-DNA, hairpins, G-quadruplexes, and cruciform DNA. These structures, influenced by nucleotide sequence, ligand binding, hydration, and super helical stress, are involved in replication, recombination, transcription, DNA repair, nucleosome assembly, and genome organization [1].

Triplex DNA was first described in 1957 when a third strand was found to bind along the major groove of a DNA duplex via Hoogsteen hydrogen bonding. Triplexes can be intermolecular or intramolecular. In H-DNA, the purine strand of the duplex folds back to pair with the pyrimidine strand in a parallel or antiparallel alignment. G-quadruplexes are non-canonical structures formed by guanine-rich sequences stabilized through Hoogsteen bonding between stacked G-tetrads (Burge et al., 2006). They are commonly found at telomeres and transcription start sites, where they influence gene regulation. Their regulatory role is conserved across diverse organisms and is linked to cancer, differentiation, and metabolism [2]. Z-DNA is a left-handed double helix that forms in alternating purine-pyrimidine sequences, particularly GC-rich regions, under physiological stress or in the presence of specific ions or chemicals. Z-DNA is stabilized by proteins such as the Zα domain of ADAR1 and ZBP1, and destabilized by agents like Actinomycin D and Distamycin A. The BZ junction marks the transition between B- and Z-DNA and is formed in negatively supercoiled regions generated during transcription [3].

Hairpin DNA forms when a single strand folds into a stem-loop via complementary base pairing, typically at palindromic sequences. These structures can form during replication or repair processes, and longer palindromes are linked to genetic instability. In histone mRNA, hairpins are essential for 3′ end processing, export, and translation, with key roles played by hairpin-binding protein and U7 snRNP [4].

Cruciform DNA arises from inverted repeat sequences and features stem-loop structures at a central branch point [5]. Found near promoters and replication origins, cruciforms facilitate chromatin remodelling, transcription, and genome stability by enabling DNA-protein interactions and bringing distant DNA elements into proximity.

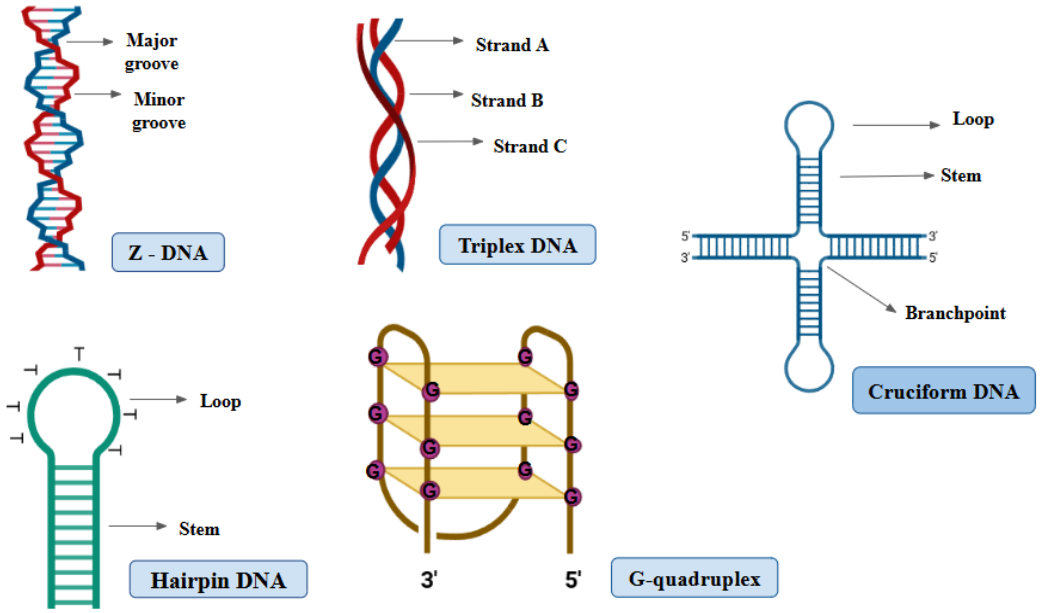

Figure 1. Non-Canonical DNA Structures.

2. Genome-Wide Databases and Computational Resources for DNA Structures

G-quadruplexes (G4s) are non-canonical, four-stranded structures that form in guanine-rich nucleic acids. The basic structural unit of a G4 is the G-tetrad—a planar arrangement of four guanine bases stabilized by Hoogsteen hydrogen bonds. With the increasing identification of G4-forming RNAs, various tools have been developed to aid in the structural determination of G4 DNA and RNA. In the field of bioinformatics, several tools have been introduced to predict G4 structures across different nucleic acid types. These tools are capable of analyzing molecular dynamics, calculating free energy, and performing molecular docking simulations to evaluate the stability and functional relevance of G4 structures [6]. Non-B DB is a comprehensive database containing predictions of 3,864,596 non-B DNA structures, including G-quadruplex motifs, across 12 mammalian genomes, including the human genome. It offers advanced search capabilities, allowing users to filter by species, DNA structure class, chromosome, gene type, and chromosomal location. Additional filters include sequence composition, motif type, and nucleotide tracts, enabling in-depth exploration of non-B DNA motifs. G4Hunter is a widely used predictive tool for identifying putative G-quadruplex sequences (PQS) in DNA or RNA, based on G-richness and G-skewness, the fraction of guanines in a sequence, and the G/C ratio between DNA strands [5]. Users can customize the search window size and threshold values, assigning weighted scores to guanine residues (e.g., G = 1, GG = 2) to refine predictions.

Non-B DNA structures arise when DNA deviates from the classical Watson-Crick B-form, often due to specific sequence motifs or environmental conditions. These alternative conformations can impair replication, increase error rates, and promote mutagenesis, leading to genome-wide variation in mutation rates. Motifs capable of forming non-B structures range from a few bases to several hundred nucleotides and are randomly distributed across the genome [7]. They are known to disrupt replication and transcription, contributing to genomic instability, particularly in cancer. In silico tools such as G4Hunter, R-loop tracker, and other structure-predicting algorithms can identify overlaps between G-quadruplexes, R-loops, and other non-B structures. These tools use algorithms that select common overlapping regions to optimize memory and computational efficiency [8]. Mapping G4-forming regions genome-wide allows researchers to identify G-quadruplexes in key regulatory areas such as promoters, telomeres, and untranslated regions (UTRs)—sites often linked to gene regulation. Furthermore, small molecules that stabilize or destabilize G4 structures are being explored as therapeutic agents, particularly in oncogenes like *MYC* or in telomeric regions. G4Hunter is instrumental in pinpointing promising G4 targets for drug development.

Table 1. Key databases with the type of structures, data sources, and key features.

Database	Type of Structures Cataloged	Data Sources	Experimental Validation	Key Features	References
G4Hunter	G-quadruplex (G4)	Genomic sequences	In-silico (scoring based on guanine content and sequence)	Predict potential G4-forming sequences,	[5,9]
Non-B DNA	Various non-B DNA structures (Z-DNA, G4, triplexes, cruciform, etc.)	Genomic sequences (e.g., human, mouse, bacterial)	Experimental data integrated alongside in silico predictions	collection of non-B DNA motifs, links to disease associations	[10]

QuadBase2	G-quadruplex (G4)	Human genome, model organisms (plants, yeast, etc.)	Experimental data from literature and high-throughput sequencing	Contains experimentally validated G-quadruplex	[11]
Triplex-Inspector	Triplex-forming oligonucleotides (TFOs) and triplex DNA	Genomic sequences (custom uploads, model organisms)	Based on sequence features, in silico predictions	Detect triplex DNA, gene regulation	[11]

3. Machine Learning and Bioinformatics Approaches for Predicting Non-B DNA Structures

Machine Learning for Detecting Gene-Gene Interactions: Machine learning (ML) is a subfield of computer science and artificial intelligence that focuses on developing algorithms capable of learning from data and making predictions. In genetics, ML can be used to predict phenotypes (such as traits or disease risks) from biomarkers like DNA sequences, making it a supervised learning problem where the genotype is the input and the phenotype is the output [12]. Machine learning is broadly categorized into three types that is supervised learning to learn from labeled training data to predict outcomes on new, unseen data. It maps input features to known outputs, enabling the model to make accurate predictions, and unsupervised learning that works with unlabeled data to identify patterns or groupings without predefined outcomes. It is used in clustering and dimensionality reduction and reinforcement learning, which involves learning through interaction with an environment, where the algorithm improves its decisions over time based on feedback or rewards [13].

3.1. Common Machine Learning Algorithms in Genomics

3.1.1. K-Nearest Neighbors (KNN)

KNN is a supervised learning algorithm used for classification and regression. It classifies a data point based on the majority class among its K nearest neighbors in the feature space. The distance metric and the value of K are critical to performance. KNN is simple and effective for datasets where similar instances belong to the same class [13].

3.1.2. Artificial Neural Networks (ANNs)

ANNs are inspired by the biological brain and consist of layers of interconnected neurons. They are effective for pattern recognition and classification tasks. Multi-layer perceptrons (MLPs), functional link ANNs, and two-class SVMs have been used to identify novel disease genes using topological features from protein–protein interaction (PPI) networks. ANNs have also been applied to gene expression datasets to distinguish between disease states [14].

3.1.3. Convolutional Neural Networks (CNNs)

CNNs are a type of deep learning architecture primarily used in image classification, including clinical and biological imaging. CNNs capture spatial relationships between features and are robust to transformations such as scaling and translation. They are increasingly used for biomarker discovery and disease prediction [15].

3.1.4. Random Forest (RF)

Random Forest is a supervised ensemble learning method based on decision trees. It uses the bagging technique to train multiple decision trees and aggregates their results for improved prediction accuracy. Random Forests are known for their robustness and ability

to handle high-dimensional genomic data. In one study, it achieved 94.74% accuracy in a binary classification task.

3.1.5. Support Vector Machines (SVMs)

SVMs are kernel-based classifiers that find the optimal hyperplane separating data into distinct groups. They are widely used in bioinformatics due to their high accuracy and ability to model complex, high-dimensional data. SVMs are particularly effective in identifying subtle differences between biological classes.

Machine learning techniques, when applied to gene-gene interaction studies, can uncover complex relationships within genomic data. By learning from large-scale, high-throughput datasets, these models facilitate the discovery of novel biomarkers, disease-associated genes, and predictive genomic signatures, making them essential tools in personalized medicine and genomic research.

The machine learning models leverage sequence-intrinsic features such as G-tract length, loop composition, and flanking nucleotide context that correlate with the propensity for adopting specific topologies rather than assuming a single static fold. This design enables the model to capture the conformational diversity inherent to polymorphic non-canonical DNA structures. Recent studies have shown that sequence-derived information alone can predict G-quadruplex (G4) topology with high accuracy. Ref. [16] provided evidence that G4ShapePredictor can predict parallel, antiparallel, and hybrid G4 structures using 482 distinct sequence–topology pairs that were validated in the lab. The results provide confidence that patterns encoded into the sequence can predict the dominant conformations that will be adopted in a biological setting, reinforcing the reliability of computational models for even DNA architectures that are dynamically equilibrating or equilibrium conformations.

Table 2. Machine Learning Techniques and Models for Detecting Gene Abnormalities.

Technique	Model Used	Purpose	Accuracy	Reference
Prediction of G-quadruplexes	Convolutional Neural Network (CNN)	Predict G-quadruplex-forming regions in DNA sequences	95.2% (AUC-ROC)	[17]
G4Boost: quadruplex identification and stability prediction	XGBoost regression model	Determine the sequences, nucleotide compositions, and estimated structural topologies of G4 motifs to forecast their secondary structure	93%	[18]
Using omics data, a method for predicting functional Z-DNA areas	convolution neural networks (CNN), Recurrent neural networks (RNN), Hybrid CNN-RNN models	**DeepZ**—Developed using chromosome accessibility, transcription factor/RNA polymerase binding, and epigenetic marker maps	86.6%	[19]
Identifying proteins that bind to DNA using features based on composition and	Random Forest, Support Vector Machine	"DNAPred_Prot" DNA-binding protein using sequence features.	91.47%	[20]

position is the focus of DNAPred_Prot.	Artificial Neural Network			
IoMT-based prediction of mitochondrial and inherited illnesses	support vector machine (SVM) K-Nearest Neighbor (KNN)	Analysis of genetic data for early and accurate diagnosis.	94.99%	[21]

4. Applications of DNA Structure Prediction in Medicine and Disease

Prediction of DNA structure has become pivotal to contemporary biology and medicine, enhancing disease mechanism insights and enabling the formulation of targeted treatments. Methods like AlphaFold have revolutionized structural biology by allowing massive protein and DNA structure predictions [16]. In cancer, mutations in genes and chromosomes affect normal protein behavior, and structure prediction assists in the identification of oncogenes, inhibitor design, and the targeting of immunotherapies. Correspondingly, in neurodegenerative diseases, structural defects like repeat expansions in Huntington's disease or methylation pattern changes in Alzheimer's disease emphasize the need for predictive models in revealing disease mechanisms and therapeutic targets [23,24]. In addition to disease-specific applications, DNA structure prediction supports personalized medicine, where mutation profiling and pharmacogenomics facilitate precision oncology and drug response prediction in individual patients [25]. Collectively, these uses prove its power to transform diagnostics, drug discovery, and precision healthcare.

5. Conclusions and Future Directions

Research on non-canonical DNA structures such as G-quadruplexes, Z-DNA, hairpins, and triplexes has transformed our understanding of genome regulation, especially their roles in DNA repair, genetic stability, and transcriptional control. The use of computational methods and machine learning has become key in identifying these structures, studying their ligand interactions, and revealing their functional importance. Combining specialized databases with advanced machine learning models improves the classification and interpretation of genomic data, leading to more accurate predictions of non-B DNA motifs.

Future efforts should focus on refining computational methods to better capture the dynamics and biological roles of these structures. Expanding algorithmic capabilities and incorporating more diverse, high-quality datasets will further enhance model performance and predictive power. These advances may pave the way for novel therapeutic strategies, particularly in cancer and neurodegenerative diseases. The integration of AI, big data analytics, and bioinformatics tools in DNA structural research holds great promise for personalized medicine. Tailoring treatments based on an individual's unique genomic architecture can revolutionize disease diagnosis and management. Ultimately, continued exploration of the relationship between DNA structure and function will be essential for unlocking new therapeutic avenues and improving patient outcomes in precision medicine.

Author Contributions:

Funding:

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement:

Conflicts of Interest:

References

1. Wang, G.; Vasquez, K.M. Impact of Alternative DNA Structures on DNA Damage, DNA Repair, and Genetic Instability. *DNA Repair* **2014**, *19*, 143–151. https://doi.org/10.1016/j.dnarep.2014.03.017.
2. Lopez, C.R.; Singh, S.; Hambarde, S.; Griffin, W.C.; Gao, J.; Chib, S.; Yu, Y.; Ira, G.; Raney, K.D.; Kim, N. Yeast Sub1 and Human PC4 Are G-Quadruplex Binding Proteins That Suppress Genome Instability at Co-Transcriptionally Formed G4 DNA. *Nucleic Acids Res.* **2017**, *45*, 5850–5862. https://doi.org/10.1093/nar/gkx201.
3. Ha, S.C.; Lowenhaupt, K.; Rich, A.; Kim, Y.-G.; Kim, K.K. Crystal Structure of a Junction between B-DNA and Z-DNA Reveals Two Extruded Bases. *Nature* **2005**, *437*, 1183–1186. https://doi.org/10.1038/nature04088.
4. Sch�mperli, D.; Pillai, R.S. The Special Sm Core Structure of the U7 snRNP: Far-Reaching Significance of a Small Nuclear Ribonucleoprotein. *CMLS Cell. Mol. Life Sci.* **2004**, *61*, 2560–2570. https://doi.org/10.1007/s00018-004-4190-0.
5. Brázda, V.; Laister, R.C.; Jagelská, E.B.; Arrowsmith, C. Cruciform Structures Are a Common DNA Feature Important for Regulating Biological Processes. *BMC Mol. Biol.* **2011**, *12*, 33. https://doi.org/10.1186/1471-2199-12-33.
6. Puig Lombardi, E.; Londoño-Vallejo, A. A Guide to Computational Methods for G-Quadruplex Prediction. *Nucleic Acids Res.* **2020**, *48*, 1–15. https://doi.org/10.1093/nar/gkz1097.
7. Georgakopoulos-Soares, I.; Morganella, S.; Jain, N.; Hemberg, M.; Nik-Zainal, S. Noncanonical Secondary Structures Arising from Non-B DNA Motifs Are Determinants of Mutagenesis. *Genome Res.* **2018**, *28*, 1264–1271. https://doi.org/10.1101/gr.231688.117.
8. Havlík, J.; Brázda, V.; Staněk, K.; Ježek, M.; Šťastný, J. Feature-Overlapper: The Tool for DNA Analysis Overlap. *Softw. Impacts* **2023**, *16*, 100498. https://doi.org/10.1016/j.simpa.2023.100498.
9. Bedrat, A.; Lacroix, L.; Mergny, J.-L. Re-Evaluation of G-Quadruplex Propensity with G4Hunter. *Nucleic Acids Res.* **2016**, *44*, 1746–1759. https://doi.org/10.1093/nar/gkw006.
10. Cer, R.Z.; Donohue, D.E.; Mudunuri, U.S.; Temiz, N.A.; Loss, M.A.; Starner, N.J.; Halusa, G.N.; Volfovsky, N.; Yi, M.; Luke, B.T.; et al. Non-B DB v2.0: A Database of Predicted Non-B DNA-Forming Motifs and Its Associated Tools. *Nucleic Acids Res.* **2012**, *41*, D94–D100. https://doi.org/10.1093/nar/gks955.
11. Dhapola, P.; Chowdhury, S. QuadBase2: Web Server for Multiplexed Guanine Quadruplex Mining and Visualization. *Nucleic Acids Res.* **2016**, *44*, W277–W283. https://doi.org/10.1093/nar/gkw425.
12. Le, D.-H. Machine Learning-Based Approaches for Disease Gene Prediction. *Brief. Funct. Genom.* **2020**, *19*, 350–363. https://doi.org/10.1093/bfgp/elaa013.
13. Sharma, B.R.; Kumar, V.; Gat, Y.; Kumar, N.; Parashar, A.; Pinakin, D.J. Microbial Maceration: A Sustainable Approach for Phytochemical Extraction. *3 Biotech.* **2018**, *8*, 401. https://doi.org/10.1007/s13205-018-1423-8.
14. Zhi, J.; Sun, J.; Wang, Z.; Ding, W. Support Vector Machine Classifier for Prediction of the Metastasis of Colorectal Cancer. *Int. J. Mol. Med.* **2018**, *41*, 1419–1426. https://doi.org/10.3892/ijmm.2018.3359.
15. Shon, H.S.; Yi, Y.; Kim, K.O.; Cha, E.-J.; Kim, K.-A. Classification of Stomach Cancer Gene Expression Data Using CNN Algorithm of Deep Learning. *J. Biomed. Transl. Res.* **2019**, *20*, 15–20. https://doi.org/10.12729/jbtr.2019.20.1.015.
16. Liew, D.; Lim, Z.W.; Yong, E.H. Machine Learning-Based Prediction of DNA G-Quadruplex Folding Topology with G4ShapePredictor. *Sci. Rep.* **2024**, *14*, 24238. https://doi.org/10.1038/s41598-024-74826-2.
17. Yang, B.; Guneri, D.; Yu, H.; Wright, E.P.; Chen, W.; Waller, Z.A.E.; Ding, Y. Prediction of DNA I-Motifs via Machine Learning. *Nucleic Acids Res.* **2024**, *52*, 2188–2197. https://doi.org/10.1093/nar/gkae092.
18. Cagirici, H.B.; Budak, H.; Sen, T.Z. G4Boost: A Machine Learning-Based Tool for Quadruplex Identification and Stability Prediction. *BMC Bioinform.* **2022**, *23*, 240. https://doi.org/10.1186/s12859-022-04782-z.
19. Beknazarov, N.; Jin, S.; Poptsova, M. Deep Learning Approach for Predicting Functional Z-DNA Regions Using Omics Data. *Sci. Rep.* **2020**, *10*, 19134. https://doi.org/10.1038/s41598-020-76203-1.
20. Barukab, O.; Khan, Y.D.; Khan, S.A.; Chou, K.-C. DNAPred_Prot: Identification of DNA-Binding Proteins Using Composition- and Position-Based Features. *Appl. Bionics Biomech.* **2022**, *2022*, 5483115. https://doi.org/10.1155/2022/5483115.
21. Rahman, A.; Nasir, M.U.; Gollapalli, M.; Alsaif, S.A.; Almadhor, A.S.; Mehmood, S.; Khan, M.A.; Mosavi, A. IoMT-Based Mitochondrial and Multifactorial Genetic Inheritance Disorder Prediction Using Machine Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 2650742. https://doi.org/10.1155/2022/2650742.

22. Yang, B.; Guneri, D.; Yu, H.; Wright, E.P.; Chen, W.; Waller, Z.A.E.; Ding, Y. Prediction of DNA I-Motifs via Machine Learning. *Nucleic Acids Res.* **2024**, *52*, 2188–2197. https://doi.org/10.1093/nar/gkae092.
23. Tabrizi, S.J.; Flower, M.D.; Ross, C.A.; Wild, E.J. Huntington Disease: New Insights into Molecular Pathogenesis and Therapeutic Opportunities. *Nat. Rev. Neurol.* **2020**, *16*, 529–546. https://doi.org/10.1038/s41582-020-0389-4.
24. Kaur, G.; Rathod, S.S.S.; Ghoneim, M.M.; Alshehri, S.; Ahmad, J.; Mishra, A.; Alhakamy, N.A. DNA Methylation: A Promising Approach in Management of Alzheimer's Disease and Other Neurodegenerative Disorders. *Biology* **2022**, *11*, 90. https://doi.org/10.3390/biology11010090.
25. Nussinov, R.; Jang, H.; Tsai, C.-J.; Cheng, F. Review: Precision Medicine and Driver Mutations: Computational Methods, Functional Assays and Conformational Principles for Interpreting Cancer Drivers. *PLoS Comput. Biol.* **2019**, *15*, e1006658. https://doi.org/10.1371/journal.pcbi.1006658.