

Construction of DNA/RNA Triplex Helices Based on GAA/TTC Trinucleotide Repeats

Jiahui Zhang¹, Ashkan Fakharzadeh¹, Feng Pan², Christopher Roland¹ and Celeste Sagui^{1,*}

¹Department of Physics, North Carolina State University, Raleigh, USA; ²Department of Statistics, Florida State University, Tallahassee, USA

*For correspondence: sagui@ncsu.edu

[Abstract] Atypical DNA and RNA secondary structures play a crucial role in simple sequence repeat (SSR) diseases, which are associated with a class of neurological and neuromuscular disorders known as “anticipation diseases,” where the age of disease onset decreases and the severity of the disease is increased as the intergenerational expansion of the SSR increases. While the mechanisms underlying these diseases are complex and remain elusive, there is a consensus that stable, non-B-DNA atypical secondary structures play an important – if not causative – role. These structures include single-stranded DNA loops and hairpins, G-quartets, Z-DNA, triplex nucleic acid structures, and others. While all of these structures are of interest, structures based on nucleic acid triplexes have recently garnered increased attention as they have been implicated in gene regulation, gene repair, and gene engineering. Our work here focuses on the construction of DNA triplexes and RNA/DNA hybrids formed from GAA/TTC trinucleotide repeats, which underlie Friedreich’s ataxia. While there is some software, such as the Discovery Studio Visualizer, that can aid in the initial construction of DNA triple helices, the only option for the triple helix is constrained to be that of an antiparallel pyrimidine for the third strand. In this protocol, we illustrate how to build up more generalized DNA triplexes and DNA/RNA mixed hybrids. We make use of both the Discovery Studio Visualizer and the AMBER simulation package to construct the initial triplexes. Using the steps outlined here, one can – in principle – build up any triple nucleic acid helix with a desired sequence for large-scale molecular dynamics simulation studies.

Keywords: DNA/RNA, Triplex helices, Molecular dynamics, Trinucleotide repeats

[Background] Simple sequence repeats (SSRs) – which represent about 3% of the entire human genome – typically consist of 1 to 6 nucleotides that repeat up to 30 times or more (Ellegren *et al.*, 2004; Subramanian *et al.*, 2003). Among the various possible SSRs, trinucleotide repeats (TRs) represent one of the most common types in the exome of all eukaryotic genomes (Toth *et al.*, 2000). Many TRs exhibit “dynamic mutations” that do not follow Mendelian inheritance, which states that mutations in a single gene may be stably transmitted between generations (Caburet *et al.*, 2005). This can lead to genetic diseases where, in successive generations, the age of disease onset decreases and disease severity increases (Mirkin, 2006). These mutations – whose probability also increases with generations – are due to the intergenerational expansion of TRs. After a certain threshold in the repeat number of TRs is reached, the probability of further TR expansion and the severity of the disease increase with the number of repeats. The dynamic mutations associated with TRs cause severe neurodegenerative and neuromuscular disorders known generically as Trinucleotide (or Triplet) Repeat Expansion Diseases

(TREDs), which lead to cell toxicity and death (Wells *et al.*, 1998; Orr *et al.*, 2007; Wells *et al.*, 2005). To date, about 50 DNA expandable SSR diseases have been identified, and their number is expected to grow. The TR expansions are believed to be caused by DNA slippage during replication, repair, transcription, or recombination.

Although the mechanisms underlying TREDs may be quite complex, some simple trends are remarkably robust. In particular, there is a correlation between the repeat number beyond the repeat threshold and the probability of further expansion and increased pathology. Another important breakthrough has been the recognition that stable, non-B-DNA secondary structure in the expanded repeats is an important factor causing the expansion and disease (McMurray *et al.*, 1999). As such, expandable repeats are known to display atypical structural characteristics such as single-stranded hairpins, Z-DNA, G-quartets, triple helix structures, and slip-stranded duplexes. It is therefore believed that understanding the structural and dynamical characteristics of these atypical secondary structures is important for ultimately unraveling the puzzle of TREDs.

While our previous work on TRs and hexanucleotide repeats in the context of C9FTD/ALS diseases (Zhang *et al.*, 2017a and 2017b) was centered on understanding single-stranded loops and hairpins (Pan *et al.*, 2017, 2018a and 2018b; Xu *et al.*, 2020), here we focus on the construction of triplexes associated with GAA/TTC TRs (Zhang *et al.*, 2020). These are associated with Friedreich's ataxia (Grabczyk *et al.*, 2000), which is caused by the expansion of GAA in the first intron of the frataxin gene. Experimentally, these repeats have been observed to form either triplexes or R-loops. DNA triplexes or triple-stranded DNA or H-DNA were first reported in 1957 (Felsenfeld *et al.*, 1957). These non-canonical three-stranded helices consist of a Watson-Crick paired helical duplex and a third strand that binds to the duplex via Hoogsteen or reversed Hoogsteen hydrogen bonds. R-loops, on the other hand, represent three-stranded nucleic acid structures consisting of a hybrid RNA:DNA duplex (formed by a template DNA and the RNA strands) in conjunction with the displaced, non-template single-stranded DNA. Both triplexes and R-loops can have cellular functions and are essential for gene therapy (Kaji *et al.*, 2001; Seidman *et al.*, 2003).

There is much to be learned about the microscopics of DNA triplexes and R-loops; in particular, many of the atomistic aspects of these atypical nucleic acid structures remain elusive. Hence, we have recently examined the structure and stability of DNA triplexes and RNA/DNA hybrids associated with GAA/TTC TRs (Zhang *et al.*, 2020). In these structures, the third strand is inserted into the major groove or minor groove in pure RNA triplexes (Szewczak *et al.*, 1998). However, since the minor groove RNA triplexes are unstable (Devi *et al.*, 2015), we only consider major groove RNA triplexes in this protocol. Our study was based on large-scale classical Molecular Dynamics (MD) simulations. The initial modeling of the triple helices was performed with the Discovery Studio Visualizer (Visualizer, 2005), and we made use of the AMBER simulation package as an optimization and sampling tool for exploring the structure and stability of DNA triplexes and selected R-loops (Case *et al.*, 2020). In this bio-protocol paper, we provide details for the initial modeling and construction of the triple helices associated with GAA/TTC TRs. Specifically, we focus on the sequence GAA/TTC(UUC) as an example.

We primarily discuss how to build up DNA-based triplexes and mixed DNA/RNA structures, as shown

below in Figure 1. With the models constructed in this bio-protocol, it is straightforward to investigate triple helices *per se* and their interactions with other biomolecules.

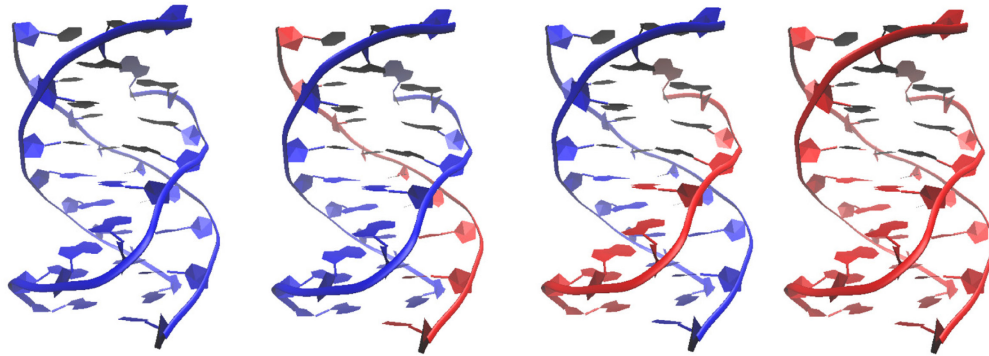


Figure 1. Structures (from left to right) of DNA triple helix, DNA-RNA:DNA hybrid triple helix, RNA-DNA:DNA hybrid triple helix, and RNA triple helix. DNA strands are colored in blue, and RNA strands are colored in red.

Software

1. Discovery Studio Visualizer version 2019, or higher
Discovery Studio Visualizer is a free software developed by Dassault Systemes BIOVIA. Access this software at: <https://discover.3ds.com/discovery-studio-visualizer-download>.
2. Amber, version 16 or higher
Amber is a suite of biomolecular simulation programs for large-scale MD studies. Access Amber at: <http://ambermd.org/>.

Procedure

- A. Initial Construction of a DNA Triple Helix
1. Open Discovery Studio Visualizer.
 2. First, build up a homopurine/homopyrimidine B-DNA double helix with the desired sequence.
The basic strategy of building a triple helix model is shown in Figure 2.

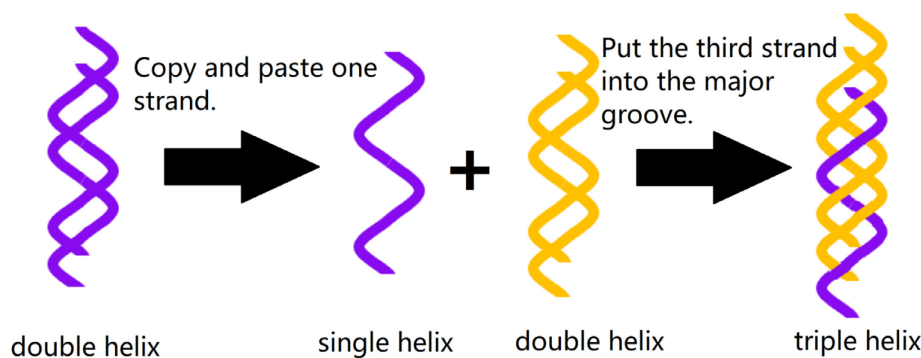


Figure 2. Scheme illustrating the construction of a triple helix

For example, suppose we want to build up the parallel sequence as shown in Figure 3(a).

3. Change the display style to arrows and rings and turn off the atom display style.
4. Select the wanted strand (all purine or all pyrimidine) and then select copy and paste. If a shifted triplex sequence is desired, then change the residues name on the third strand to the desired sequence. Next, copy and paste the pyrimidine strand from the duplex, as shown in Figure 3(b), to get the sequence Figure 3(c).

As the desired triplex involves a shifted sequence, rename the residue name to obtain Figure 3(d).

5. Carefully put the pasted strand into the major groove of the DNA duplex with the desired orientation, as shown in Figure 4 and in Video 1. Then our example sequence thus becomes Figure 3(e).

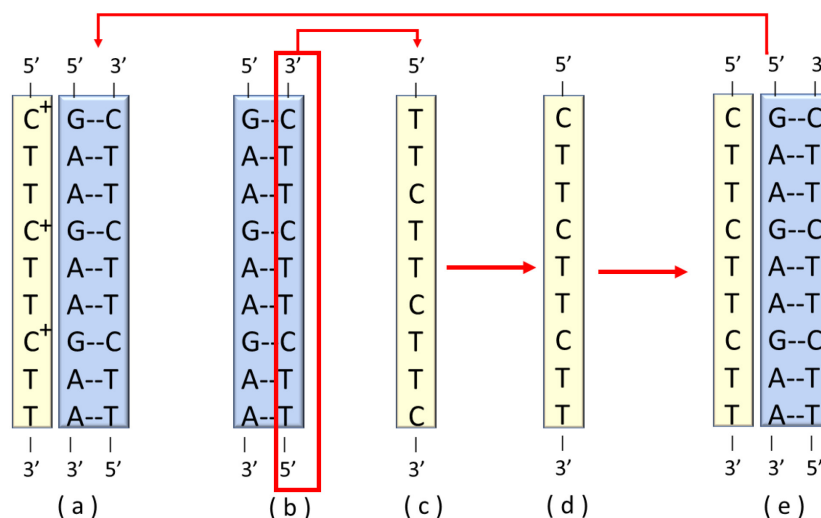


Figure 3. Sample sequence for our procedure. (a) represents the triplex we want to build. (b) is initial duplex from which we obtained the strand shown in (c). The shifted strand obtained from the previous step is shown in (d). This shifted strand is placed into the major groove of the duplex as shown in (e). Protonating the cytosines in the third strand finally leads to (a).

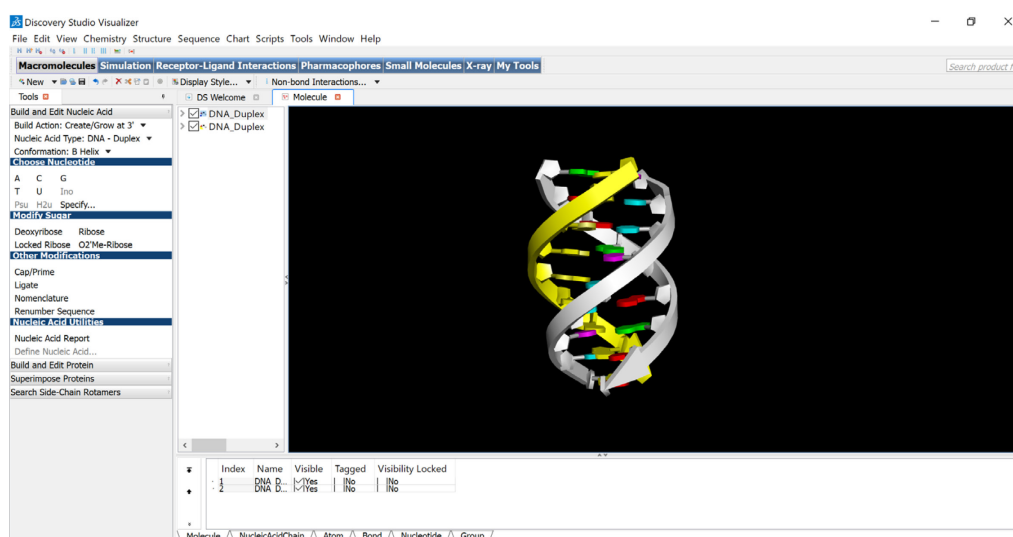
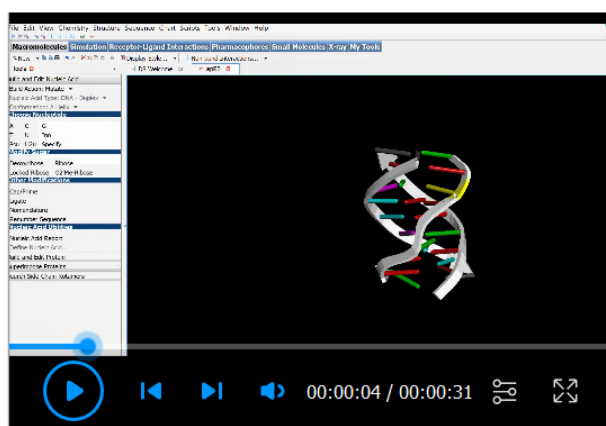


Figure 4. Snapshot for the placement of the third strand into the major groove of the DNA duplex



Video 1. A short illustration of how we build up the initial model of a triple helix

6. Roughly adjust the position of each residue to avoid any overlap between the residues, as shown in Video 1. For example, when the oxygen atoms on the phosphate groups are too close, we manually slightly move the position of one of the oxygen atoms.
7. Change the name of the third-strand bases to construct the desired sequence. Name protonated DC just DC, protonated DA just DA, *etc.* (same for RNA). We will illustrate how to deal with the protonated cases later.
8. Change the display style back to atom/ball and stick.
9. Precisely adjust each atom on the third strand to avoid high potential energy collisions during subsequent MD runs and to construct the hydrogen bond structure as illustrated in Figure 5 and Figure 6.

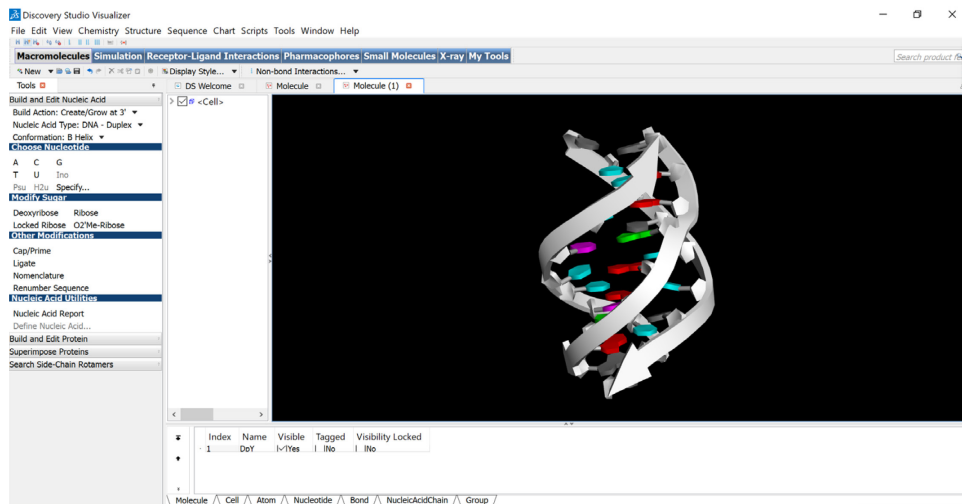


Figure 5. Final conformation for the example given in Figure 3 and built through steps 1-14.

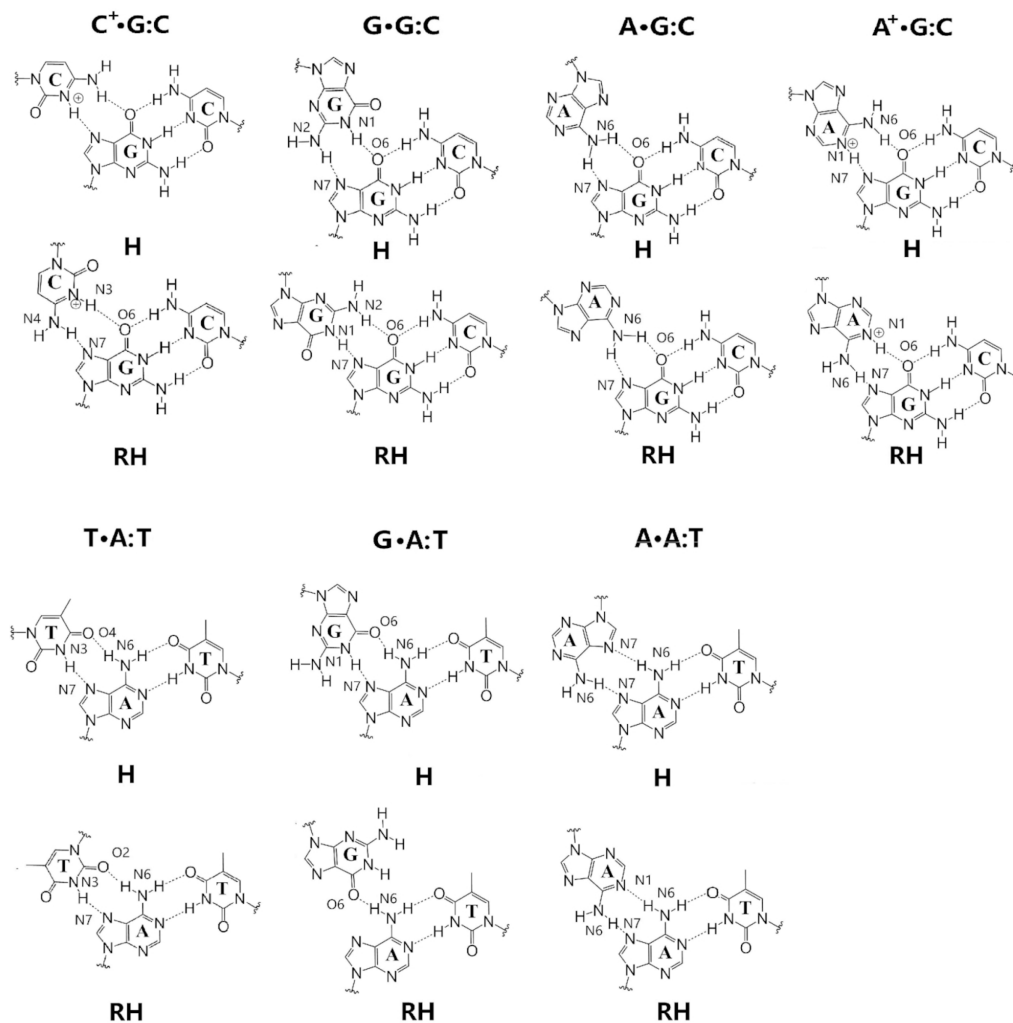


Figure 6. Initial hydrogen bond constraints for a DNA triple helix

10. Select edit, select. Change the property to be element, then select hydrogen. After selecting all hydrogen atoms, type *Del* to delete all hydrogen atoms as some of the hydrogen atoms' names are different than those in Amber and will cause errors.
11. Save the structure as PDB.
12. Edit the PDB file, deleting all the connection relations at the end of the PDB file.
13. If there are protonated bases in the desired sequence, change the corresponding residue name DC to DCP, DA to DAP. If the protonated C is at 5', rename it as D5C. If the protonated C is at 3', rename it as D3C. If the protonated A is at 5', rename it as D5A. If the protonated A is at 3', rename it as D3A, *etc.* While this nomenclature seems somewhat arbitrary and inconsistent, it is mandated by the fact that tLeap (included in Amber) only recognizes three letters at most. Then our example sequence will be Figure 3(a). The configuration is shown in Figure 5.
14. Go to section E.

B. Initial Construction of DNA·RNA:DNA Hybrid Triple Helix

1. Open Discovery Studio Visualizer.
2. Build up a homopurine/homopyrimidine B-DNA or A-DNA double helix with the desired sequence.
3. Change the display style to arrows and rings and turn off the atom display style.
4. Repeat steps 4-7 in section A.
5. Select the desired strand and change it from DNA to RNA. This is done by selecting the desired part of nucleic acid and then clicking the "Ribose" on the options manual "Modify Sugar."
6. Repeat steps 8-14 in section A.

C. Initial Construction of RNA·DNA:DNA Hybrid Triple Helix

1. Open Discovery Studio Visualizer.
2. Load the DNA triple helix that has been stabilized after MD (in our case, at the end of 1 μ s MD simulation [Zhang *et al.*, 2020]) with the desired sequence (where T's will be exchanged by U's as described below). Possible initial hydrogen bond patterns for this structure are given in Figure 7.

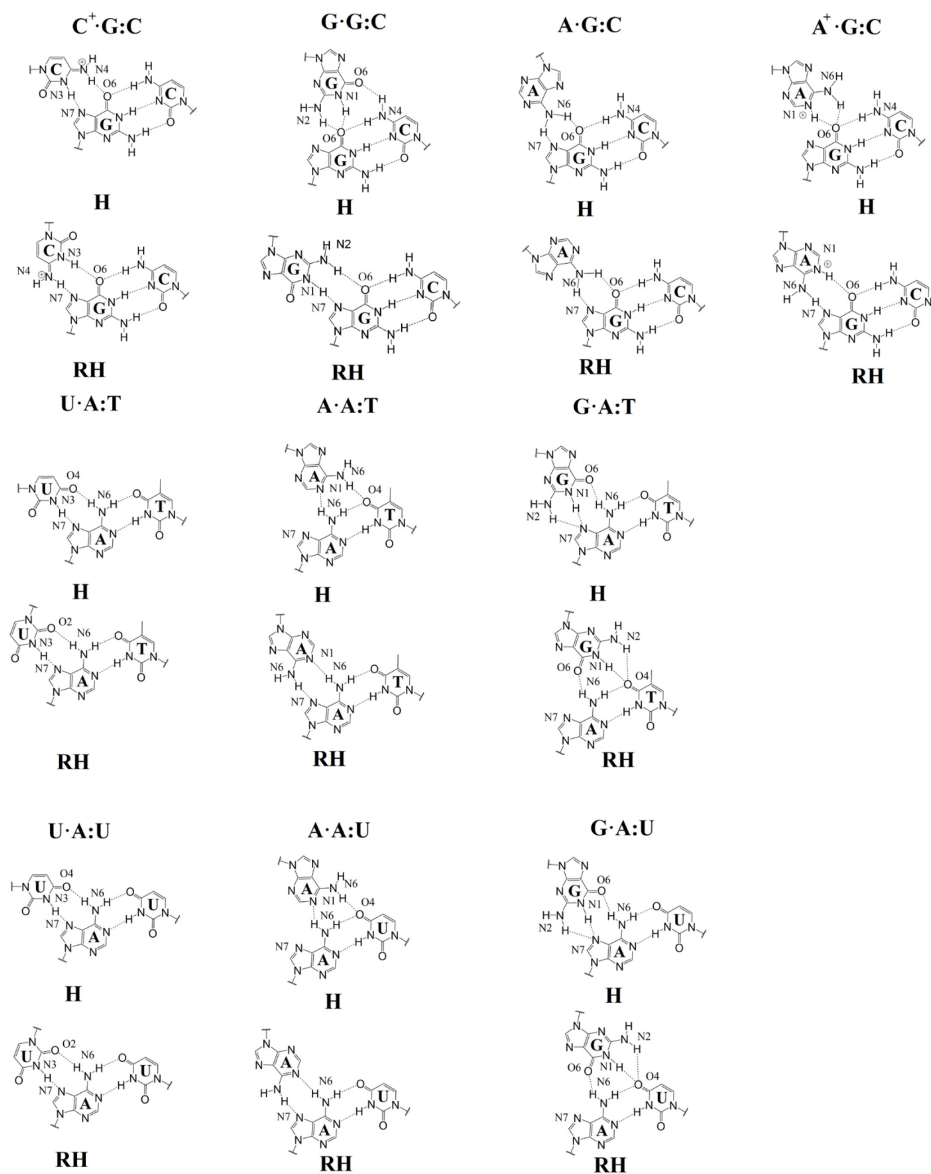


Figure 7. Initial hydrogen bond constraints for RNA-DNA:DNA and RNA triplex helices

3. Change the terminal residue, such as DG5, back to DG in order to avoid errors.
4. If there are protonated bases, change them back to ordinary bases.
5. Change the wanted part from DNA to RNA.
6. Repeat steps 10-12 in section A.
7. If there are protonated bases in the wanted sequence, change the corresponding residue name C to CP and A to AP. If the protonated C is at 5', rename it as C5P. If the protonated C is at 3', rename it as C3P. If the protonated A is at 5', rename it as A5P. If the protonated A is at 3', rename it as A3P.
8. Go to section E.

D. Initial Construction of RNA Triplex Helix

Repeat all steps of section C.

E. Details of Molecular Dynamics Simulations

1. Use tLeap in the Amber package to generate the topological file, initial coordinate file, and complete PDB file. We use the BSC1 force field (Ivani *et al.*, 2016), BSC0 (Pérez *et al.*, 2007) + OL3 (Zgarbová *et al.*, 2011) force field, and protonated Amber forcefield (Weiner *et al.*, 1986).
2. Use periodic boundary conditions.
3. The unit cell size should be large enough to avoid the self interaction between nearest nucleic acids, whose distance must be larger than the electrostatic and van der Waals cutoff (9 Å). In our example, there are about 840-870 atoms in each triplex, and we need to create an octahedral box that respects this distance. For this, we set the minimum distance between the nucleic acid and the box boundary as 8 Å, which results in a box size ~90,000 Å³. After this, TIP3P (Jorgensen *et al.*, 1983) water molecules are added randomly to the box. (Figure 8)

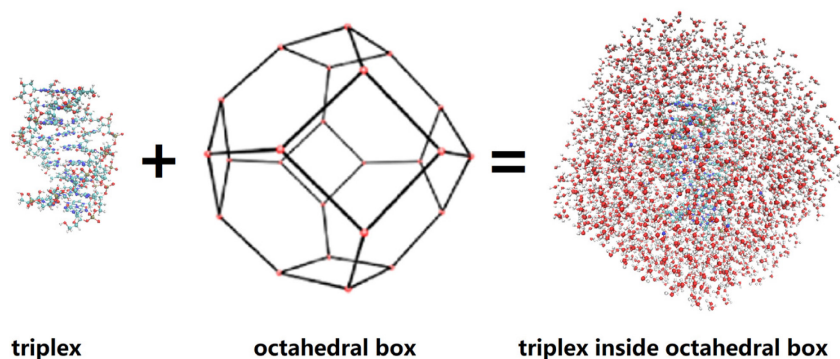


Figure 8. Illustration of an octahedral box of triplex

4. We use Na⁺ ions (Joung *et al.*, 2008) to neutralize the system. The number of Na⁺ equals the number of negative charges in the triplex.
5. Add hydrogen bond constraints, as shown in Figure 6 (with some T replaced by U) or Figure 7. Add the hydrogen bond constraints as instructed by the website: <https://ambermd.org/tutorials/advanced/tutorial4/index.htm>. Edit: \${AMBERHOME}/dat/map.DG-AMBER if there is any error reported. Add some sentences like this to the file:

RESIDUE C3P

MAPPING H3 = H3

Example hydrogen bond constraint input file can be written as:

1 DG5 H21	18 DC3 O2	1.9	1.9
1 DG5 H1	18 DC3 N3	1.9	1.9
1 DG5 O6	18 DC3 H41	1.9	1.9
1 DG5 O6	18 DC3 N4	2.9	2.9
1 DG5 N1	18 DC3 N3	2.9	2.9

1 DG5 N2	18 DC3 O2	2.9	2.9
9 DA3 N1	10 DT5 H3	1.9	1.9
9 DA3 H61	10 DT5 O4	1.9	1.9
9 DA3 N6	10 DT5 O4	2.9	2.9
9 DA3 N1	10 DT5 N3	2.9	2.9
27 D3A H1	1 DG5 O6	1.9	1.9
27 D3A N1	1 DG5 O6	2.9	2.9
27 D3A H61	1 DG5 N7	1.9	1.9
27 D3A N6	1 DG5 N7	2.9	2.9
19 DG5 H21	10 DT5 O4	1.9	1.9
19 DG5 N2	10 DT5 O4	2.9	2.9
19 DG5 H1	10 DT5 O4	1.9	1.9
19 DG5 N1	10 DT5 O4	2.9	2.9
19 DG5 O6	9 DA3 H62	1.9	1.9
19 DG5 O6	9 DA3 N6	2.9	2.9

6. Perform MD simulations to optimize the initial structures. Set the electrostatic cutoff to be 9.0 Å. Set the van der Waals cutoff to be 9.0 Å. Use Langevin dynamics with a coupling parameter of 1.0 ps⁻¹ to control the temperature. Use the SHAKE algorithm to deal with bonds involving hydrogen atoms. The MD simulation for nucleic acids is explained in detail in Šponer and Filip (2006).
7. Minimize the energy for the initial conformations obtained by modeling: initially, keep the nucleic acid and ions fixed; then, slowly (in steps) lift the constraints, allowing them to move.
8. Then gradually raise the temperature from 0 to 300 K over a 50 ps run with a 1 fs time step with the nucleic acid and ions constrained.
9. Then use a 100 ps run at constant volume to gradually reduce the restraining harmonic constants for nucleic acids and ions.
10. Then use a 1 μs run at constant pressure to gradually optimize the volume of the unit cell; the Berendsen pressure coupling method is the one used.
11. The final structure of the MD simulation may be used as the initial structure for further research.

Notes

1. Models built up either from B-DNA and A-DNA initial structures will converge to the same structure provided that triplex is stable (unstable triplexes fall apart). The duplex part of a given triple helix is somewhere between that of B- and A-DNA structures.
2. When changing a protonated DNA residue to a protonated RNA residue, check the position of O2' atom on the sugar ring.
3. Models built up with different hydrogen bond constraints will converge to a similar structure after an unrestrained MD simulation provided that the sequence is stable. The initial hydrogen bond

constraint is just a way to keep the third strand attached to the duplex part.

4. While we use AMBER ver. 20 (Case *et al.*, 2020) for our MD runs, these may also be performed using other simulation packages such as NAMD (Nelson *et al.*, 1996).

Acknowledgments

Funding was provided by National Institute of Health (NIH) grant R01GM118508. Original paper behind this work is Zhang *et al.* (2020).

Competing interests

There are no conflicts of interest or competing interest.

References

1. Caburet, S., Cocquet, J., Vaiman, D. and Veitia, R. A. (2005). [Coding repeats and evolutionary "agility"](#). *Bioessays* 27(6): 581-587.
2. Case, D. A., Belfon, K., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, T. E., *et al.* (2020). Amber20. University of California: San Francisco, CA, USA.
3. Devi, G., Zhou, Y., Zhong, Z., Toh, D. F. K. and Chen, G. (2015). [RNA triplexes: from structural principles to biological and biotech applications](#). *Wiley Interdiscip Rev RNA* 6(1): 111-128.
4. Ellegren, H. (2004). [Microsatellites: simple sequences with complex evolution](#). *Nat Rev Genet* 5(6): 435-445.
5. Felsenfeld, G., Davies, D. R. and Rich, A. (1957). [Formation of a three-stranded polynucleotide molecule](#). *J Am Chem Soc* 79(8): 2023-2024.
6. Grabczyk, E., and Usdin, K. (2000). [The GAA•TTC triplet repeat expanded in Friedreich's ataxia impedes transcription elongation by T7 RNA polymerase in a length and supercoil dependent manner](#). *Nucleic Acids Res* 28(14): 2815-2822.
7. Ivani, I., Dans, P. D., Noy, A., Pérez, A., Faustino, I., Hospital, A., *et al.* (2016). [Parmbsc1: a refined force field for DNA simulations](#). *Nat Methods* 13(1): 55.
8. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. and Klein, M. L. (1983). [Comparison of simple potential functions for simulating liquid water](#). *J Chem Phys* 79(2): 926-935.
9. Joung, I. S. and Cheatham III, T. E. (2008). [Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations](#). *J Phys Chem B* 112(30): 9020-9041.
10. Kaji, E. H. and Leiden, J. M. (2001). [Gene and stem cell therapies](#). *Jama* 285(5): 545-550.
11. McMurray, C. T. (1999). [DNA secondary structure: a common and causative factor for expansion in human disease](#). *Proc Natl Acad Sci U S A* 96(5): 1823-1825.

12. Mirkin, S. M. (2006). [DNA structures, repeat expansions and human hereditary disorders](#). *Curr Opin Struct Biol* 16(3): 351-358.
13. Mirkin, S. M. (2007). [Expandable DNA repeats and human disease](#). *Nature* 447(7147): 932-940.
14. Nelson, M. T., Humphrey, W., Gursoy, A., Dalke, A., Kalé, L. V., Skeel, R. D. and Schulten, K. (1996). [NAMD: a parallel, object-oriented molecular dynamics program](#). *Int J Supercomput Appl High Perfor Comput* 10(4): 251-268.
15. Orr, H. T., and Zoghbi, H. Y. (2007). [Trinucleotide repeat disorders](#). *Annu Rev Neurosci* 30: 575-621.
16. Pan, F., Man, V. H., Roland, C., and Sagui, C. (2017). [Structure and dynamics of DNA and RNA double helices of CAG and GAC trinucleotide repeats](#). *Biophys J* 113(1): 19-36.
17. Pan, F., Zhang, Y., Man, V. H., Roland, C. and Sagui, C. (2018). [E-motif formed by extrahelical cytosine bases in DNA homoduplexes of trinucleotide and hexanucleotide repeats](#). *Nucleic Acids Res* 46(2): 942-955.
18. Pan, F., Man, V. H., Roland, C. and Sagui, C. (2018). [Structure and dynamics of DNA and RNA double helices obtained from the CCG and GGC trinucleotide repeats](#). *J Phys Chem B* 122(16): 4491-4512.
19. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T. E., Lughton, C. A. and Orozco, M. (2007). [Refinement of the AMBER force field for nucleic acids: improving the description of \$\alpha/\gamma\$ conformers](#). *Biophys J* 92(11): 3817-3829.
20. Seidman, M. M. and Glazer, P. M. (2003). [The potential for gene repair via triple helix formation](#). *J Clin Invest* 111(4): 487-494.
21. Šponer, J. and Filip, L. (2006). *Computational Studies of RNA and DNA*. Vol. 2. Springer Science & Business Media.
22. Szewczak, A. A., Ortoleva-Donnelly, L., Ryder, S. P., Moncoeur, E. and Strobel, S. A. (1998). [A minor groove RNA triple helix within the catalytic core of a group I intron](#). *Nat Struct Biol* 5(12): 1037-1042.
23. Subramanian, S., Madgula, V. M., George, R., Mishra, R. K., Pandit, M. W., Kumar, C. S., and Singh, L. (2003). [Triplet repeats in human genome: distribution and their association with genes and other genomic regions](#). *Bioinformatics* 19(5): 549-552.
24. Tóth, G., Gáspári, Z., and Jurka, J. (2000). [Microsatellites in different eukaryotic genomes: survey and analysis](#). *Genome Res* 10(7): 967-981.
25. Visualizer, D. S. (2005). Discovery Studio Visualizer. 2. Accelrys Software Inc.
26. Weiner, S. J., Kollman, P. A., Nguyen, D. T. and Case, D. A. (1986). [An all atom force field for simulations of proteins and nucleic acids](#). *J Comput Chem* 7(2): 230-252.
27. Wells, R. D., Dere, R., Hebert, M. L., Napierala, M. and Son, L. S. (2005). [Advances in mechanisms of genetic instability related to hereditary neurological diseases](#). *Nucleic Acids Res* 33(12): 3785-3798.
28. Wells, R. D. and Ashizawa, T. (Eds.). (2011). [Genetic instabilities and neurological diseases](#). (Vol. 31). Elsevier.

29. Xu, P., Pan, F., Roland, C., Sagui, C. and Weninger, K. (2020). [Dynamics of strand slippage in DNA hairpins formed by CAG repeats: roles of sequence parity and trinucleotide interrupts.](#) *Nucleic Acids Res* 48(5): 2232-2245.
30. Zgarbová, M., Otyepka, M., Šponer, J., Mládek, A., Banáš, P., Cheatham III, T. E., and Jurecka, P. (2011). [Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles.](#) *J Chemical Theory Comput* 7(9): 2886-2902.
31. Zhang, J., Fakharzadeh, A., Pan, F., Roland, C., and Sagui, C. (2020). [Atypical structures of GAA/TTC trinucleotide repeats underlying Friedreich's ataxia: DNA triplexes and RNA/DNA hybrids.](#) *Nucleic Acids Res* 48(17): 9899-9917.
32. Zhang, Y., Roland, C., and Sagui, C. (2017a). [Structure and dynamics of DNA and RNA double helices obtained from the GGGGCC and CCCCGG hexanucleotide repeats that are the hallmark of C9FTD/ALS diseases.](#) *ACS Chem Neurosci* 8(3): 578-591.
33. Zhang, Y., Roland, C., and Sagui, C. (2017b). [Structural and dynamical characterization of DNA and rna quadruplexes obtained from the GGGGCC and GGGCT hexanucleotide repeats associated with c9ftd/als and sca36 diseases.](#) *ACS Chem Neurosci* 9(5): 1104-1117.