E2EDNA: Simulation Protocol for DNA Aptamers with Ligands

Michael Kilgour[1], Tao Liu[1], Brandon D. Walker[2], Pengyu Ren[2], Lena Simine[1,]*

[1]Department of Chemistry, McGill University, 801 Sherbrooke St. W, Montreal, Quebec, H3A 0B8, Canada.

[2]Department of Biomedical Engineering, The University of Texas at Austin, 78712, Texas, USA

[*]Correspondence to: lena.simine@mcgill.ca

We present E2EDNA, a simulation protocol and accompanying code for the molecular biophysics and materials science communities. This protocol is both easy to use and sufficiently efficient to simulate single-stranded (ss)DNA and small analyte systems that are central to cellular processes and nanotechnologies such as DNA aptamer-based sensors. Existing computational tools used for aptamer design focus on cost-effective secondary structure prediction and motif analysis in the large datasets produced by SELEX experiments. As a rule, they do not offer flexibility with respect to the choice of the theoretical engine or direct access to the simulation platform. Practical aptamer optimization often requires higher accuracy predictions for only a small subset of sequences suggested e.g., by SELEX experiments, but in the absence of a streamlined procedure this task is extremely time and expertise intensive. We address this gap by introducing E2EDNA, a computational framework that accepts a DNA sequence in the FASTA format and the structures of the desired ligands, and performs approximate folding followed by a refining step, analyte complexation, and molecular dynamics sampling at the desired level of accuracy. As a case study we simulate a DNA-UTP (uridine triphosphate) complex in water using the state-of-the-art AMOEBA polarizable force field. The code is available at https://github.com/InfluenceFunctional/E2EDNA.

1. Introduction

DNA aptamers are short (~10-100 residues) single-stranded nucleotide (ssDNA) sequences. One popular use for DNA aptamers is as sensors or 'aptasensors' for a wide variety of molecular ligands including antibiotics[1], neurotransmitters[2,3], steroids[4], metals[5], proteins[6], nucleosides, including most famously ATP [7,8], and other small molecules [9–11]. The advantage of aptasensors that makes them particularly attractive is the demonstrated potential for stable and selective sensing in crowded biochemical environments, e.g., in blood or in polluted water. The difficulty lies in designing sequences which reliably and selectively bind a desired target analyte. Promising aptasensor candidates are selected using the Systematic Evolution of Ligands by Exponential Enrichment (SELEX) protocol which iteratively enriches DNA libraries with sequences exhibiting preferential affinity toward a target ligand. The candidate aptasensors identified by SELEX contain sequences that are promising but often far from optimal for real-life applications. The key to improving them lies in identifying the causal relationship between sequence and sensing performance. This relationship depends primarily on the 3D folded structure of the aptamer and the structural rearrangements that may be caused by the binding of a ligand of interest.

Several powerful computational frameworks were initially developed for the study of protein-RNA structure and binding [12], with many basic tools carrying over to the study of DNA aptamers and combining with entirely new approaches. For example, APTANI and APTANI2[13] are two methods for selecting potentially relevant aptamers from SELEX datasets through a sequence-structure analysis. Both tools contain modules to predict specific secondary structures in each selection round and to rank aptamers by motifs embedded in their predicted structures. AEGIS, a platform for aptamer design, combines fully automated SELEX

experiments with computational structural characterization for efficient discovery of new aptasensors for disease targets. Aptamer structural prediction with AEGIS is carried out using a novel deep learning approach: a generative adversarial network trained on structure predictions following the ViennaRNA/Rosetta powered iGEM aptamer analysis protocol [14–16], which guesses the possible secondary structures of a transcripted RNA strand, before converting to DNA and making a final prediction.

Computational analyses of this type are very sensitive to the underlying physical models which predict the folded structure of the aptamer and the response to the presence of a ligand. Our goal in this paper is to present a high-accuracy computational pipeline from sequence to bound aptamer-analyte complex that may be used to inform aptamer design efforts, including automated in-silico design platforms. The capability of our E2EDNA design protocol does not hinge on any particular or assumed model. Where E2EDNA takes input from such models, such as in the prediction of aptamer secondary structure, we directly test their predictions using all-atom, explicit water molecular dynamics simulations, using appropriate high-accuracy force fields.

The paper is structured as follows: In Section 2, we present our modular computational framework. We then demonstrate its use with a case study characterization of a given aptamer structure and its binding to a charged uridine triphosphate (UTP) molecule in water, along with detailed analysis in Section 3. We conclude the paper with a summary and an outlook in Section 4.

2. E2EDNA Protocol

E2EDNA (End-to-End-DNA) allows for prediction of the structure of a given aptamer and its binding affinity for a given target molecule or 'analyte'. Beginning from basic information – the aptamer FASTA sequence, and the structure of the analyte – the E2EDNA protocol incorporates 2D and 3D aptamer structure prediction and evaluation, and analyte binding analysis.

2.1.Pipeline Outline

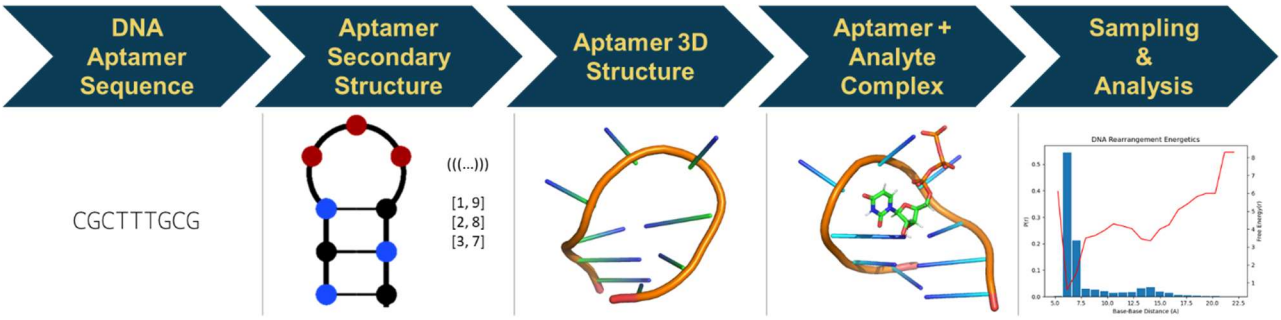

Figure 1: Schematic of the E2EDNA end-to-end aptamer-analyte binding pipeline for UTP complexing with a very simple hairpin.

Figure 1 graphically illustrates the steps of our end-to-end protocol. One begins with the DNA sequence one wishes to test against a given analyte. The pipeline uses a secondary structure prediction tool to predict the most likely base-pairing motifs at the experimentally relevant conditions. A folding tool takes the sequence and list of bases to be paired, and, using strong fictitious forces and an inexpensive force field, pulls the sequence from a fully extended conformation into one which satisfies the given base pairing conditions, which is a rough prediction for the 3D structure.

In general, secondary structure prediction tools cannot always identify with high confidence the complete aptamer structure at a given condition. Further, even an accurate secondary structure prediction necessarily misses key details of the 3D conformation of the aptamer. E2EDNA

evaluates the quality of the proposed configuration or configurations and makes a final prediction via high-accuracy all-atom molecular dynamics simulation, with an example discussed in Section 3.1. Once a consensus 3D structure is discovered, the folded aptamer is complexed with the analyte and sampled via molecular dynamics simulation to determine binding affinity. Analysis of analyte binding to the aptamer, and any influence the analyte may have on aptamer structural reorganization is then carried out by comparison of complexed and free aptamer trajectories.

2.2.Pipeline Components

2.2.1. Secondary Structure Prediction

A crucial difficulty in aptamer-analyte binding analysis is the pre-folding of the aptamer to the correct equilibrium structure. In general, a single DNA sequence may adopt a very wide array of folded structures [17], and brute force approaches such as naïve molecular dynamics search are prohibitively computationally expensive.

Despite decades of work on this problem, due to the complexity of the underlying physical system, secondary structure prediction is not always possible with high confidence with existing tools. Indeed, in sequences with length exceeding a few dozen bases, de novo prediction without experimental input becomes extremely difficult, as the number of possible stable structures may become large, and likely structures may be difficult to discriminate [12]. Common software packages approach this issue by issuing an ensemble of predicted structures with associated probabilities, when appropriate, or by adding base-by-base pairing probabilities to overall structure predictions. One may also solicit predictions from various packages which each employ

different numerical methods such as template-based models, fragment assembly, and minimization of empirical potentials [12].

E2EDNA leverages such purpose-built models to a-priori predict a structure or set of structures a given aptamer has a realistic probability of adopting, before analyzing and refining down to a final structural prediction. In the absence of experimental data, we assess the quality of predicted structures via molecular dynamics simulation with an appropriate force-field, similar to approaches which have been used in RNA structure evaluation[18]. MD sampling then allows us to compare the stability of proposed structures on the nanosecond-microsecond timescale.

In our E2EDNA implementation, we primarily use the 'NUPACK' and 'seqfold' Python packages for secondary structure prediction [19], which implement a range of modern empirical energy models to identify the minimum free energy structure at a given temperature and ionic strength. Both are remarkably fast and easy-to-use, with straightforward I/O and highly useful utility functions. One has only to supply the sequence FASTA string and associated simulation conditions.

2.2.2. Coarse 3D Structure Prediction

Given a proposed aptamer secondary structure in the form of a list of paired bases, we fold it from an extended initial condition into the prescribed secondary structure. We use the program, MacroMoleculeBuilder (MMB) [20], which folds the structure through inexpensive directed simulation, with user-scalable attractive fictitious forces between the paired bases. Using such a tool, we can fold from an arbitrary DNA sequence to a 3D structure which agrees with the predicted secondary structure in a matter of minutes.

Depending on the size and complexity of a given aptamer structure, a customized user-specified annealing schedule may be required. We supply a script with a robust protocol which is efficient even for multi-hairpin structures exceeding 60 bases in length, along with a script which automatically generates MMB input files from secondary structure and sequence information.

Given the strong fictitious forces used to fold the structure in MMB, the output of this simulation is only a rough prediction of the aptamer 3D structure based on the secondary structure prediction, which is analyzed and refined by subsequent MD simulation. Beyond refining the details, extended MD sampling is also used to verify the stability of a predicted secondary structure. For this refinement, we use the MD protocol detailed in Section 2.2.3.

2.2.3. Molecular Dynamics Sampling

In order to accurately capture the nuanced dynamics of the highly charged of the analyte in our test case (see Section 3), we employ the advanced electrostatic and polarizable force field, AMOEBA[21]. We used the Tinker9 software suite running on Compute Canada NVIDIA V100 GPUs for the bulk of our simulations, as well as for simulation setup and post-processing. Detailed simulation parameters, derived from POLTYPE 2, in the form of Tinker keyfiles, are available in our GitHub repository and the Supporting Information (see the Supporting Information also for parameterization details).

2.2.4. Binding Analysis

Once a suitable representative 3D aptamer structure has been isolated through analysis of free aptamer MD trajectories, one must complex the aptamer with the analyte. Our code provides the option for users to initialize the analyte at a preset distance and random direction from the

aptamer center-of-geometry, or to directly place the analyte at a certain coordinate. We analyze aptamer folding and analyte binding via simulation trajectories with and without the analyte and compare the equilibrium probability densities along a series of user-identified reaction coordinates.

In the test case examined in Section 3, we used reaction coordinates which describe base-pairing, large-scale structural rearrangements, and the distance from the analyte to predicted binding sites. Given sufficient sampling time we compute the equilibrium density, P(r), and accompanying free energy profile, F(r) along the reaction coordinate and interrogate the stability of the folded structure and the influence of the analyte on the aptamer complex.

The free energy profile, as a function of the reaction coordinate distance, is given by,

Equation 1

$$F(r) = -kT \ln P(r),$$

where k is the Boltzmann constant, and T, the temperature. The probability distribution P(r) is computed via histogram of time-series analysis of the relevant reaction coordinates r, which are tracked using the Python package, MDAnalysis. Selection of reaction coordinates in this implementation is done on an aptamer-by-aptamer basis.

3. Discussion

Following the method outlined in Section 2, we demonstrate the E2EDNA protocol on a medium-length aptamer in the presence of uridine triphosphate (UTP) in phosphate-buffered saline (PBS). The sequence of the aptamer we have chosen for this test study is:

TATGCATGTGGGCGACGCAGTGCCCGTGGGATTTACTTGCAC

In this section, we will generate a representative 3D structure and evaluate its binding affinity for negatively charged UTP(-4) near bases 39-40.

3.1.Identification and Evaluation of Aptamer Fold

The first step in E2EDNA is identifying the candidate secondary structure, comprising a list of paired bases in the equilibrium structure, to be folded and evaluated in 3D. In Figure 2 we show the NUPACK/seqfold predicted secondary structure and accompanying pair-by-pair NUPACK pairing probabilities, as given by the color. We also show in this figure the user-defined reaction coordinates we follow during simulation trajectories to determine the stability of key features of the structure.

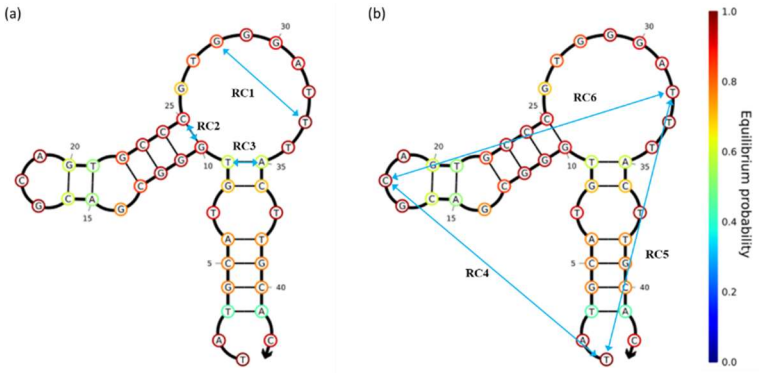

Figure 2: (a) and (b) show the proposed aptamer secondary structure with local (a) and global (b) rearrangement reaction coordinates overlaid. The colors of the bases identify their equilibrium pairing probability according to NUPACK, with redder as more probable and bluer as less probable.

The ‘local’ reaction coordinates, RC’s 1-3, were selected to monitor the openness of the loop (RC1) from base 25-35, as well as adjacent nominally paired areas (RC2 & RC3). The ‘global’ coordinates, RC’s 4-6, were selected to assist in identification of the correct overall 3D structure. By tracking the distances between the three key secondary structural features (the open loop and the two, mostly paired ‘arms’), we can characterize the relations between 3D structural motifs.

To generate initial configurations for MD trajectories, we generate an MMB script from this set of paired bases, which MMB uses to fold a 3D structure which satisfies all the base pairing conditions. While MMB automatically includes certain physical features such as helical stacking interactions between consecutive base pairs, the resulting structure does not generally correspond to the relaxed one found in solution. See an illustration of this difference in the SI.

Evaluation of the secondary structure prediction and identification of representative 3D structure begins with the MMB output structure. In this case, we ran five parallel MMB folds and MD evaluations with a time-step of 2.0 fs and a total of 20 ns of sampling per-run. We exclude the first 2 nanoseconds of each trajectory as equilibration and present in Figure 3 the normalized free energy surfaces for each of the reaction coordinates identified in Figure 2. The temperature was 310K, pH 7.4 and ionic strength 163 mM. Relevant MMB and Tinker control files are provided in the GitHub repository.

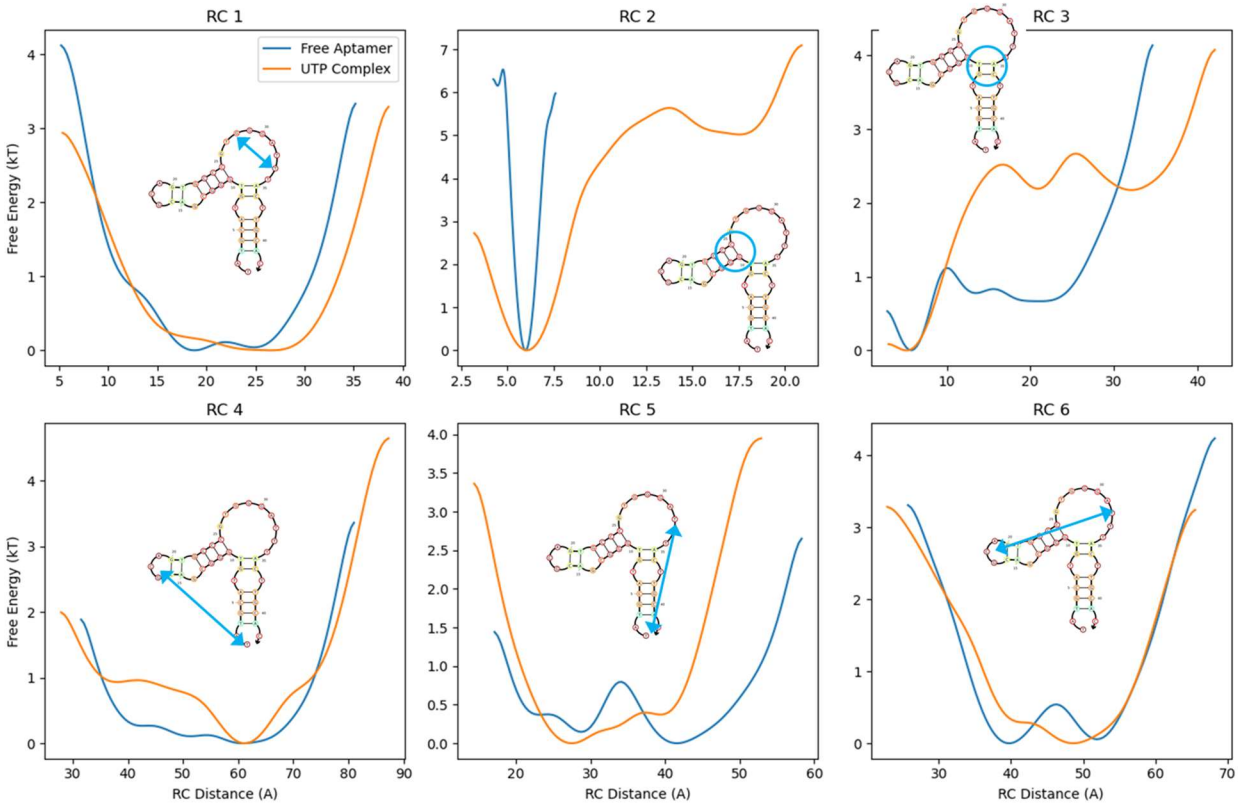

Figure 3: Aptamer 3D conformation reaction coordinates and free energy in units of kT at 310K for two states: the free aptamer, and aptamer with UTP placed in the vicinity of base 39. The surfaces were generated using Equation 1, with the probability density computed by concatenation of the several equilibrated MD trajectories and the free energy minimum set independently to zero for each curve for easy visual comparison. Free energy profiles have been smoothed by a Gaussian kernel to aid readability.

Figure 3, subplots (a-c) allow us to evaluate the correctness of the predicted secondary structure. Following RC 1, we can see that there is no ring-closing; the bases of the loop do not pair to form a hairpin. From RC2 and additional visual scrutiny, we can see that the hairpin from bases 10-25 is very stable throughout every free aptamer trajectory. Finally, in RC3 we see at least two minima, one with the bases 9 and 35 paired, and one very wide basin where they have detached.

Thus, we can basically validate the predicted NUPACK secondary structure, noting that the prediction of a weaker base 9-35 contact was also notably correct.

3.2.Complexation and Analysis of Aptamer-Analyte Binding

Before proceeding with simulation and evaluation of binding of the analyte molecule, we must isolate a representative 3D structure with which to complex it. We accomplish this by analysis of the local (RC 1-3) and global (RC 4-6) reaction coordinates. After identifying the rough minima of each reaction coordinate, we can filter all our trajectories for structures which satisfy most or all of them concurrently, and so retrieve samples representative of the equilibrated 3D structure.

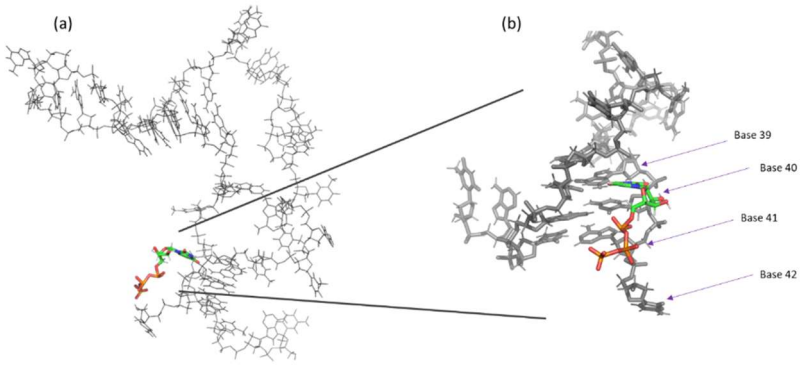

Figure 4: (a) Example initial configuration for analysis of UTP binding to representative aptamer structure, with the UTP adjacent to base 39, with a zoom-in on a rotated view in (b). The UTP is initialized edge-on to base 39, which is well ensconced in a helix.

We chose base 39 as a potential binding site and we run five additional 20 ns trajectories starting from the representative structure, with the UTP initialized near base 39, and watch for any binding or reorganization. See Figure 4(a) for an example initial structure.

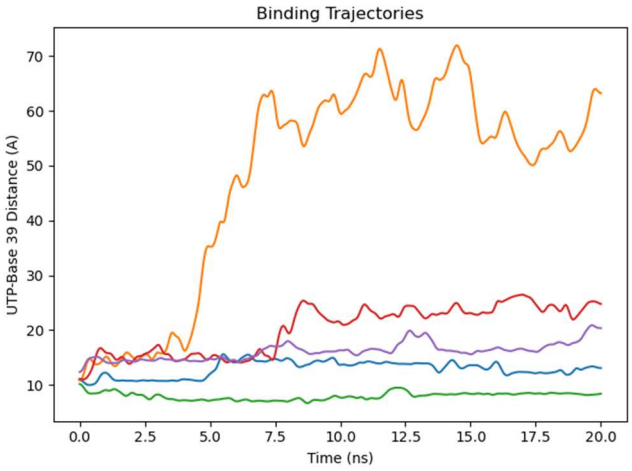

Figure 5: Several DNA-UTP trajectories where we show the distance between the UTP analyte and base 39 as a function of time.

From these simulations we glean the following: 1) From the atomistic details of the binding between UTP and the aptamer, we observe that the binding does not involve π-stacking or hydrogen bonding and overall is rather weak. 2) Figure 4 shows the overall configurational response of the aptamer to the binding event. The shifts in free energy minima in along the given reaction coordinates indicate shifts in the overall configuration of the aptamer in the presence/absence of a ligand. 3) We track the distance between the UTP and the binding site as shown in Figure 6 in order to collect information about the stability of the complex. Furthermore, additional questions may be easily posed to the atomistic simulations and the relevant observables computed from the trajectory data via user-specified reaction coordinates.

4. Summary and Outlook

We have presented and demonstrated E2EDNA, an end-to-end computational pipeline for the high-accuracy characterization and evaluation of DNA aptamer 3D structure, and aptamer-analyte binding under experimental conditions. Following this protocol, a secondary structure for

a given aptamer is proposed, folded in 3D, and evaluated via extended sampling with explicit-solvent, all-atom molecular dynamics simulations. Once a representative aptamer structure is identified, one may complex the aptamer with the analyte of interest and evaluate their binding characteristics using further molecular dynamics simulations.

Our implementation of this protocol requires only the FASTA sequence of the aptamer to be analyzed, and the structure of the analyte molecule to get started. If necessary, automated parameterization of the analyte is straightforward using POLTYPE2. In our implementation, secondary structure prediction, folding and molecular dynamics sampling are all automated, though structural analysis (e.g., selection of important reaction coordinates) and precise analyte placement still must be done manually using supplied input parameters. Molecular dynamics is by far the most computationally intensive component of this pipeline, therefore, for efficient deployment, we recommend using Tinker9 on a GPU platform. With this infrastructure, we have achieved ~5-10 ns/day of sampling on the aptamer complexes similar to the one from Section 3 on NVIDIA V100 GPUs.

Bringing together several disparate computational tools, E2EDNA presents a straightforward approach to aptamer structural and functional evaluation, with all predictions ultimately tested using high-accuracy all-atom explicit solvent molecular dynamics simulation. This protocol provides a necessary building block for future, in-silico studies on aptasensor design and evaluation. We hope it will be useful both in the context of exploring and verifying the binding mechanism for experimental aptasensors, as well as in guiding computational efforts for design of entirely new aptasensor platforms.

DATA AND SOFTWARE AVAILABILITY

Any data generated and analyzed for this study that are not included in this article, the Supporting Information or the GitHub repository are available from the authors upon request. All the software components required to run E2EDNA are available free of charge, with installation instructions in the GitHub README.

ASSOCIATED CONTENT

Supporting Information.

Parameterization details and coarse vs. relaxed 3D structures (PDF)

AMOEBA parameters for UTP(-4) (.doc)

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

AUTHOR INFORMATION

Pengyu Ren is a co-founder of qubit pharmaceuticals.

ACKNOWLEDGMENT

Funding from NSERC Discovery grant RGPIN-2019-24734 and an NSERC PDF for Michael Kilgour are greatly appreciated. Brandon Walker and Pengyu Ren are grateful for support by the National Institutes of Health (Nos. R01GM106137). Computations were made on the supercomputer Beluga, managed by Calcul Quebec (https://www.calculquebec.ca/) and Compute Canada (https://www.computecanada.ca/). The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI).

REFERENCES

[1] A. Mehlhorn, P. Rahimi, Y. Joseph, Aptamer-based biosensors for antibiotic detection: A review, Biosensors. 8 (2018). https://doi.org/10.3390/bios8020054.

[2] K. Sinha, C. Das Mukhopadhyay, Quantitative detection of neurotransmitter using aptamer: From diagnosis to therapeutics, J. Biosci. 45 (2020). https://doi.org/10.1007/s12038-020-0017-x.

[3] H. Hou, Y. Jin, H. Wei, W. Ji, Y. Xue, J. Hu, M. Zhang, Y. Jiang, L. Mao, A Generalizable and Noncovalent Strategy for Interfacing Aptamers with a Microelectrode for the Selective Sensing of Neurotransmitters In Vivo, Angew. Chemie. 132 (2020) 19158–19162. https://doi.org/10.1002/ange.202008284.

[4] S.B. Ebrahimi, D. Samanta, B.E. Partridge, C.D. Kusmierz, H.F. Cheng, A.A. Grigorescu, J.L. Chávez, P.A. Mirau, C.A. Mirkin, Programming Fluorogenic DNA Probes for Rapid Detection of Steroids, Angew. Chemie Int. Ed. (2021) 1–7. https://doi.org/10.1002/anie.202103440.

[5] W. Zhou, R. Saran, J. Liu, Metal Sensing by DNA, Chem. Rev. 117 (2017) 8272–8325. https://doi.org/10.1021/acs.chemrev.7b00063.

[6] R. Kirby, E.J. Cho, B. Gehrke, T. Bayer, Y.S. Park, D.P. Neikirk, J.T. McDevitt, A.D. Ellington, Aptamer-based sensor arrays for the detection and quantitation of proteins, Anal. Chem. 76 (2004) 4066–4075. https://doi.org/10.1021/ac049858n.

[7] Z. Zhang, O. Oni, J. Liu, New insights into a classic aptamer: Binding sites, cooperativity and more sensitive adenosine detection, Nucleic Acids Res. 45 (2017) 7593–7601. https://doi.org/10.1093/nar/gkx517.

[8] D.E. Huizenga, J.W. Szostak, A DNA Aptamer That Binds Adenosine and ATP,

Biochemistry. 34 (1995) 656–665. https://doi.org/10.1021/bi00002a033.

[9] S. Amaya-González, N. de-los-Santos-Álvarez, A.J. Miranda-Ordieres, M.J. Lobo-Castañón, Aptamer-based analysis: A promising alternative for food safety control, Sensors (Switzerland). 13 (2013) 16292–16311. https://doi.org/10.3390/s131216292.

[10] F. Li, Z. Yu, X. Han, R.Y. Lai, Electrochemical aptamer-based sensors for food and water analysis: A review, Anal. Chim. Acta. 1051 (2019) 1–23. https://doi.org/10.1016/j.aca.2018.10.058.

[11] E.J. Cho, J.W. Lee, A.D. Ellington, Applications of aptamers as sensors, Annu. Rev. Anal. Chem. 2 (2009) 241–264. https://doi.org/10.1146/annurev.anchem.1.031207.112851.

[12] I. Tuszynska, D. Matelska, M. Magnus, G. Chojnowski, J.M. Kasprzak, L.P. Kozlowski, S. Dunin-Horkawicz, J.M. Bujnicki, Computational modeling of protein-RNA complex structures, Methods. 65 (2014) 310–319. https://doi.org/10.1016/j.ymeth.2013.09.014.

[13] J. Caroli, M. Forcato, S. Bicciato, APTANI2: Update of aptamer selection through sequence-structure analysis, Bioinformatics. 36 (2020) 2266–2268. https://doi.org/10.1093/bioinformatics/btz897.

[14] A.M. Fernanzes-Arias, Aptafolding : Gan That Folds Aptamers, Universidad Politecnica de Madrid, 2020.

[15] R. Lorenz, S.H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, I.L. Hofacker, ViennaRNA Package 2.0, Algorithms Mol. Biol. 6 (2011). https://doi.org/10.1186/1748-7188-6-26.

[16] C.Y. Cheng, F.C. Chou, R. Das, Modeling complex RNA tertiary folds with Rosetta, 1st ed., Elsevier Inc., 2015. https://doi.org/10.1016/bs.mie.2014.10.051.

[17] M.R. Dunn, R.M. Jimenez, J.C. Chaput, Analysis of aptamer discovery and technology,

Nat. Rev. Chem. 1 (2017). https://doi.org/10.1038/s41570-017-0076.

[18] E. Capriotti, T. Norambuena, M.A. Marti-Renom, F. Melo, All-atom knowledge-based potential for RNA structure prediction and assessment, Bioinformatics. 27 (2011) 1086–1093. https://doi.org/10.1093/bioinformatics/btr093.

[19] J.N. Zadeh, C.D. Steenberg, J.S. Bois, B.R. Wolfe, M.B. Pierce, A.R. Khan, R.M. Dirks, N.A. Pierce, NUPACK: Analysis and design of nucleic acid systems, J. Comput. Chem. 32 (2011) 170–173. https://doi.org/10.1002/jcc.21596.

[20] S.C. Flores, M.A. Sherman, C.M. Bruns, P. Eastman, R.B. Altman, Fast flexible modeling of RNA structure using internal coordinates, IEEE/ACM Trans. Comput. Biol. Bioinforma. 8 (2011) 1247–1257. https://doi.org/10.1109/TCBB.2010.104.

[21] X. Ma, X. Kong, S. Zhang, E. Hovy, MaCow : Masked Convolutional Generative Flow, (2019) 1–10.