2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Automating the Design of Cancer Specific DNA Probes Using Computational Algorithms

Jiacheng Zhang, George Alexandrou, Chris Toumazou and Melpomeni Kalofonou

***Abstract*— This paper introduces a novel Python script which automates the design process of cancer variant-specific DNA probes, based on the amplification method LAMP (Loop-Mediated Isothermal Amplification). With just an input of the DNA sequence and the mutation base location, the script outputs suggestions of two best fitting primer sets for a given target, together with an estimated working efficiency. The script also implements a feature of 'script training', using experimentally-validated primers as a benchmark for primer design optimisation. The proposed script has been tested using the gene sequences of *ESR1* p.E380Q and *ESR1* p.Y537S cancer specific mutations, with the results to closely resemble the experimentally validated primer sets. Creating a rapid LAMP primer design utility allows LAMP to be more easily used as a molecular method for assay development in Lab-on-Chip (LoC) systems to track mutational profiles of variant-specific assays.**

I. INTRODUCTION

Cancer is a global disease threatening lives of numerous developed and developing countries [1]. Despite the rapid advancements of technology in cancer treatment, cancer is still challenging to cure due to its heterogeneity. Heterogeneity occurs by the accumulation of somatic mutations at various rates across tumour cells during tumorigenesis, leading to challenges in treating cancer, as sub-populations of tumour cells that become resistant can repopulate the tumour [2], [3]. Heterogeneity can also cause minimal residual disease (MRD) or malignant cells left after treatments that become dormant in distant sites or in the patients' blood. MRD is hard to clinically determine and can later become reactivated, causing relapse to the patient [4]. The issue of heterogeneity, treatment resistance and relapse potential have sparked the field of liquid biopsies, which utilise circulating tumour DNA (ctDNA) in patients' blood to determine the genetic makeup of tumours at a continuous level [5].

Polymerase Chain Reaction (PCR), has been traditionally used as the molecular method for detection of DNA mutations [6]. Newer generation PCRs, such as digital droplet PCR (ddPCR), divides a sample into many partitions, whereby each partition contains one or two target molecules which undergo individual PCR reactions [7]. Finally sequencing based technologies, whether targeting panels or whole genome analysis, are better for discovering novel mutations but come with a higher analysis cost and time duration. Alternative approaches for cancer diagnostics to sequencing and PCR are considered to be technological platforms and in particular LoC based systems that use chemical sensors, ISFETs (Ion-Sensitive Field-Effect Transistors), as DNA sensing elements [8]–[11], offering high accuracy of detection, better cost effectiveness, portability and accessibility, specifications that are heightened even more now by the COVID-19 pandemic [12]. More recently studies have shown that using LAMP in tandem with ISFET enabled LoCs can facilitate detection of mutational changes which can be applied to liquid-biopsy testing in cancer [13]–[15].

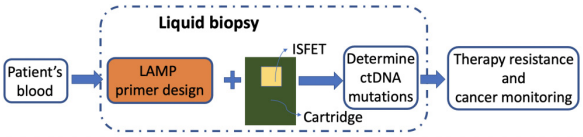

Fig. 1. Overview of how the proposed Python script is used in liquid-biopsy cancer diagnostics.

LAMP has several benefits compared to its PCR counterpart. Firstly, LAMP operates at a constant temperature, eliminating the need for a costly thermal cycler [16]. LAMP is also label-free, offers quicker sample-to-result time and produces higher amplification yield compared to PCR. However, LAMP primers are more complex to design, particularly because LAMP utilises 6 primers recognising various regions of the DNA target.

Current available online tools that can design LAMP primers are not fully automated, e.g. Primer Explorer v5 [17]. Users need to manually adjust the primer locations to create loop primers and Primer Explorer v5 does not provide primer-dimer analysis. Being a web tool, Primer Explorer v5 also forces the user to land on several pages during the primer designing process. Usually, users would have to use both Primer Explorer v5 and Multiple Primer Analyser to check for possible primer dimerisation formations [18]. Not only do these limitations add complexes to the whole primer design experience, but they add the need to switch between tools and perform visual inspections on lengthy DNA sequences which make the process extremely time-consuming and error-prone. Moreover, web tools are restricted to specific browsers and plugins, limiting cross-platform usage.

A novel and updated Python script which automates the process of designing LAMP primers is hereby proposed, based on a previously reported in-silico tool [19]. The script designs target-specific primers with multiple user-adjustable parameters such as primer lengths, melting temperature (Tm)

*This work was supported by the Cancer Research UK (Multidisciplinary Award C54044/A25292), the Leventis Foundation and the Val O'Donoghue scholarship.

All authors are with the Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, SW7 2BT, UK (corresponding author e-mail: m.kalofonou@imperial.ac.uk)

and GC content. It outputs two best sets of LAMP primers to the user together with their estimated efficiencies. Fig. 1 illustrates briefly the workflow and steps followed for designing and testing cancer specific assays, from sample to clinical interpretation, whereby the LAMP primer design step is critical for detecting accurately ctDNA mutations.

II. LAMP PRIMER SPECIFICATION: DESIGN & RANKING

This section introduces the basic principles of the Python script designing LAMP primer sets. The script comprises of two main sections: the primer designing section and the primer set ranking section.

A. Overall Script Architecture

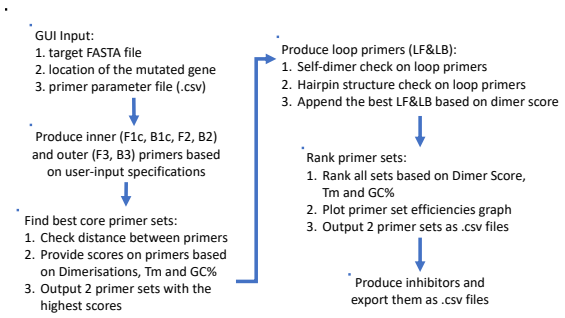

Fig. 2. Workflow of the Python script to design LAMP primers.

As shown in Fig. 2, the script starts by designing the necessary core and loop primers according to user-specifications. All the possible combinations of primers are filtered through numerous checks such as cross-dimer, self-dimer, hairpin loop, Tm and GC content. Each primer set is ranked with a score as an indication of how likely it is to perform well in LAMP based DNA amplification. The script outputs two best sets according to its algorithm and these are then compared with a stored primer set (if available) to estimate the quality of the designed sets.

B. Graphical User Interface

To ease the LAMP primer design process for the user, a simple GUI is presented upon code execution, of which is depicted in Fig. 3. A FASTA format of the target DNA sequence, selection of 'Wild Type' or 'Mutant Type' of the input sequence, the location of the mutated gene and a CSV primer parameter file are necessary information needed for the code to design primers successfully. Once executed, the GUI closes and the results are presented as an output in a new pop-up window.

C. Designing LAMP based Core Primers

The four main LAMP primers are consisted of the Forward Inner Primer (FIP), Backward Inner Primer (BIP), Forward Outer primer (F3) and Backward Outer primer will (B3) as illustrated in Fig. 4. FIP and BIP are double domain primers, whereby they each recognises two of the six regions

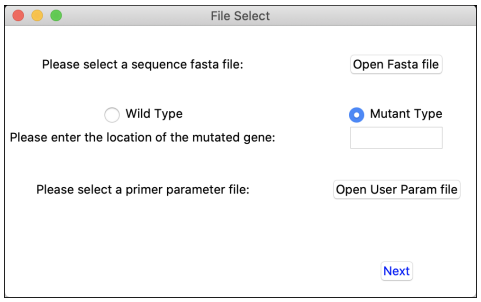

Fig. 3. Graphical user interface prompts user to input the required information.

of the target DNA sequences. FIP recognises the F2 and F1c regions while BIP recognises the B2 and B1c regions of the template sequence ('c' means complementary, it is used to differentiate between complementary DNA sequences).

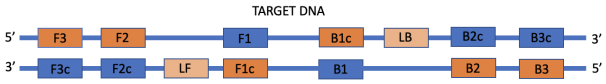

Fig. 4. LAMP primers: primers highlighted in orange are the ones designed by the script. Core primers are highlighted in dark orange and loop primers are highlighted in light orange.

Each primer is created adhering strictly to the user-input parameters. Once the core primers are created, they are put into sets. First, all possible primer set combinations are created using the generated primers. Then, each primer set is undergoing the filtering process which checks for distances, cross-dimerisation and Tm differences. These checks are the same as the later 'Primer Set Ranking' stage but with only four core primers. This first stage of filtering is essential in narrowing down the possible core primer sets into a total of four final sets which eases the loop primer design.

LAMP uses loop primers to accelerate DNA amplification. These primers are generated based on the location of the core primers. It should be noted that these primers should not overlap with each other at any bases.

D. Inhibitor Primers

Recently, it has been reported that two novel primer configurations (inhibitor primers) added in a regular LAMP assay can aid DNA amplification [20]. The main function of these extra primers is to inhibit non-specific LAMP (nspLAMP) reaction while allowing a specific LAMP (spLAMP) reaction to occur uninterruptedly. These are the FB and BB primers, which can be derived from the F1c and B1c primers.

III. LAMP PRIMER SET RANKING

The two best evaluated primer sets are then going through a novel ranking system which outputs a graph, indicating the estimated efficiencies of the designed primer sets. To rank the sets, numerous parameters have been taken into account such as dimer formation, Tm and %GC content.

A. Primer Dimerisations

LAMP primers can bind to bases other than the target DNA sequence, creating undesirable primer dimerisations. For example, cross-dimers are formed when primers of distinct types bind to each other, reducing the number of effective primers which trigger DNA amplification. Other forms of dimers such as self-dimers and hairpin loops are also to be avoided during the design process.

The proposed script suggests the integration of novel functions to evaluate the probability of each primer to undergo dimerisations. This is done by inspecting the primer bases and the ΔG associated with them.

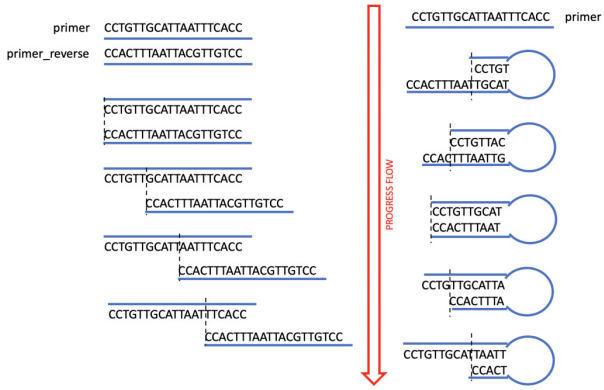

Fig. 5. Workflow explaining the self-dimer (left) and hairpin loop (right) checks mechanisms.

Fig. 5 depicts the mechanisms used in the script to detect the probability of self-dimer and hairpin loop formations of each designed primer. Both are quite similar in the sense that the primer is sliding across itself to check for base matches. As shown in Fig. 6, the more consecutive base matches there are, the more likely the primer is to form either or both of the dimers.

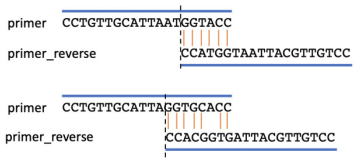

Fig. 6. Primers with more consecutive base matches have a higher tendency to bind to each other, forming unwanted secondary structures.

B. Melting Temperature

Currently, there is no single most accurate way of calculating the melting temperature of the DNA sequences. The Biopython module offers various thermodynamics values such as values from Breslauer et al. and Sugimoto et al. [21], [22]. Since different ways of calculating Tm yield different results, a relative Tm approach is adapted in this script. This way, the discrepancies between the absolute values can be deducted.

According to [17], Tm of F1c/B1c/F3/B3 should be 5°C higher than F2/B2. Therefore, the script calculates the average differences between these primers and evaluate a Tm score based on how far the difference is from the 5°C mark. The further the difference, the lower the Tm score is given to the primer set. This parameter is a strong estimation of how well a set of primers can work together.

C. GC content

Ideally, effective primers have their GC content in the range of 40%-60%. Primers fall into this range would score full points in terms of GC efficiency. The further away the GC content is from that range, the lower the efficiency score. For example, the script gives a score of 50% in GC efficiency for primers with 20-30% or 70-80% GC content. However, this is not the prominent parameter to estimate the quality of primer sets.

D. Distance

All primers must not overlap at any other bases other than the mutated base. This is to achieve allele-specific LAMP primer set creation. If this overlap rule is breached, that particular primer set would have its 'Distance Efficiency' score set to 0, otherwise, it is set to 100. Hence, this is just a binary indication regarding primer-overlapping.

E. Experimentally validated sets used as benchmark

In order to test the quality of the script-proposed primer sets, two experimentally validated primer sets are included in the script to act as benchmarks for primer set ranking. These sets were validated through laboratory testing and had received positive results in DNA amplification. The experimentally validated set goes through the same 'Primer Set Ranking' system as the script-designed set. If the overall efficiency score of the proposed primer set is 20% lower than that of the lab-based one, it is rejected.

IV. RESULTS

The script proposes two sets of LAMP primers for the input target DNA sequence. Table I illustrates an example of the output .csv file for one of the two input gene sequences (*ESR1* p.Y537S). Multiple F1c and B1c primers are included in the output as their efficiencies can only be tested out in the laboratory settings. For each F1c and B1c primer, five primers of varying lengths are generated. However, in Table I, only four B1c primers are proposed, implying one of them was removed during the filtering process. FB1 and BB1 are inhibitor primers derived from F1c and B1c primers respectively. '1' indicated that the starting base of the inhibitor is one base before its original primer.

Fig. 7 shows that at least 80% of the primers generated by the script are more than 50% similar compared to the experimentally-validated ones in terms of base sequences. Even though base location is not the single most prominent aspect of the primer set qualities, it can be used as a guideline of how closely it resembles the already-proven primer set. It should be noted that the experimentally validated primer

TABLE I

EXAMPLE OUTPUT OF THE PRIMER SET FILE.

Primer Type	Sequence	Length
F3	TCGGGTTGGCTCTAAAGTA	
F2	TCTGTGTCTTCCCACCTACA	
F1c	TAGAGGGGCACCACGT	16
F1c	TAGAGGGGCACCACGTTC	
F1c	TAGAGGGGCACCACGTTCTT	
F1c	TAGAGGGGCACCACGTTCTTGC	
F1c	TAGAGGGGCACCACGTTCTTGCAC	
B1c	ATGACCTGCTGCTGG	15
B1c	ATGACCTGCTGCTGGAG	
B1c	ATGACCTGCTGCTGGAGATGC	
B1c	ATGACCTGCTGCTGGAGATGCTG	
B2	ATGCCCCTCCACGGCT	
B3	CAGTGGCCAAGTGGCTT	
LF	CTGTACAGATGCTCCATGCCTTTG	
LB	CCGCCTACATGCGCCCAC	
FB1	ATAGAGGGGCACCACG	16
BB1	TATGACCTGCTGCTG	15

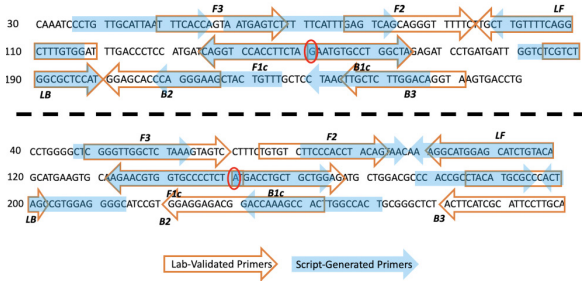

Fig. 7. Visual comparison between script generated primer set with the experimentally validated primer set. Top: primers for target *ESR1* p.E380Q; Bottom: primers for target *ESR1* p.Y537S.

set may not be the most efficient set for LAMP based DNA amplification. That means the script may be able to generate a theoretically better working sets through simulations.

In Fig. 8, 'Set 1' and 'Set 2' are the primer sets designed by the script while 'Test Set' is the experimentally validated LAMP primer set. It can be seen that all sets have similar efficiency scores. This means that the primers in 'Test Set' have similar properties to the script-designed primers. Note that the 'Test Set' has a lower 'Dimerisation Score' (DS Efficiency %) compared to the script-generated sets. Though it is an indication of a higher probability of undergoing dimerisations, this set has been validated to perform spLAMP reaction successfully. Hence, the script-designed 'Set 1' and 'Set 2' primers are also expected to perform spLAMP reaction well.

The 'OVERALL Efficiency (%)' for the script-generated primers is close to the 'Test Set', signaling similar expected working efficiencies between these sets. This parameter is calculated based on the 'DM, TM and GC Efficiencies (%)', providing the overview of the sets' spLAMP reaction effectiveness.

V. CONCLUSIONS

This study has demonstrated a quick and reliable way to design LAMP primers through an updated Python script.

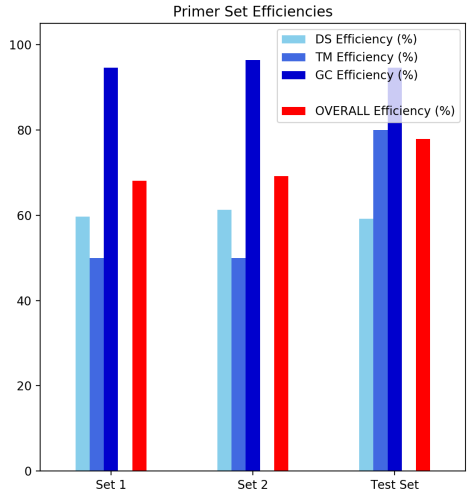

Fig. 8. Example output of the primer set efficiencies for *ESR1* p.E380Q.

New dimerisation-checking functionalities and a novel ranking system have been implemented to allow the integration of the whole design workflow into a single tool. Compared to existing publicly-available tools, this script offers a more complete yet flexible approach in LAMP primer design for mutational detection, which is directly linked to liquid-biopsy based assays. The best designed primer sets are compared with pre-stored experimentally validated sets, providing an insight to the users of the estimated working efficiencies of the script-designed sets. Through simulations with DNA sequences of well-studied cancer genes, *ESR1* p.E380Q and *ESR1* p.Y537S, the primer sets designed by the script deemed to have similar working efficiencies in DNA amplification. Future work can include the expansion of the proposed script to a larger pool of experimentally validated primer sets, to enhance accuracy and reliability of prediction. This encourages the rapid development of LAMP primers which eases the workflow of using LoC platforms for continual cancer profile tracking.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] D. Urbach, M. Lupien, M. R. Karagas, and J. H. Moore, "Cancer heterogeneity: origins and implications for genetic association studies," *Trends in Genetics*, vol. 28, no. 11, pp. 538–543, 2012.

[3] F.-I. D. Dimitrakopoulos, A. G. Antonacopoulou, A. E. Kottorou, S. Maroussi, N. Panagopoulos, I. Koukourikou, C. Scopa, M. Kalofonou, A. Koutras, T. Makatsoris *et al.*, "NF-kB2 genetic variations are significantly associated with non-small cell lung cancer risk and overall survival," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[4] K. Pantel and C. Alix-Panabières, "Liquid biopsy and minimal residual disease—latest advances and implications for cure," *Nature Reviews Clinical Oncology*, vol. 16, no. 7, pp. 409–424, 2019.

[5] M. R. Openshaw, K. Page, D. Fernandez-Garcia, D. Guttery, and J. A. Shaw, "The role of ctDNA detection and the potential of the liquid biopsy for breast cancer monitoring," *Expert review of molecular diagnostics*, vol. 16, no. 7, pp. 751–755, 2016.

[6] A. Schoenfeld, Y. Luqmani, D. Smith, S. O'reilly, S. Shousha, H. Sinnett, and R. Coombes, "Detection of breast cancer micrometastases in axillary lymph nodes by using polymerase chain reaction," *Cancer research*, vol. 54, no. 11, pp. 2986–2990, 1994.

[7] N. Eastley, A. Sommer, B. Ottolini, R. Neumann, J.-L. Luo, R. K. Hastings, T. McCulloch, C. P. Esler, J. A. Shaw, R. U. Ashford *et al.*, "The circulating nucleic acid characteristics of non-metastatic soft tissue sarcoma patients," *International journal of molecular sciences*, vol. 21, no. 12, p. 4483, 2020.

[8] N. Moser, T. S. Lande, C. Toumazou, and P. Georgiou, "ISFETs in CMOS and emergent trends in instrumentation: A review," *IEEE Sensors Journal*, vol. 16, no. 17, pp. 6496–6514, 2016.

[9] M. Kalofonou and C. Toumazou, "An ISFET based analogue ratiometric method for DNA methylation detection," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2014, pp. 1832–1835.

[10] D. Ma, J. Rodriguez-Manzano, S. de Mateo Lopez, M. Kalofonou, P. Georgiou, and C. Toumazou, "Adapting ISFETs for epigenetics: An overview," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 5, pp. 1186–1201, 2018.

[11] M. Kalofonou and C. Toumazou, "A Low Power Sub-μW Chemical Gilbert Cell for ISFET Differential Reaction Monitoring," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 4, pp. 565–574, 2014.

[12] J. Rodriguez-Manzano, K. Malpartida-Cardenas, N. Moser, I. Pennisi, M. Cavuto, L. Miglietta, A. Moniri, R. Penn, G. Satta, P. Randell *et al.*, "A handheld point-of-care system for rapid detection of SARS-CoV-2 in under 20 minutes," *medRxiv*, 2020.

[13] M. Kalofonou, K. Malpartida-Cardenas, G. Alexandrou, J. Rodriguez-Manzano, L.-S. Yu, N. Miscourides, R. Allsopp, K. L. Gleason, K. Goddard, D. Fernandez-Garcia *et al.*, "A novel hotspot specific isothermal amplification method for detection of the common PIK3CA p. H1047R breast cancer mutation," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020.

[14] G. Alexandrou, N. Moser, J. Rodriguez-Manzano, P. Georgiou, J. Shaw, C. Coombes, C. Toumazou, and M. Kalofonou, "Detection of breast cancer ESR1 p. E380Q mutation on an ISFET lab-on-chip platform," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.

[15] G. Alexandrou, N. Moser, K.-T. Mantikas, J. Rodriguez-Manzano, S. Ali, R. C. Coombes, J. Shaw, P. Georgiou, C. Toumazou, and M. Kalofonou, "Detection of multiple breast cancer ESR1 mutations on an ISFET based lab-on-chip platform," *IEEE Transactions on Biomedical Circuits and Systems*, 2021.

[16] N. Tomita, Y. Mori, H. Kanda, and T. Notomi, "Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products," *Nature protocols*, vol. 3, no. 5, pp. 877–882, 2008.

[17] T. Notomi, H. Okayama, H. Masubuchi, T. Yonekawa, K. Watanabe, N. Amino, and T. Hase, "Loop-mediated isothermal amplification of dna," *Nucleic acids research*, vol. 28, no. 12, pp. e63–e63, 2000.

[18] "Multiple Primer Analyzer — Thermo Fisher Scientific-UK."

[19] G. Alexandrou, J. Rodriguez-Manzano, K. Malpartida-Cardenas, P. Georgiou, C. Toumazou, and M. Kalofonou, "In-silico automated allele-specific primer design for loop-mediated isothermal amplification," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.

[20] K. Malpartida-Cardenas, J. Rodriguez-Manzano, L.-S. Yu, M. J. Delves, C. Nguon, K. Chotivanich, J. Baum, and P. Georgiou, "Allele-specific isothermal amplification method using unmodified self-stabilizing competitive primers," *Analytical chemistry*, vol. 90, no. 20, pp. 11 972–11 980, 2018.

[21] K. J. Breslauer, R. Frank, H. Blöcker, and L. A. Marky, "Predicting DNA duplex stability from the base sequence," *Proceedings of the National Academy of Sciences*, vol. 83, no. 11, pp. 3746–3750, 1986.

[22] N. Sugimoto, S.-i. Nakano, M. Yoneyama, and K.-i. Honda, "Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes," *Nucleic acids research*, vol. 24, no. 22, pp. 4501–4505, 1996.