Integration of Computer Science into the Creation of DNA Encoded Libraries

Tom Zhang and Eric He

Uploaded January 30, 2019

Abstract

Using computer programming techniques to assist in the field of drug discovery, we have developed a program for the batch generation of DNA sequences that can be used in DNA-encoded library (DEL) screening. Using the C language, a computer program was created, which is filtered by setting several conditions, and the generated coding sequence is filtered by the condition and output, and finally the coding sequence whose parameters meet the requirements is obtained. These sequences can then be used as DNA tags in the DEL as parts of the small molecule compounds.

Background

In 1992, Brenner and Lerner first proposed the idea of DNA-encoded library (DEL) and laid out basic theory for the process. The DNA-encoding compound uses the DNA sequence as a tag for recording the structural information of the compound. In the compound library, each compound molecule is linked to a DNA tag of a unique sequence to record each compound in the compound library. Although their ideas are now somewhat primitive, the use of DNA tags to encode information is a key breakthrough and innovation point [1]. Early technologies were not recognized by the pharmaceutical industry due to the high cost and low depth of sequencing. However, DEL technology has made great progress after the next-generation sequencing technology has greatly reduced the cost of sequencing and the sequencing depth has been greatly improved.

In recent years, DEL has proven to be an effective way to discover new ligands and understand biological systems [2]. DEL consists of two parts of a chemical small molecule and a tag DNA sequence molecule covalently bound. [3]. By affinity screening, the DEL library will be incubated with a target protein and washed [4]. When some DEL molecules with no affinity are washed away, those molecules that have affinity for the target protein will be screened and enriched, and then decoded by high-throughput sequencing to retrieve the corresponding compound. Thus, DNA coding is very important for the entire DEL technology system. The control of the quality of DNA coding sequences is particularly important. Once the coding is mutated or misinterpreted (such as a hairpin structure), the label will not be effectively connected to the small molecule, and it will not be recognized, which will affect the drug development process.

There are two kinds of coding techniques for constructing DEL, which are used in drug development: DNA-encoded compound library and DNA-directed compound library. [5] The former has a simple coding method, is easy to be used in industry and research and development, and has been mentioned in patents and literature. The principles of design exist, but so far there is no special program available; the latter is more complicated and has been introduced in detail in previous reports [6]. With the increasing use of DEL technology in the academic and industrial users (Figure

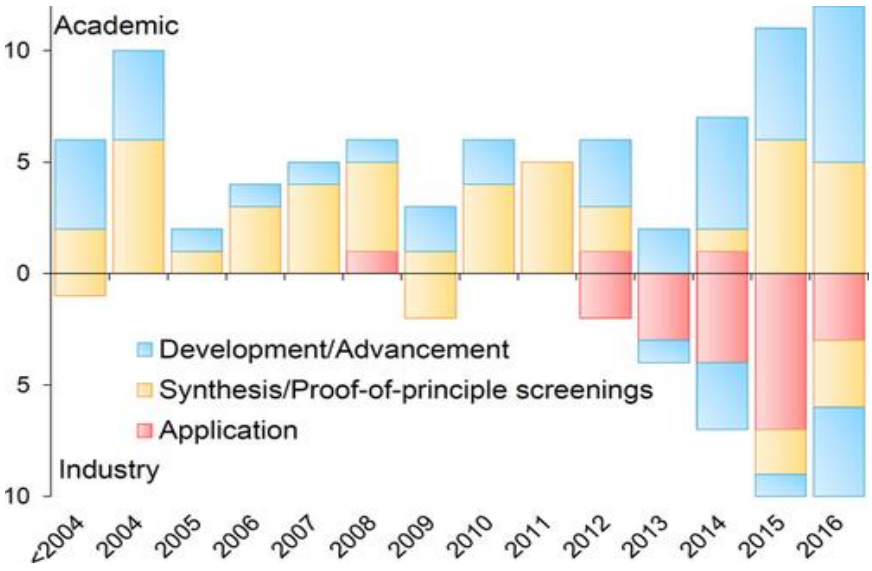

Figure 1: Publications related to DECL from academia and industry [7]

1), it is urgent to be able to develop a corresponding program to solve the process of coding design.

Presently, as more and more DEL technology is widely recognized by academics and industry, the number of corresponding published articles has also increased year by year. In particular, the pharmaceutical industry has been very active in the past two years. The number of articles published each year is equal to that of the academic community. For this reason, more examples of DEL specific applications have been reported, which further promoted the development of this field.

In the DNA-encoded compound library technology (shown in Figure 2), which is widely used, nucleic acids are used to encode the building blocks used in each cycle. The strategy of combining chemical Split-&-Pool can be used in a shorter period. A large number of compound libraries are produced within the cycle.

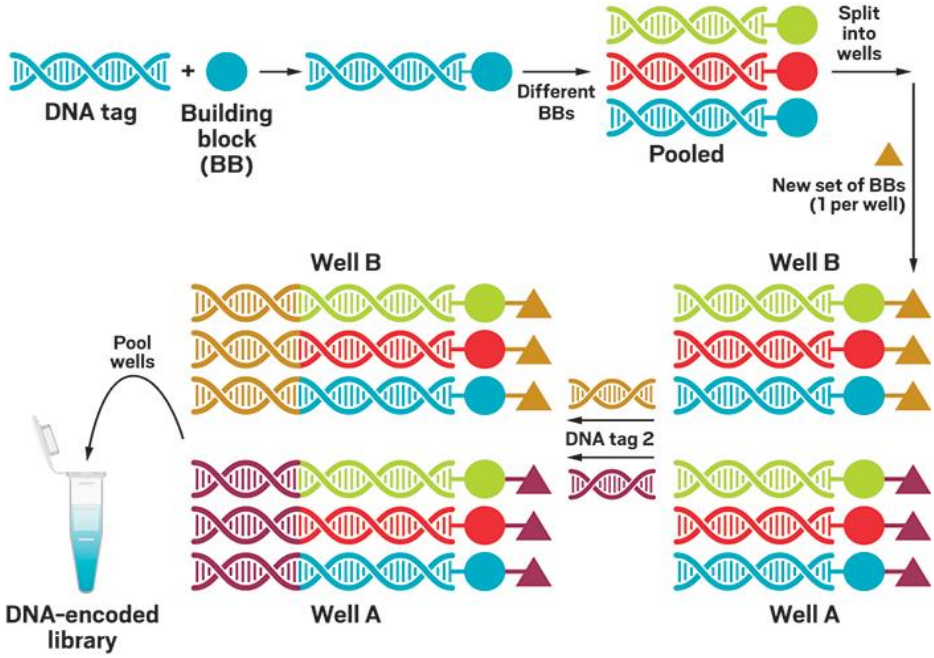

Figure 2: C&EN, Cover Story, How DNA-encoded libraries are revolutionizing drug discovery With the bar-coding technology, drugmakers leverage the chemistry of large numbers, Volume 95 Issue 25 | pp. 28- 33 Issue Date: June 19, 2017)

In DEL technology, the diversity of compound libraries is limited by the diversity of DNA coding. Considering the capacity of two to three rounds of compound libraries is between 10 ^ 3-10 ^ 6 and even as high as 10 ^ 15 [5], We fixed the length of the coding sequence to 9 bases, covering the size of the compound library in an order of magnitude. The target structure of a single DNA double strand is as follows: in the middle is a coding region of 9 base pairs in length, and the 3' end of each single-stranded DNA is a preset 2 base length sticky linker sequence. The purpose of the program we designed was to generate sequences in batches, and to ensure that each pair of DNA codes generated had good stability in the calculation parameters, eliminating the sequence of consecutive single/double base repeats, hairpin structures, and sliding sequences. It can also meet the requirements of GC content and melting temperature as well as under the premise of forming a GC clamp, and ensure the validity and integrity of the sequence as much as possible.

Methods and Discussion

The factor determining whether DNA coding can be correctly linked is the number and order of four bases A, T, C, and G in two single-stranded DNA. The flow of our program is roughly this:

a) The user pre-sets the base sequence of the 2 bases before and after (the length is 2)

b) The user inputs the n pairs of sequences that they want

c) The program will randomly create a new 9 nucleotide DNA encoding and simultaneously generate its reverse complementary sequence.

d) The program will determine if the generated sequence meets the filter criteria

e) n pairs of DNA sequences obtained by screening conditions are returned to the user as a result

The program will use multiple factors to determine if the DNA strands are properly connected and the stability of the connection. Assuming that a DNA strand can be erroneously linked and at the same time has some stability, this deformed DNA strand will be rejected. The goal of the program is to determine if the two DNA strands can be correctly paired, and if paired, then calculate the stability after pairing. This can be applied to a variety of fields including pharmaceutical and DNA computing.

Specific procedures:

This program takes into account the following factors:

1. Single Base Repeat
2. Double Base Repeat
3. GC Content
4. GC Clamp
5. Sliding Sequences
6. Melting Temperature
7. Hairpin Structures

Among them, it is necessary to avoid the generation of sequences with 1, 2, 5, and 7 features. At the same time, it is necessary to meet the requirements for the production of GC clamps, the GC content requirements, and the melting temperature requirements.

1. Single/Double Base Repeat

A single base nucleotide repeat of four or more identical nucleotides, such as "aaaa", increases the risk of guiding errors in the PCR. Three or more double base nucleotide repeats of the same two nucleotides, such as "acacac", also increase the risk of guiding errors [8]. We

constructed a new searching program for the purpose of testing 1000 DNA sequences generated by the original program for generating the DEL library. The search result was 0. Therefore, it was demonstrated that the original DEL library generation program did not generate a single/double base DNA sequence.

```
      for (int x = 0; x < 2000; x++) {
      //the above is the first loop
    for (int y = 0; y < 8; y++) {
            //the above is the second loop
       if (message[x][y] == 'a' || message[x][y] == 't' || message[x][y] == 'c' || message[x][y] == 'g') {
          if (message[x][y] == message[x][y + 1] && message[x][y] == message[x][y + 2] && message[x][y] == message[x][y +
3]) {
            flag = 1;
            printf("%s %d", message[x], x)
            break;
          }
           }
       }
    if (flag == 1) {
       goto output;
    }
}
 for (int a = 0; a < 2000; a++) {
 //the above is the third loop
    for (int b = 0; b < 6; b++) {
            //the above is the fourth loop
       if (message[a][b] == 'a' || message[a][b] == 't' || message[a][b] == 'c' || message[a][b] == 'g') {
          if (message[a][b] == message[a][b + 2] && message[a][b] == message[a][b + 4]) {
            if (message[a][b + 1] == message[a][b + 3] && message[a][b + 1] == message[a][b + 5]) {
               flag = 2;
               printf("%s %d", message[a], a);
               break;
```

```
            } else {
                flag = 0;
            }
        }
            }
    }
    if (flag == 2) {
        break;
    }
}
 output:
    printf("%d\n", flag);
```

In this code, “message” is a two-dimensional character array that contains the individual marking number of sequences in the first dimension and then a nucleotide of A, T, C, or G in the second dimension. In total, there are 2000 individual single-strands of DNA containing 11 nucleotides each. Due to the way that arrays are structured, the first member of an array is counted as being in the 0^{th} position instead of the 1^{st}. Adapting to this, the first part of the code, which accounts for the single-base runs, is a “for-loop” that increases from 0–1999 with an increment-per-loop of 1. This first loop is used to cycle through all 1000 of the duplexes. The second loop searches through each individual single-stranded DNA’s nucleotides for single-base repeats, from the 1^{st} actual nucleotide to the 8^{th} (the 0^{th} array member to the 7^{th}). If the program finds four of the same consecutive nucleotides at any point, the Boolean “flag” will be made true. The program will then exit the loop and output “1” as to identify an instance of a single-base run. Next are the 3^{rd} and 4^{th} loops, these loops are used to spot instances of double-base nucleotide repeats, and will return a “2” in the event that a double-base repeat is found. As expected, the program returned “0” every time.

2. GC content

Another important parameter for successful PCR is the GC content of the DNA duplex. The GC content is the percentage of G or C in the total duplex (excluding the 2 sticky ends), and the reason for its utility is that the stability provided by the GC base pair chemical bond is greater than the stability of AT base pairs. Based on the average GC content of the 2876 template sequences used in an experiment and the recommended GC content, the recommended GC content in one sequence was 49.0 ± 11.3% [9].

However, please note that a more important factor in the design sequence is the relationship of GC and Tm between the template and sequence [10]. The program will test the GC content of the 1000 sequences generated by our program by examining how many G or C bases among the 9 nucleotide pairs in the center. There were 473 sequences in the sampler 1000 sequence with a GC content of 4/9 (44.44%) and 527 sequences with a GC content of 5/9 (55.56%) (Fig. 2). The average GC content in these DNA sequences was 50.3%, which was completely within the recommended range.

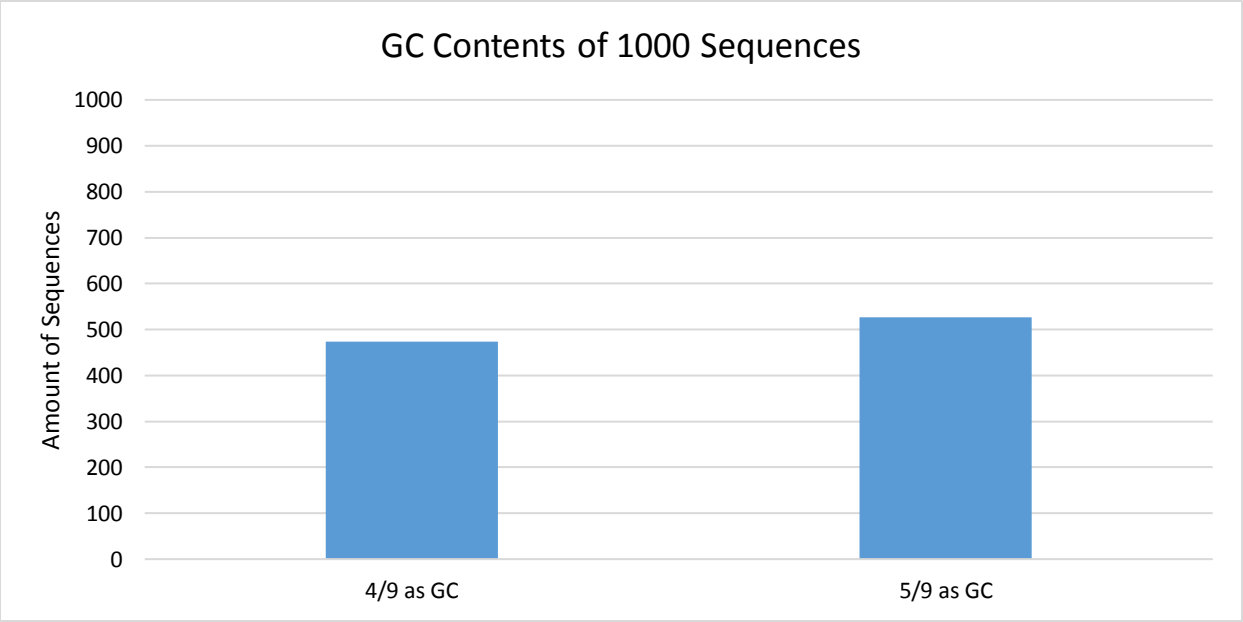

Figure 3: Number of 1000 product sequences and their GC contents

3. GC clamp

The GC-clamp is 5 bases at the 3' end of the DNA sequence in which there should be 2-3 GC bases. Since the GC has two hydrogen bonds with three hydrogen bonds instead of the AT pair, the presence of GC bases promotes a more stable structure. However, single nucleotide repeats of G or C of length 3 should be avoided, as single-base repeats increase the chance of mispriming [8]. Also, avoid exceeding having the total number of GCs be more than 3 because the GC content at the end will exceed the recommended range mentioned earlier.

The following program is to build a generator with GC clamp:

```
int clamp = 0;
    for (int cl = 6; cl <= 10; cl++) {
        if (DNA[cl] == 3 || DNA[cl] == 4) {
            clamp++;
        }
    }
    if (clamp < 2 || clamp >= 4) {
        fitness--;
    } else {
        fitness++;
    }
    for (int cla = 6; cla < 9; cla++) {
        if (dna2[cla] == 'c' && dna2[cla + 1] == 'c' && dna2[cla + 2] == 'c') {
            fitness--;
        } else if (dna2[cla] == 'g' && dna2[cla + 1] == 'g' && dna2[cla + 2] == 'g') {
            fitness--;
        }
    }
```

4. Sliding Sequences

Use a unified view of nearest neighbor thermodynamics and parameters from seven laboratories, we have integrated Gibb's Free Energy (G) into our sequence design [11]. The meaning of G is basically all the useful energy inside the system, $G^o = H^o - TS^o$. The G of the double helix DNA is the energy required to destroy the structure, and it is also the DNA double strand's free energy after sliding minus the original DNA double strand free energy. We considered the slipping of each individual sequence by comparing the original sequence to their most stable post-sliding state, resulting in the ΔG difference of 1000 DNA sequences (Figure 4). As shown, the values in the scatter plot are all greater than 0, indicating that the Gibbs free energy of the original sequence produced is always more stable than the sliding sequence.

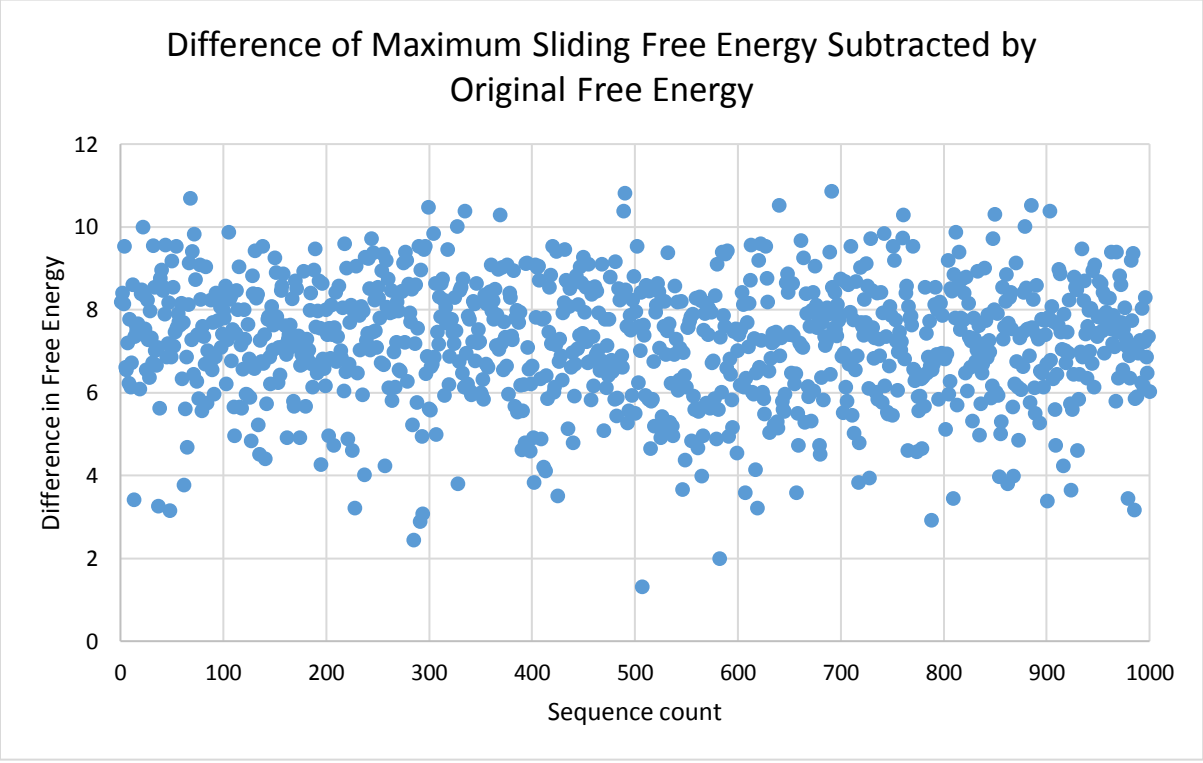

Figure 4: Shows that the resulting sliding position of the sequence does not have a more stable structure than the original sequence produced.

5. Melting temperature (T_m)

The melting temperature (T_m) of the chain is a crucial aspect to consider when generating a sequence. Experiments have validated the effect of mismatched base pairs on the melting temperature of oligomeric double-stranded DNA. For example, when the experimental group was paired with am-3 DNA with mismatched base pairs, the average difference in T_m of the am-3 group was 10 degrees Celsius lower than that of the control group [12]. T_m represents the stability of the chain. The ideal temperature range is 52-58 ° C, wherein all of the sequences' T_m's produced by our program is within (Figure 5). For a Tm formula with a sequence of less than 14 nucleotides, the formula is $\boldsymbol{T_m = (wA + xT) * 2 + (yG + zC) * 4}$ where w, x, y and z are the number of A, T, G, and C [13].

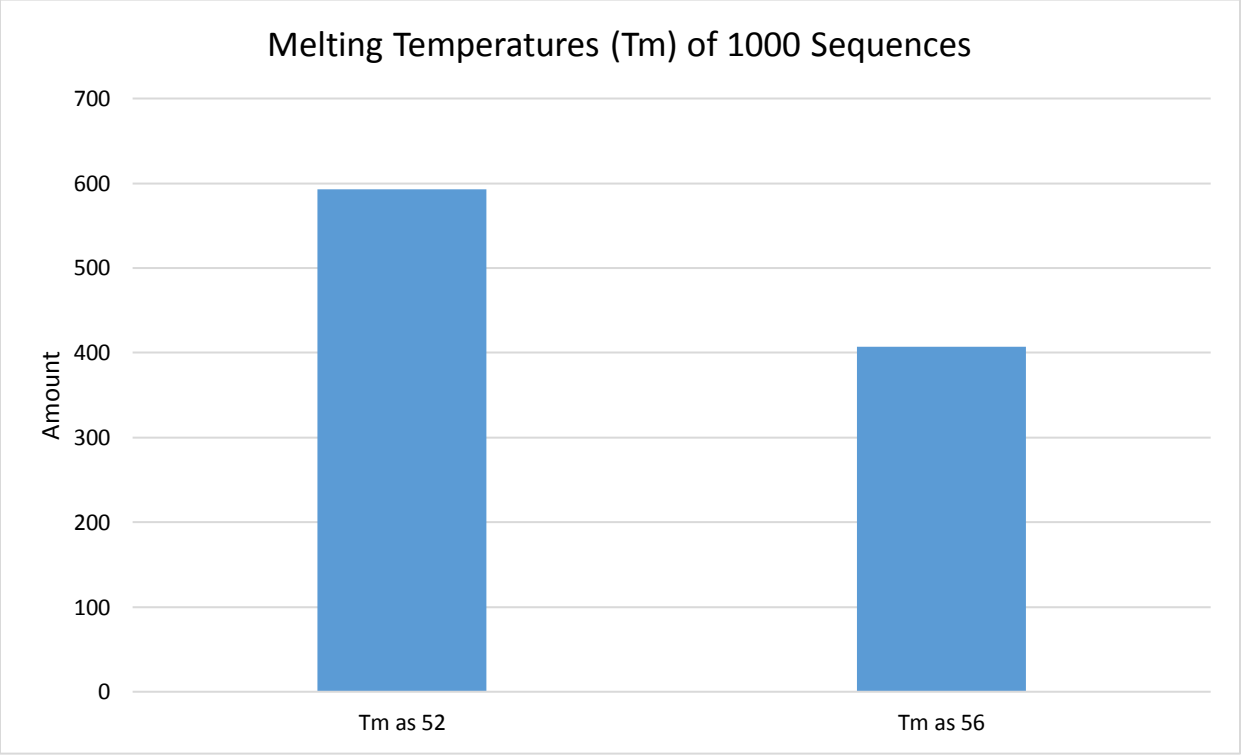

Figure 5: Shows that melting temperatures of all sequences are within recommended range

Below is the code for calculating melting temperature：

```
float at = 0, tm = 0, newgc = 0;
   for (int aa = 2; aa < 11; aa++) {
      if (DNA[aa] == 2 || DNA[aa] == 5) {
         at++;
      } else if (DNA[aa] == 3 || DNA[aa] == 4) {
         newgc++;
      }
   }
   for (int rep = 0; rep < 9; rep++) {
      if (DNA2[rep] == 2 || DNA2[rep] == 5) {
         at++;
      } else if (DNA2[rep] == 3 || DNA2[rep] == 4) {
         newgc++;
      }
   }
   tm = (at * 2) + (newgc * 4);
   if (tm <= 58 && tm >= 52) {
      fitness++;
   } else {
      fitness--;
   }
```

6. Hairpin Structures

Hairpins are a type of unwanted secondary structure. This structure may occur if there are at least 3 paired nucleotides on either side of a single strand. To bypass this hairpin problem, the program does not use ΔG to determine if a hairpin is likely to affect the efficiency of the PCR. Instead, the program eliminates the possibility of a hairpin by zero by filtering out any pair of three pairs of consecutive duplexes on any one chain. This will subtract ΔG and T_m from the equation, so there is no need to use complex methods in determining the validity of the hairpin structure.

The following is the program to avoid issuing a hairpin structure:

```
for (int y = 0; dna[y + 5] != '\0'; y++) {
   flag = 0;
   for (int j = y + 5; dna[j] != '\0'; j++) {
      flag = 0;
      if (DNA[y] + DNA[j] == 7) {
         flag++;
         for (int i = 1; i < 3; i++) {
               if (DNA[y + i] + DNA[j - i] == 7) {
                        flag++;
               } else{
               break;
               }
          }
     }
      if (flag == 3) {
         flag3 = 1;
         break;
      } else if (flag < 3) {
         flag3 = 0;
      }
   }
   if (flag3 == 1) {
      fitness--;
   }
}
```

In this code, “dna” is the name of a character array containing 11 members, those being the A, T, G, and C nucleotides. “DNA” is the corresponding array in integer values. Nucleotides were given values so that when integer values of matching AT or CG pairs were added together, they would have a sum of 7, as to identify them. If the farthest ends of the hairpin are pairs of

A & T or C & G then this potential hairpin has a possibility of forming. In the next loop, the interior 2 pairs of the hairpin are checked for their correspondence. If the integer "flag" is 3, it shows that all 3 pairs in the hairpin are matching. As a result, the duplex must be discarded, at which point the program is quick to reloop to the beginning to generate a new random sequence. A similar code exists for the second single-strand DNA in the duplex, with minor changes in variable names, but has the same effect of filtering out duplexes with hairpins.

Conclusion

The DNA double-stranded structure of 2+9/9+2 (2 nucleotide sticky end + 9 central nucleotides) structure can be efficiently obtained by our program. At the same time, the 1000 DNA double-stranded structures generated by the program meet the requirements for DNA tag design. Single base/double base repetition is avoided, formation of hairpin structure is avoided, and two-way sliding is avoided. It can meet the requirements of GC content, melting temperature, and successfully generate a GC clamp for every sequences.

Our research provides a technical basis for the future use of programming technology to construct an efficient, diverse, and reasonable DNA double-strand structure in the DEL library, and provides ideas for the efficiency of future drug screening.

References

[1]. Brenner, S. and R.A. Lerner, Encoded combinatorial chemistry. Proc Natl Acad Sci U S A, 1992. 89(12): p. 5381-3.

[2]. Salamon, H., et al., Chemical Biology Probes from Advanced DNA-encoded Libraries. ACS Chem Biol, 2016. 11(2): p. 296-307.

[3]. Litovchick, A., et al., Encoded Library Synthesis Using Chemical Ligation and the Discovery of sEH Inhibitors from a 334-Million Member Library. Sci Rep, 2015. 5: p. 10916.

[4]. Shi, B., et al., Recent advances on the encoding and selection methods of DNA-encoded chemical library. Bioorg Med Chem Lett, 2017. 27(3): p. 361-369.

[5]. Kleiner, R.E., C.E. Dumelin and D.R. Liu, Small-molecule discovery from DNA-encoded chemical libraries. Chem Soc Rev, 2011. 40(12): p. 5707-17.

[6]. Usanov, D.L., et al., Second-generation DNA-templated macrocycle libraries for the discovery of bioactive small molecules. Nat Chem, 2018. 10(7): p. 704-714.

[7]. Yuen, L.H. and R.M. Franzini, Achievements, Challenges, and Opportunities in DNA-Encoded Library Research: An Academic Point of View. Chembiochem, 2017. 18(9): p. 829-836.
[8]. Reichova, N. and J. Kypr, Expansion during PCR of short single-stranded DNA fragments carrying nonselfcomplementary dinucleotide or trinucleotide repeats. Mol Biol Rep, 2003. 30(3): p. 155-63.
[9]. Yakovchuk, P., E. Protozanova and M.D. Frank-Kamenetskii, Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res, 2006. 34(2): p. 564-74.
[10]. Benita, Y., et al., Regionalized GC content of template DNA as a predictor of PCR success. Nucleic Acids Res, 2003. 31(16): p. e99.
[11]. SantaLucia, J.J., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci U S A, 1998. 95(4): p. 1460-5.
[12]. Wallace, R.B., et al., Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. Nucleic Acids Res, 1979. 6(11): p. 3543-57.
[13]. MARMUR, J. and P. DOTY, Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. J Mol Biol, 1962. 5: p. 109-18.