Long DNA Strands Synthesis Optimizing

Robert Nowak[1] and Marcin Romaniuk[1]

[1] Warsaw University of Technology, Institute of Electronic Systems, Warsaw, Poland,
email: r.m.nowak@elka.pw.edu.pl

Abstract. The method described in this paper helps to syntheses DNA (deoxyribonucleic acid) molecules with length about 1000 bp, using typical techniques enable to create strands of length up to 70 bp. The given DNA strand is divided into smaller fragments, and next these fragments are connected by proposed protocol in genetic laboratory. The evolutionary algorithm is used to find the optimal solution.

The freely accessible application called longdna, based on presented ideas was implemented and tested on simulated and real data.

1 Introduction

A double helix of DNA is formed from two separate DNA strands, connected together (head-to-toe) by hydrogen bonds. Each strand has a natural orientation denoted (according to chemical convention) as 5' and 3' end, therefore DNA strands may be viewed as sequence of nucleotides or bases, in which nucleotides namely: adenine, cytosine, guanine, and thymine abbreviated to `A`, `C`, `G`, and `T` respectively. The hydrogen bond is selective, adenine bonds only with thymine, so `A` is complementary to `T`, and guanine bonds with cytosine, pairs `(G,C)` are complementary.

Hybridization is the process of combining complementary single-stranded nucleic acids into a single molecule, double helix of DNA, using hydrogen bonds between bases. This process may be reverted in denaturation or melting reaction, by heating or by changing pH of solution. DNA melting (or denaturation) temperature, denoted T_m, is understood as the temperature at which a DNA double helix dissociates into single strands. The T_m is defined as temperature at which half of the DNA molecules in mixture are in single stranded state.

The reactions performed in genetic laboratories use enzymes (proteins), which catalyse particular chemical process on DNA. In this work the enzyme called DNA ligase is widely used. This protein links together DNA strands by creating covalent bonds between 3' and 5' ends of contiguous molecules. The opposite effect, i.e. breaking or cutting DNA strand, is catalysed by restrictazes. The DNA polymerase is an enzyme that is able to produce the new strand against the existing DNA template using base-pairing interactions. The free nucleotides are added to 3' end of newly-forming strand, and the newly-polymerized molecule is complementary to the template. From computer scientist's point of view furher information about DNA and genetic operations can be found in [1, 6].

The DNA strand of given sequence could be chemically synthesised by adding repeatedly monomers with protected 5' end. Despite the yield of this step is about 98%, the length of created strand is limited [10]. Necessity of synthesise long DNA molecules, e.g. in bacterial production of substances like human insulin, are needed to develop techniques to obtain such molecules from smaller parts.

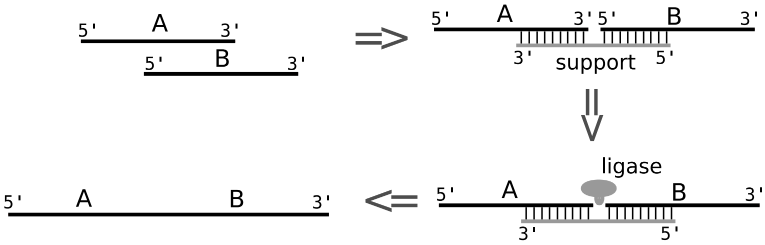

Figure 1. Typical method [2] to create longer DNA molecules using ligation. The long sequence is divided into fragments which are synthesised separately, then the support DNA strands complementary to contiguous fragments are added and the hybridisation is performed. Next, the fragments are connected by using DNA ligase. Finally denaturation returns the designed long DNA strand. If the double stranded molecule is necessary, the DNA polymerase is used.

The commonly technique, used in genetic laboratories, to obtain long DNA molecules, recently patented [2], is to connect many shorter DNA strands by using hybridisation and ligation as depicted in fig. 1. The main disadvantage of this technique is long period and high cost, because in single reaction (showed in fig. 1) only two fragments are combined, thus if the final molecule is build from the n fragments (of length between 15 and 70 bp), the $n - 1$ reactions must be performed.

Presented algorithm considers possibility of performing many various connections in the same ligation, i.e. it optimizes the length of fragments to make possible perform many cycles of typical algorithm concurrently, thus it is less time-consuming and costly. Proposed protocol could replace currently used techniques.

2 Algorithm

In presented approach the evolutionary algorithm finds the optimal protocol of long DNA synthesis. The solution includes the length and sequences used DNA strands, i.e. fragments and supporting mulecules, as well as the way of performing the reaction (in one or more steps). For example, if the input sequence is divided into four fragments A, B, C, D and synthesis involves one ligation (one step), the protocol is presented in fig. 2. All fragments are joined in one ligation, thus the protocol is faster than typical method.

Algorithm to find protocol

The individual in EA represents the sequences of molecules used in creation: the sequences of input molecule fragments and the sequences of supporting molecules. The fragments are generated from input sequence, by cutting it into parts of proper length (between 15 and 70 nucleotides), furthermore the fragments length can be various. The supports are complementary to two contiguous fragments (fig. 2).

The protocol of synthesis, indicating fitness, is calculated for given set of molecules (fragments and supports) by algorithm depicted in fig. 3. This algorithm found the

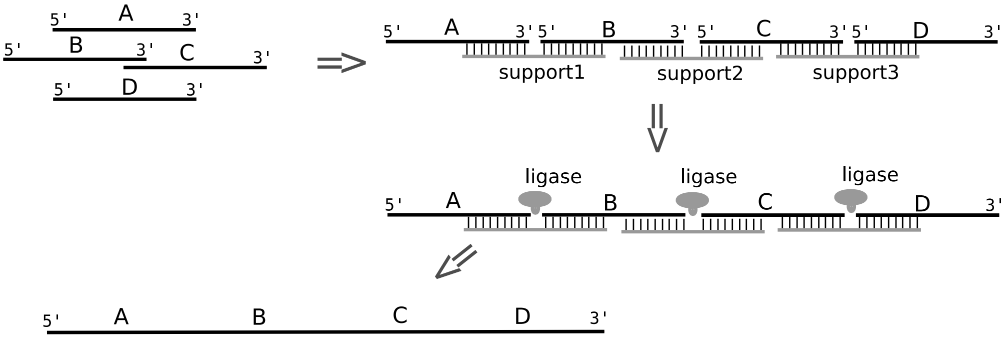

Figure 2. Improved protocol for creating long DNA molecule, example for production from four fragments in one probe. Synthesised short fragments are added to solution with mixture of support molecules, which hybridise and connect contiguous fragments by hydrogen bonds. Next, the DNA ligase is added and the covalent bonds are created in various places, so all fragments are joined. Finally, denaturation is performed.

subsets of molecules, called probes, which could be connected properly in one cycle showed in fig. 2. Initially the one probe synthesis is proposed (when whole synthesis is performed in one cycle), thus the collection In (containing probes) has one element with all fragments.

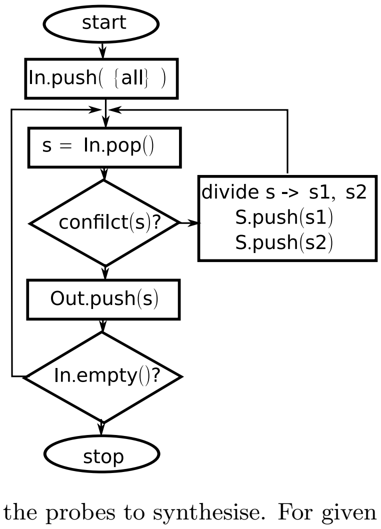

Figure 3. Algorithm finding the probes to synthesise. For given set of fragments and supports it calculate minimal number of probes by separation molecules which hybridise incorrectly.

Algorithm considers the one probe in a loop. The actual probe is analysed to find improper connections between molecules (the request for conflicts in fig. 3). This task involves checking each pair of strands (fragments or supports) from actual probe by modified Needleman-Wunsch algorithm [3, 7] which returns the maximum scored connections between molecules. For these bindings the melting temperature is calculated, and pairs are ordered by this parameter. Because the temperature in probe, after heating, to denaturate every DNA molecule, decrease slowly as shown in fig. 4, the pair of highest melting temperature is taken into account. If this pair represents improper connection between two fragments or between fragment and improper support or between two supports, the conflict is reported and the algorithm separates such molecules in two probes, then each probe is considered separately. Otherwise, if considered pair, of currently highest melting temperature in probe, represents proper connection between fragment and support, the two sequences are eliminated from further analyses and the next pair is taken into consideration. If actual probe is empty the no conflict message is reported.

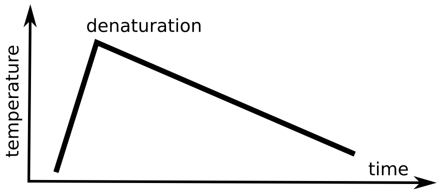

Figure 4. Temperature in probe. Firstly dilution is heated to denaturate, then cooled slowly.

Finally the set of probes is returned (denoted as Out in fig. 3), when each probe contains fragments and supports, that could be connected properly in one ligation. If the number of probes is 2 or more the additional ligations are required to connect the results of synthesis in separated probes.

Function of fitness

The fitness function for individual is a weighted sum of two terms: the number of used probes and the sum of length of used molecules, as depicted in equation 1. The weights α and β are fixed experimentally.

$$F(x) = -\alpha \cdot N(x) - \beta L(x) \tag{1}$$

The number of probes for solution x, denoted by $L(x)$ is returned by algorithm to find the protocol described before. This term represents the time efficiency of proposed protocol. It is assumed, that reactions are performed in series (nonparallel). It should be mentioned, that the longest process is the ligation, it takes up to 24 hours.

The summary length of used molecules, representing the cost of order, is calculated as shown in equation 2, where $|f_i|$ denotes length of i-th fragment and $|supp_i|$ length of i-th supports. All fragments gives the input sequence, so only supports length can be optimized.

$$L(x) = \sum_{i=0}^{k} |f_i| + \sum_{j=0}^{k-1} |supp_i| = |in| + \sum_{j=0}^{k-1} |supp_i| \tag{2}$$

Evolutionary algorithm

The initial population of randomly generated individuals is modified (after selection) by mutation. Implemented EA does not use recombination. The mutation changes:

- the proposed fragments length and sequences, by moving the cutting points on the input sequence
- the proposed supports sequences, by moving the place of hybridisation or by modifying its length

In presented solution the simple (roulette wheel), tournament and marginal reproduction were tested as well as roulette wheel and elite selection. The algorithm stops after given numbers of generations.

Algorithm to find connections between DNA strands

The dynamic programming algorithm proposed by Needleman and Wunsch [3, 7] is used to find the connections between two nucleotide sequences. This algorithms use similarity matrix between nucleotides presented in tab. 1 and the linear penalty for a gap equal -2. The maximum scored double stranded molecule is returned. For example if the sequences are `GAATTC` and `TAATC`, the maximum scored connections are depicted in fig. 5.

Table 1. Similarity matrix used by algorithm to find the alignment between DNA strands

-	**A**	**C**	**G**	**T**
A	-1	-1	-1	2
C	-1	-1	3	-1
G	-1	3	-1	-1
T	2	-1	-1	-1

The presented method does not consider the stiffness of the DNA strands, thus it can return connections that can not be formed, because of very high loop energy (like in position 2 or 3 in fig. 5). It should be mentioned, that only the connections between different molecules are considered by algorithm to find the protocol. The effect of self-overlap (second structures) occurs, when connections between bases on the same strands are found and they are neglected in current version.

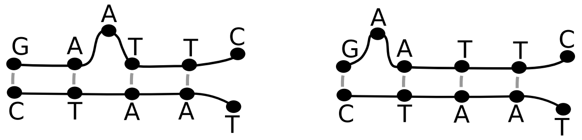

Figure 5. The maximum scored connections of sequences `GAATC` and `TAATC` found by Needleman-Wunsch algorithm. This algorithm is used to eliminate improper joins in hybridisation.

Melting temperature prediction

The melting temperature T_m, temperature when the hydrogen bonds between the bases are breaking, depends on the molecule length as well as on the nucleotide sequence of that molecule. In presented approach the melting temperature is calculated using the equation 3 [6]. This method is accurate, but very imprecise: it counts the number of the cytosine-guanine pairs n_{gc} (with three hydrogen bonds) and number of the adenine-thymine pairs n_{at} (two hydrogen bonds).

$$T_m[^\circ\mathrm{C}] = 3 \cdot n_{gc} + 2 \cdot n_{at} \tag{3}$$

3 Application longdna

Presented algorithm was used in application called longdna. The implementation in C++ depends on the Boost libraries (http://www.boost.org) and faif library (http://faif.sourceforge.net). The source code and binaries for Windows NT/2000/XP and Debian Linux are available at project web page: http://staff.elka.pw.edu.pl/~rnowak2/longdna.

Application was firstly tested on simulated data sets. The short strands were generated to check the proper implementation of used algorithms. Next, the EA with various parameters (population size, reproduction and succession type, fitness weights) was tested, to determine the best set of parameters to find optimal protocol to synthesise long DNA molecules. Finally, the EA was used to optimize the fragments for real sequences. Results are shown in tab. 2.

Table 2. Results for real sequences generated by EA, when population: 20 individuals, reproduction: tournament, succession: elite, stop after 1000 generation, fitness $\alpha = 1$, $\beta = 0.1$. The human insulin and yeast ubiquitin can be produced from 13 strands (7 fragments) in one probe, the human growth hormone needs 3 probes and 4 independent reactions.

input		solution		
name	length	probes	fragments	nucleotides
human insulin	228	1	7	339
yeast ubiquitin	240	1	7	357
human growth hormone	654	3	23	1185

The results were checked by external program [8] to find the bad connections or secondary structures. Because this tool did not report any improprieties, it is concluded that presented solutions are valuable for biologists.

The experiment in genetic laboratory, using proposed method is planned in Institute of Biotechnology and Antibiotics.

4 Conclusions

Presented application can effectively create long DNA strand, which lower costs than typical currently methods used in genetic laboratories. There are many other techniques helps to synthesis long DNA molecules [5], which were used mainly in whole-genome sequencing projects [9]. To find an optimal decomposition to variable-length fragments is believed to be NP-complete and was computed by various algorithms: e.g. simulated annealing [4].

It should be mentioned, that many aspects of used biological reactions are not modeled in current implementation, so the results might be inaccurate. Planned improvements mainly take advantage of better models of genetic reactions. Firstly, the kinetics aspects of hybridisation and ligation (diffusion, zipping, unzipping) should be considered. Secondly, the more precise method for T_m prediction, using the nearest-neighbour DNA/DNA duplexes should be applied. Thirdly, improper connections could be created between nucleotides belonging to the same strand, so the secondary structures should be considered by algorithm for most probable alignment finding. Finally, more accurate similarity matrix may be utilised. On the other hand, the better models for biological reactions may lead to growth the necessary computational power.

Additionally, the use of other optimizing algorithm also could improve results generated by presented application, especially the usage of local optimum finding (e.g. hill climbing) in connection with EA. In the EA the maximum age of individuals is planned to be implemented. To achieve better solution more parameters of developed protocol are considered to bring optimization, e.g. the temperature in probe, which is currently static (as depicted in fig. 4). The other improvements are: the graphics user interface and the import and export module for popular data formats.

The application would be confirmed by larger studies involving experimental work on DNA molecule synthesis, especially the experiment in genetic laboratory using proposed method is required.

Bibliography

[1] M. Amos. *Theoretical and experimental DNA computation.* Springer, 2005.

[2] W. Stemmer at al. Us patent no. 6,368,861, 2002.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis.* Cambridge University Press, 2002.

[4] D. Hoover and J. Lubkowski. Dnaworks: an automated method for designing oligonucleotides for pcr-based gene synthesis. *Nucleic Acids Res*, 30, 2002.

[5] S. Huntsman. Towards the batch synthesis of long dna. Technical Report ADA409078, Institute for Defense Analyses, Alexandria, US, 2002.

[6] K. Lila, R. Kitto, and G.Gloor. A computer scientist's guide to molecular biology. *Soft Computing*, 5:95–101, 2001.

[7] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[8] S. Rozen and H. Skaletsky. Primer3 on the www for general users and for biologist programmers. *Methods Mol Biol*, 132:365–386, 2000.

[9] N. Shevchuk, A. Bryksin, A. Yevgeniya, C. Cabello, M. Sutherland, and S. Ladisch. Construction of long dna molecules using long pcr-based fusion of several fragments simultaneously. *Nucleic Acids Res.*, 32, 2004.

[10] M. Stryer. *Biochemistry.* Freeman, 1995.