VERTICALLY INTEGRATED

PROJECT (VIP) REPORT

APPLICATIONS TO NEUROMORPHIC SOLUTIONS TO HIGH DIMENSIONAL DNA SUBSEQUENCE MATCHING VIA ALIGNMENT DATA

*Ethan Jacob Moyer*

ENGR 370: Applications, Algorithms, & Architecture for

Neuromorphic Computing

*Dr. Anup Kumar Das*

Drexel University

4 May 2020

**APPLICATION**

BACKROUND:

Originally postulated by Watson and Crick in the 1950s, the process of cellular deoxyribonucleic acid (DNA) synthesis has been long understood as a trademark to biological studies (Cooper). Over the past few decades, understanding this process has led to the discovery and development of many synthetic manufacturing methods, namely syntenic gene synthesis (Hughes). As a result, other areas of genomics have accumulated massive research interest. From the start of the Human Genome Project in 1990 to the discovery of Cluster Regularly Interspaced Palindromic Repeat (CRISPR) Cas9 system in 2005, the field of genomics has proven to advance at incredible rates as the reservoir of information pertaining to DNA and gene research increases ("The Human Genome Project”) ("CRISPR Timeline"). Although the public adoption of genomic technologies and tools is arguably in the near feature, the cost of producing and researching novel genes remains as the field’s largest determinant (Tang). In a 2017 publication, two University of Texas laboratory and biological researchers, Randall A. Hughes and Andrew D. Ellington, assert that by decreasing the “DNA read-write cost gap,” humanity could claim this century as the “century of DNA.” Closing this gap will in fact allow researchers to more discover novel genes and genetic systems with far less resources. With the advent of such research, both industrial and academic disciplines involved may witness the advancement of certain fields including “novel therapeutics, materials, [and] biosensing” (Hughes).

One process that the two researchers highlight in their paper is oligonucleotide synthesis (also known as gene synthesis), which ligates, or glues, short single-stranded DNA (ssDNA) together to chemically synthesize genes. Although this general process can be carried out by many different procedures in lab, column-based (or column enclosed) oligonucleotide synthesis remains as an efficient and cost-effective method by which DNA is synthesized. Through a “four-step chain elongation cycle,” ssDNA ranging from 60 to 100 nucleotides are chemically synthesized. In their order, these steps include deprotection, coupling, oxidation, and capping. After these short molecules have been synthesized, nucleotides in each molecule are aligned through a specifically designed base pairing process. An enzyme commonly used in cellular DNA synthesis called DNA Polymerase is used to both fill-in any gaps in the molecule created during the base pairing process and proofread the sequence. The product is a synthetic gene sequence that is identical to one that would be synthesized through a cellular process; the difference, however, is that laboratory researchers can control this synthetic sequence. With current technology, this method can synthesize individual ssDNA at “scales ranging from 10 to 1000 nmol” with costs anywhere between “$0.05 and $0.17 per base” (Hughes). Because of these constraints on synthesis, scaling with larger sequences is a challenge.

RESTRICTION SYNTHESIS:

Stemming from the need to decrease the cost of synthetic synthesis, the rationale behind this research begins with the introduction of a hypothesized catabolic synthesis method (referred to as restriction synthesis): a process that breaks down a DNA into fragments using many endonucleases in order to synthesize a gene. This process is an extension of the BioBricks™ standard proposed by Dr. Thomas Knight. In their publication, BioBricks™ is defined as a process that “employs iterative restriction enzyme digestion and ligation reactions to assemble small basic parts into larger composite parts.” These larger components are built using fragments flanked by two restriction enzymes, Xbal and SpeI (Anderson). One key difference between restriction synthesis and BioBricks™ is that the latter restricts digestion to those two enzymes: XbaI and SpeI. Conversely, restriction synthesis relies on a dataset of more than 200 enzymes all with unique restriction cutting sites. In flanking subsequences with a greater number of available restriction enzymes, restriction synthesis was originally thought to build many more composite subsequences, which are homologous to those referenced by the BioBricks™ standard.



The development of an algorithmic simulation of restriction synthesis began in 2016. Since the project’s inception, the estimated cost of synthesis has shown to decrease rapidly as the simulation grew more representative of the proposed lab procedure. With costs now as low as $0.019/bp, this simulation is ready for its next stage of development, predictive subsequence classification, which will be major topic of this paper.

In regard to synthesizing DNA, catabolic synthesis is counterintuitive compared to the anabolic methods. Nevertheless, this process generates low simulated costs because of its use of restriction enzymes. During the simulation of synthesis, the algorithm searches for applicable sequences. As it will be explained more in depth later in the report, an applicable sequence is one that can be flanked and digested by two enzymes and be end-to-end compatible with sticky-ended sequences before and after it. Therefore, the total cost of synthesizing a fragment is determined by the individual unit cost of the enzymes used, and the cost per base pair ($/bp) is determined by the length of the flanked sequence. In this way, longer sequences spread the overhead of the cost for each iteration, decreasing the total for the entire process. In other conventional methods such as anabolic gene synthesis, the cost of production is fixed to the sum of each singular reaction of nucleotide-to-nucleotide elongation. As the main determinant of production, this cost metric is integral to increasing the efficiency of the gene synthesis process.

ALGORITHM:

Restriction synthesis is currently modeled computational based on an iterative restriction enzyme digest and sticky end alignment process. Given a gene and a larger DNA sequence (also known as the set), the algorithm simulates the process by first searching for fragments (or subsets) in the set that sequentially appear in the gene. These matches are considered subsequence matches, which is one of three conditions required for a fragment to be used for the synthesis process. The length of these fragments varies with each search depending on what subsets are available in the set. It is more likely, for instance, to find a subsequence match of length seven compared to a subsequence match of length fourteen. Taking this into account, each subsequence search begins with parsing for subsequences of length t, and then continues by decrease this value until a subsequence match is found (t is set to 16 by default). Through this method, each iteration for a subsequence match ensures that the longest possible subsequence match is found.

An instance match, unlike a subsequence match, is based on whether two enzymes exist that can flank and digest the fragment in the set. This condition is critical because while there exist thousands of subsequence matches, only a select few may be applicable for the synthesis process—in order for a subsequence to be included from the set, it needs to be removed by flanked enzymatic digest (Anderson). Using the New England Biolabs’ restriction enzyme database, the algorithm has access to 204 unique restriction enzymes and their respective restriction sites (New England Biolabs). Despite this low number of enzymes, many have versatile recognition sequences, cutting a variety of DNA patterns. Access to these versatile enzymes is one of the reasons why this process can successfully complete on a variety of genes.

In addition to a subsequence match and an instance match, a third match is required: a ligation match. Some restriction enzymes cut DNA in a Z-like manner, where the product of the enzymatic digest contains nucleotide overhangs, commonly referred to as sticky ends. Because many of the enzymes create these sticky ends on digested fragments, a ligation match is needed to ensure that sequential fragments are compatible. For instance, a subsequence that ends in an overhang of AGTA on the 5’ end needs to be followed with an overhang that begins with TCAT on the 3’ end. Also, fragments cut with enzymes that produce blunt ends, or ends without any overhangs, need to be followed with fragments that have these produced blunt ends as well. This final end-to-end compatibility indicates whether a fragment is complete match and whether the search process continues.

These three main conditions, subsequence match, instance match, and ligation match, dictate whether a subset in the set can be including in the synthesis process. Specifically, the two former conditions rely entirely on the sequence of the gene and the set. For this reason, both DNA sequences need to be known prior to performing this method. This assumption would require costly sequencing if each gene was synthesized from a different set. In other words, each unique gene sequence would have a respective sequenced set from which subsequences would be pulled. While this solution would allow for incredibly accurate synthesis, it would massively increase the cost of this method. Another solution is to have one singular set that can be applied to a category of genes, ones that are alike in their general nucleotide composition. While this solution would allow for prices to remain constantly low, it requires choosing which subsets of fragments to include a generalized set. This proposed solution is the most cost-effective out of the two. Because of the algorithm’s high reliance on subsequence matching, this will be the main focus of the paper.

**REFERENCES**

Anderson, J Christopher et al. “BglBricks: A flexible standard for biological part assembly.” Journal of biological engineering vol. 4,1 1. 20 Jan. 2010, doi:10.1186/1754-1611-4-1. Accessed Web. 24 Oct 2019.

Cooper, Geoffrey. "DNA Replication." *Sinauer Associates* (2000). Accessed Web. 27 Feb. 2020.

"CRISPR Timeline." *Broad Institute*. 2015. Accessed Web. 3 Mar. 2020.

Goodfellow, Ian, et al. Deep Learning. The MIT Press, 2017. p. 326

Hughes, Randall A, and Andrew D Ellington. “Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology.” Cold Spring Harbor perspectives in biology vol. 9,1 a023812. 3 Jan. 2017, doi:10.1101/cshperspect.a023812. Accessed Web. 24 Oct 2019.

New England Biolabs. “Restriction Endonucleases.” NEB, www.neb.com/products/restriction-endonucleases. Accessed Web. 29 Oct 2019.

Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – Accessed Web. 24 Jan. 2020. Available from: https://www.ncbi.nlm.nih.gov/nuccore/

Tang, Nicholas, Siying Ma, and Jingdong Tian. "New Tools For Cost-Effective DNA Synthesis." Synthetic Biology (2013): 3-21. Accessed Web. 1 Mar. 2020.

"The Human Genome Project." Genome.gov. *National Human Genome Research Institute*. 2020. Accessed Web. 3 Mar. 2020.