DNA media storage

Christy M. Bogard, Eric C. Rouchka*, and Benjamin Arazi

Computer Engineering and Computer Science, University of Louisville, Louisville, KY 40292, USA
(Christy.Bogard @louisville.edu) (Eric.Rouchka @louisville.edu) (Benjamin.Arazi @gmail.com)

1 Introduction

In 1994, University of Southern California computer scientist Dr. Leonard Adleman solved the Hamiltonian path problem using DNA as a computational mechanism. He proved the principle that DNA computing could be used to solve computationally complex problems. Because of the limitations in discovery time, resource requirements, and sequence mismatches, DNA computing has not yet become a commonly accepted practice. However, advancements are continually being discovered that are evolving the field of DNA computing. Practical applications of DNA are not restricted to computation alone. This research presents a novel approach in which DNA could be used as a means of storing files. Through the use of multiple sequence alignment combined with intelligent heuristics, the most probabilistic file contents can be determined with minimal errors.

2 DNA representation of information

How one approaches a problem is often defined in how the problem is represented. Various representations lend themselves to a set of predefined actions that can easily shape one's perspective and approach in the quest for the solution. For example, when presented with the problem of determining the time at which a thrown ball is at a given height, it is often easier to decipher the two solutions from a graphical representation of the problem as opposed to an algebraic representation.

While computer scientists have long used the notion of a binary bit to represent digital information – 1 to indicate the state is present and 0 to indicate the state is not – geneticists use a quaternary alphabet to encode information, using the symbols A, C, G, and T. Translating between the computer scientist's alphabet and the geneticist's representation is easily accomplished through a direct substitution of two binary base pairs encoding for a single quaternary character.

Digital → DNA			
00 →A	01 →C	10 →G	11 →T

Figure 1: Conversion between digital bit-based and DNA-based alphabet.

3 Adleman and the Hamiltonian path problem

A Hamiltonian path is defined as a route through an undirected graph which visits each vertex in the graph exactly once.. The Hamiltonian path problem (HPP) aims to find the lowest cost Hamiltonian path within a graph. One specific variant of HPP is the Traveling Salesman Problem (TSP), where the vertices in the graph represent different cities, and the edges represent the cost to travel between a set of cities.

In 1994, University of Southern California computer scientist Dr. Leonard Adleman solved the Hamiltonian path problem using DNA as a computational mechanism [1,2]. Adleman began by using 20-mer oligonucleotide sequences to uniquely represent each city. Paths were represented using complementary 20-mer oligonucleotide sequences generated by combining the last 10 bases of the starting city with the first 10 bases of the ending city. When the oligonucleotide sequences were combined, DNA's desire to form a stable double helix structure enables the paths to be constructed through the combination of the city sequences with the complementary edge sequences.

Once all representations of the cities and corresponding paths were in place, a large number of copies were generated to produce all possible combinations of cities and edges, in effect generating all possible paths through the graph. Paths that did not meet all of the problem rules – i.e. those that did not consist of exactly seven edges, and those with duplicated cities within the generated path – were systematically eliminated. Any remaining generated paths are valid Hamiltonian paths through the graph. If no generated paths remain, then the graph does not contain any Hamiltonian paths.

* Corresponding author, E. C. Rouchka can be reached at 1-502-852-1695.

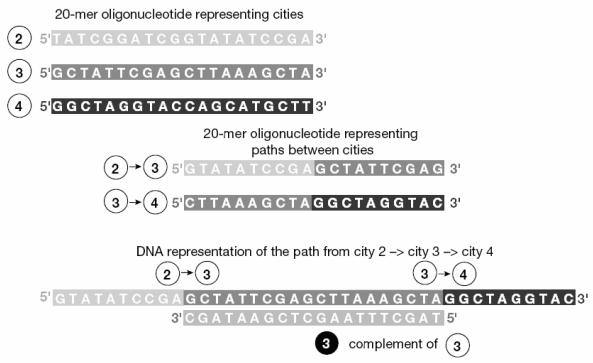

Figure 2: DNA Representation of the Traveling Salesman Problem. Image adapted from Parker, 2003 [3].

3.1 Limitations of Adleman's solution

Adleman's solution to the Hamiltonian path problem proved that DNA could in principle be used to solve NP-complete problems. While DNA has the ability to quickly enumerate all possible solutions to the seven – city Hamiltonian path problem in only a few hours, it requires a lengthy discovery time to experimentally determine the DNA solution, taking nearly seven days to complete. Secondly, while DNA requires significantly less space to store a single bit – requiring only 1 nm^3 compared to current methods requiring 10^{12} nm^3 – it is not easily scalable; the DNA required to enumerate the Hamiltonian path problem for 200 cities would exceed the weight of the earth [3]. Finally, since Adleman's experiment was limited to only seven cities, he could represent the cities with distinct sequences as to minimize the number of alignments that would result in solutions that do not exist. However, as the number of cities increases, it becomes more difficult to uniquely represent the cities in such a manner as to avoid mismatched alignments. Therefore, additional error-checking will be required to ensure accurate solutions.

4 Using multiple sequence alignment in error reduction

DNA allows for a drastic reduction in storage space per bit compared with traditional digital computing. As a result, redundant storage capabilities and parallel processing on the exact same data are feasible. However, if the storage or computation results in inconsistencies, determining which are correct and which are not is problematic. The bioinformatics technique of multiple sequence alignment yields insight into how the issue of data integrity can be solved.

4.1 Multiple sequence alignment

Multiple sequence alignment is the process of finding a representative, or consensus, model of the similarities between three or more sequences. Like pairwise sequence alignment, it finds an optimal solution for the model conditions placed upon it. If the conditions are changed, then the model may or may not hold. For a set of highly conserved sequences, the multiple sequence alignments are easily seen, even with the naked eye. As the sequences diverge, so does the complexity of finding the best alignment [4].

Once a multiple sequence alignment is in place, it can be described using a number of different approaches. The most useful of these represents the alignment as a statistical model, known as a profile Hidden Markov Model (HMM) [4]. HMMs have the power to represent the alignment through states for insertions, deletions, and matches/mismatches found within the alignment. For the match/mismatch and insertion states, an associated emission probability is given to the observed characters for a particular position.

4.2 Multiple sequence alignment for error reduction

Since multiple sequence alignment is sensitive to sequence similarities, it can be used to combine the multiple copies of the same file to find the most probabilistic contents. There are three scenarios that can be discovered: (1) areas completely conserved among all of the sequences, (2) areas highly conserved among the sequences, and (3) areas that are not conserved among the sequences. Each of these scenarios directly corresponds with the level of error within the region.

First, consider areas that are completely conserved. In this case, no mutations have occurred in any of the file copies, and as such, the region is completely 100% free of errors.

For highly conserved areas, discrepancies indicate potential errors that have been introduced. Since a multitude of copies have been stored, it is probable the majority of sequences will be highly correlated. Thus, the emission properties of the associated Hidden Markov Model state will clearly indicate which one of the bases is most probable of being emitted as it will have a significantly higher emission over the remaining bases.

Finally, consider areas that are not conserved among the sequences. It may not be possible to determine the most probabilistic emission because a significant number of discrepancies have been introduced into the region. Since there can be no determination as to what the sequence was originally, this region represents the

system state of irrecoverable errors. In such circumstances, alternatives must be employed to determine the state.

4.3 Improving multiple sequence alignment

The genetic code allows for a three-base nucleotide sequence (codon) to encode for one of twenty amino acids. Since there are four possible bases (A, C, G, T) for each of the three possible bases of the amino acid, there are a total of sixty-four possible combinations, meaning there are multiple codon representations that encode for a single amino acid. Consequently, alignment of the translated amino acid sequences has a greater probability of defining more highly conserved regions that may be indeterminate at a DNA sequence level. Alignment of regions of low conservation can potentially be improved by aligning the corresponding translated amino acid sequences.

4.4 Heuristic improvements of the algorithm

Knowing that the aligned sequences are very similar, if not identical, there are number of heuristics that can be applied to reduce the computational, storage, and time complexity required for the multiple sequence alignment. Continuing with the example of the storage of a file, it is reasonable to assume that the majority of sequences being aligned will be of the same length within a bounded threshold. Thus, one can quickly eliminate sequences which are disproportionately longer or shorter than majority of sequences being aligned.

Additionally, since the sequences are highly similar, the alignment will probabilistically follow the diagonal of the dynamic programming alignment matrix [5,6]. Thus, performing a bounded alignment along the diagonal are calculated will reduce the computational complexity and the storage complexity required for all of the pairwise sequence alignments performed. It is reasonable to assume that the threshold could be set between 5-10% and still produce highly accurate results.

Finally, an intelligent agent could be introduced to retain the probabilities of the identical alignments without the actual storage of the alignments. The frequencies of the identical sequences must be retained in order for the Hidden Markov Model emissions to be representative of the aligned sequences, but the actual alignment does not need to be retained.

Acknowledgements

This project was made possible by NIH – NCRR grant P20RR16481 and NIH – NIEHS grant 2P30ES014443 – 01A1. Its contents are solely the responsibility of the authors and do not represent the official views of NCRR, NIEHS, or NIH. The authors would also like to acknowledge the support provided by Hank and Becky Conn through the University of Louisville Conn Fellowship.

References

[1] Adleman, Leonard M. (1994): 'Molecular Computation of Solutions to Combinatorial Problems', Science (November 1994): 1021-1024.

[2] Amos, Martyn (2005): Theoretical and Experimental DNA Computation, Netherlands: Springer.

[3] Parker, Jack (2003): 'Computing with DNA', European Molecular Biology Organization (January 2003): 7-10.

[4] Rabiner, L. R. and B. H. Juang (1986): 'An Introduction to Hidden Markov Models', *IEEE ASSP Magazine*, (Jan 1986): 4-16.

[5] Carillo, H and D. Lipman (1988): 'The Multiple Sequence Alignment Problem in Biology', *SIAM Journal on Applied Mathematics* (October 1988): 1073-1082.

[6] Myers, Eugene W (1991): 'An Overview of Sequence Comparison Algorithms in Molecular Biology', University of Arizona, Department of Computer Science, Technical Report TR 91-29.