Practical Algorithms for Universal DNA Primer Design: An Exercise in Algorithm Engineering

Patricia Evans [1] **Todd Wareham** [2]

Keywords: Closest Substring problem, Hamming distance, DNA primer design, algorithm engineering, parameterized computational complexity theory

1 Problem and Motivation

Given two strings s and s' of the same length over an alphabet Σ, the *Hamming distance* $d_H(s, s')$ between s and s' is the number of position at which the symbols in s and s' differ. Consider the following problem:

CLOSEST SUBSTRING [6]
Input: A set S of strings over an alphabet Σ such that the length of each string $s \in S$ is n and positive integers l and k such that $1 \leq l \leq n$ and $1 \leq k \leq l$.
Question: Is there a string $c \in \Sigma^l$ such for each string $s \in S$, there is a substring s' of s such that $d_H(s', c) \leq k$?

This problem occurs in the design of PCR primers for groups of DNA sequences, as the strength of hybridization of a primer s to a DNA sequence s' is strongly correlated with the number of mismatches between the sequences of s and the hybridizing substring of the same length in s' [10] (see [6] for further applications).

2 Previous Work

When $m = 2$, the CLOSEST SUBSTRING problem is solvable in $O(knm)$ time [5, Section 12.2.5]. The general version of the problem is known to NP-hard and has deterministic [6] and randomized [7] polynomial time approximation algorithms that find solution-strings that are within multiplicative factors of 2 and $2 - \frac{2}{2|\Sigma|+1}$ mismatches of the optimal solution, respectively. Given the sensitivity of hybridization strength to the number of mismatches, this may not be acceptable in practice. A polynomial time approximation scheme has recently been derived [8] but its time complexity makes it unusable in practice.

There are many primer design programs described in the literature, *e.g.*, [3, 9]. Unfortunately, these programs either require that given sequences be aligned or do not specify the algorithm used (and are hence probably invoking heuristics that are not guaranteed to give all or even any optimal solutions).

[1] Faculty of Computer Science, University of New Brunswick, Fredericton, NB Canada E3B 5A3. E-mail: `pevans@unb.ca`

[2] Department of Computer Science, Memorial University of Newfoundland, St. John's, NF Canada A1B 3X5. E-mail: `harold@cs.mun.ca`

3 Our Approach

As Closest substring is NP-hard, it is unlikely that the problem can be solved exactly using polynomial time algorithms. However, instances of the primer design problem encountered in practice typically have restricted values for $|\Sigma|$, l, k, and m, *i.e.*, $|\Sigma| = 4$, $l \leq 25$, $k \leq 2$, and $m \leq 20$. Hence, it seems likely that algorithms whose running-times are non-polynomial purely in terms of these parameters may yet be usable in practice. Our research has focused on exploring the space of possible non-polynomial time algorithms for Closest substring within the framework of parameterized computational complexity theory [1, 2]. To date, we have derived algorithms whose running times are non-polynomial in terms of

- $|\Sigma|$ and l, *i.e.*, $O(|\Sigma|^{2l} + lmn)$;
- $|\Sigma|$, l, and k *i.e.*, $O(|\Sigma|^{l+k} l^k + lmn)$; and
- $|\Sigma|$, l, k, and m *i.e.*, $O((\min(n-l, |\Sigma|^l))^m |\Sigma|^{km+k} l^k ml + lmn + mn \log n)$.

Theoretical analyses have been supplemented with experience gained by implementing and running these algorithms, which has pinpointed how the actual time and space requirements can be reduced by applying various techniques, *e.g.*, bitwise encoding of solution-strings.

4 Results

We have run our algorithms against two datasets provided by Drs. Steve Carr and Dawn Marshall of the Department of Biology at Memorial University of Newfoundland. These datasets consist of the complete mitochondrial genomes of 4 species of seals and 12 species of fish, respectively ($n \approx 17,000$). Various interesting patterns have been found, *e.g.*, several unusually long perfectly conserved regions, as well as primers that were not obvious to Marshall, who has extensive experience in designing such primers by conventional methods.

References

[1] Bodlaender, H., Downey, R., Fellows, M., Hallett, M., and Wareham, T. 1995. Parameterized complexity analysis in computational biology. *CABIOS* 11(1):49–57.

[2] Downey, R. and Fellows, M. 1999. *Parameterized Complexity Theory.* Berlin: Springer-Verlag.

[3] Dopazo, J., Rodríguez, A., Sáiz, J.C., and Sobrino, F. 1993. Design of primers for PCR amplification of highly variable genomes. *CABIOS* 9:123–125.

[4] Frances, M. and Litman, A. 1997. On Covering Problems of Codes. *Theory of Computing Systems* 30:113–119.

[5] Gusfield, D. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press.

[6] Lanctot, J.K., Li, M., Ma, B., Wang, S., and Zhang, L. 1999. Distinguishing String Selection Problems. In: *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'99)*, ACM Press. 633–642.

[7] Li, M., Ma, B., and Wang, L. 1999. Finding similar regions in many strings. In: *Proceedings of the 34th Annual ACM Symposium on the Theory of Computation (STOC'99)*, ACM Press.

[8] Ma, B. 2000. A Polynomial Time Approximation Scheme for the Closest Substring Problem. In Giancarlo, R. and Sankoff, D., editors, *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching (CPM 2000)*, Lecture Notes in Computer Science no. 1848, Berlin: Springer-Verlag. pp. 99–107.

[9] Proutski, V. and Holme, E.C. 1996. Primer Master: a new program for the design and analysis of PCR primers. *CABIOS* 12:253–255.

[10] Wetmur, J.G. 1991. DNA Probes: Applications of the Principles of Nucleic Acid Hybridization. *Critical Reviews in Biochemistry and Molecular Biology* 26:227–259.