A COMPARATIVE STUDY OF CROSS-CORRELATION METHODS FOR ALIGNMENT OF DNA SEQUENCES CONTAINING REPETITIVE PATTERNS

Andrzej K. Brodzik

The MITRE Corporation, Bedford MA 01730

ABSTRACT

In this work we consider the problem of global DNA sequence alignment. One of the best known and most efficient computational techniques used for this task is the cross-correlation method. We compare efficacy of evaluating periodic DNA sequence misalignment using the standard magnitude-and-phase cross- correlation technique with the lesser known phase-only cross-correlation method. We prove that for a periodic DNA sequence whose length is a prime number the standard approach leads to significant sidelobes in the cross-correlation, the magnitude of which increases with the length of the sequence, while the phase-only approach allows attaining a perfect cross-correlation with zero sidelobes. Numerical experiments on synthesized data are included and robustness of the phase-only method to random DNA insertions and imperfect DNA fragment matches is discussed.

1. INTRODUCTION

One of the outstanding problems in genomic signal processing is that of DNA sequence alignment. Typically, an unknown collection of smaller length (few tens to few thousands of bases) DNA fragments is acquired, which is then compared with one of several known collections of DNA fragments contained in the library. Either or both of these collections might be incomplete, unordered and contain errors, including symbol mismatches and symbol deletions. Finding a match between the unknown and one of the known DNA collections allows identification of an individual's DNA, or DNA fingerprinting [4].

A different challenge is posed by the problems of pathogen detection and gene finding. In these cases, instead of a library of known DNA fragments, a specific DNA pattern is given that is either part of a pathogen signature or is informative of a start of a coding region. This pattern needs to be compared with the DNA sequence under investigation. A match of the pattern with a specific region of the DNA sequence confirms previous pathogen exposure or identifies an exon [15].

In this work we focus on periodic DNA sequence alignment using the cross-correlation methods. While in general the distribution of DNA symbols in a sequence can be random, it is often of interest to compare DNA sequence fragments containing repetitive patterns. This is due to the fact that many relevant genetic phenomena such as mutations, genetic diseases, or start of a coding region, can be associated with occurrence of periodically spaced DNA symbols [9], [15]. In fact, DNA repeats are estimated to comprise more than one half of the human genome [12].

The methods most often used for DNA sequence alignment are based on dynamic programming. Unfortunately, dynamic programming techniques suffer from high computational cost and algorithmic complexity, which reduces their utility when applied to any significant subset of the three billion letters of the human genome [13]. This is in contrast to the cross-correlation or matched filter (MF) methods, which due to the availability of the Fourier transform techniques, offer simplicity of implementation and computational efficiency [6], [14]. While the high efficiency of the cross-correlation method makes it an attractive alternative to dynamic programming, its wider use is limited by the lack of robustness to partial sequence mismatches and by an ambiguous misalignment reading when applied to periodic sequences.

In this paper we propose to replace the MF technique with its modification, the symmetric phase-only MF (SPOMF). SPOMF was proposed two decades ago in optical signal processing [10], and heuristic arguments have been made that the method is superior to the standard approach in terms of misalignment resolution and robustness to noise. Since then it has been successfully applied to image registration [7], watermarking [11], and sonar [5]. We prove that when applied to a periodic DNA data sequence whose length is a prime number, SPOMF produces a perfect cross-correlation, while the standard MF produces sidelobes whose magnitude is proportional to the length of the sequence. To our knowledge this is the first theoretical result that proves superiority of SPOMF for a particular class of signals and the first application of phase-only filtering in genomics. To illustrate the theoretical developments we include numerical experiments in which we compare the two methods in robustness to random insertions and local mismatches.

2. MF AND SPOMF

Define the cyclic cross-correlation, or MF, of two real discrete sequences x and y by

$$z(n) = x(n) * y(n) = \sum_{m=0}^{N-1} x(n+m)y(m), \quad 0 \leq n < N, \quad (1)$$

with $n + m$ taken modulo N. Take $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ to be the discrete Fourier transforms of x, y and z, respectively.
Since

$$\mathbf{z}(k) = \mathbf{x}(k)\bar{\mathbf{y}}(k), \quad (2)$$

(1) can be efficiently implemented by using the Fourier transformed sequences, i.e.,

$$z(n) = \mathrm{DFT}^{-1}\left\{\mathbf{x}(k)\bar{\mathbf{y}}(k)\right\}. \quad (3)$$

Approved for Public Release; Distribution Unlimited; Case # 05-0073.

Define SPOMF of x and y by

$$w(n) = \mathrm{DFT}^{-1}\left\{\frac{\mathbf{x}(k)\bar{\mathbf{y}}(k)}{|\mathbf{x}(k)\bar{\mathbf{y}}(k)|}\right\}, \quad \mathbf{x}(k) \text{ and } \bar{\mathbf{y}}(k) \neq 0. \tag{4}$$

3. REGULAR PERIODIC DNA SEQUENCES

DNA sequence is a symbolic string of characters 'a', 'c', 'g' and 't'. Various methods of mapping a symbolic DNA sequence to a numeric sequence have been proposed, including the use of complex and hypercomplex number systems [2], [3], [6]. For the sake of simplicity, but without a loss of generality, in the next two sections we will consider only single symbol DNA sequences, which will be represented by binary numbers. Furthermore, we will only consider periodic sequences (in the sense specified below), which can be conveniently modeled as combs. In the last section we apply the developed formalism to the 4-symbol DNA data and discuss processing of non-periodic DNA sequences.

Definition 1 Take $N,\ P,\ S \in Z^+$, such that P is a divisor of N, and $0 \leq S < N$. A regular P-periodic comb is an N-point sequence

$$x_{N,P}(n) \;=\; \begin{cases} 1, & n \bmod P = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

An S-shift of a regular P-periodic comb is a sequence $x_{N,P,S}(n) = x_{N,P}(n-S)$, with $n-S$ taken modulo N. The first result shows that if $x(n)$ is a comb then $\mathbf{x}(k)$ is also a comb.

Theorem 1 Take $N,\ P,\ S \in Z^+$, such that $\bar{P} = N/P \in Z,\ \ 0 \leq S < N$. The Fourier transform of an S-shift of a P-periodic comb $x_{N,P,S}$ is a scaled and modulated $\bar{P}$-periodic comb $\bar{P}e^{2\pi ikS/N}x_{N,\bar{P}}$.

Proof

$$\begin{aligned} \mathbf{x}_{N,P,S}(k) &= \sum_{n=0}^{N-1} x_{N,P,S}(n)e^{2\pi ink/N} \\ &= e^{2\pi ikS/N}\sum_{s=0}^{\bar{P}-1} e^{2\pi isk/\bar{P}}, \\ &= \begin{cases} \bar{P}e^{2\pi ikS/N}, & k \text{ a multiple of } \bar{P}, \\ 0, & \text{otherwise.} \ \square \end{cases} \end{aligned}$$

We can now compute MF and SPOMF of regular combs.

Theorem 2 Take $x = x_{N,P,0} = x_{N,P}$ and $y = x_{N,P,S}$. The MF of x and y is

$$z(n) \;=\; \begin{cases} \bar{P}, & (n+S) \bmod P = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Proof

$$z(n) = \frac{1}{N}\sum_{k=0}^{N-1} e^{-2\pi ink/N}\mathbf{x}(k)\bar{\mathbf{y}}(k) = \frac{\bar{P}^2}{N}\sum_{l=0}^{P-1} e^{-2\pi il(n+S)/P},$$

which leads to (6). □

Theorem 2 shows that MF of a regular comb provides a measure of DNA sequence misalignment provided $S < P$, which is of limited use. Due to theorem 1 SPOMF of a regular comb is not defined, since insertion of $\mathbf{x}_{N,P}$ into (4) results in division by zero. One way to circumvent this problem is to assign to the "divide by zero" points some fixed small value. Like MF however, this approach provides a measure of DNA sequence misalignment, only when $S < P$. Moreover, there is a performance penalty associated with the zeroes of the DFT, manifesting itself in sidelobes of the cross-correlation sequence. As will be shown in the next section, both obstructions can be removed by restricting the length of the comb to a prime number.

4. IRREGULAR PERIODIC DNA SEQUENCES

Definition 2 Take $N,\ P,\ S \in Z^+$, N an odd prime. An irregular P-periodic comb is an N-point sequence

$$x'_{N,P}(n) \;=\; \begin{cases} 1, & n \bmod P = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

As before, $x'_{N,P,S}(n) = x'_{N,P}(n-S)$, with $n-S$ taken modulo N. The restriction of N to primes is not very severe. For example, there are 25 primes between 1 and 100, 21 primes between 101 and 200, and 18 primes between 201 and 300, all of them fairly uniformly distributed [1]. Moreover, as with the power of two length DFT, there are fast algorithms for computing a prime length DFT. We consider the DFT of an irregular comb next.

Theorem 3 Take $N,\ P,\ S \in Z^+$, N an odd prime. The Fourier transform of an irregular P-periodic comb $x'_{N,P,S}$ is

$$\mathbf{x}'_{N,P,S}(k) = \begin{cases} \lfloor \frac{N}{P} \rfloor + 1, & k = 0, \\ \frac{1-\exp(2\pi ikP(\lfloor N/P \rfloor+1)/N)}{1-\exp(2\pi ikP/N)}e^{2\pi ikS/N}, & \text{else,} \end{cases} \tag{8}$$

where $\lfloor N/P \rfloor$ is the largest integer not greater than N/P.

Proof We have

$$\mathbf{x}'_{N,P,S}(k) = \sum_{n=0}^{N-1} x'_{N,P,S}(n)e^{2\pi ink/N} = e^{2\pi ikS/N}\sum_{s=0}^{\lfloor \frac{N}{P} \rfloor} e^{2\pi iskP/N},$$

which leads to (8). □

Corollary 1 $\mathbf{x}'_{N,P,S}(k) \neq 0\ \forall k$.

Proof Follows directly from theorem 3. □

In effect, by selecting a prime N, the zero obstruction is removed. The next three results characterize performance of MF of an irregular comb and are key in this paper.

Theorem 4 Set $\delta = \lfloor \frac{N}{P} \rfloor + 1$. The MF of irregular P-periodic comb $x'_{N,P,S}$ is given by

$$z'(n) = \frac{\delta^2}{N} + \frac{1}{N} \sum_{k=1}^{N-1} e^{-2\pi i k(n+S)/N} \frac{1-\cos(2\pi k P \delta / N)}{1-\cos(2\pi k P/N)} \quad (9)$$

Proof Follows directly from inserting $\mathbf{x}'_{N,P,S}$ and $\mathbf{x}'_{N,P}$ (theorem 3) into (3). □

Theorem 5 Take N a prime, $N \geq 3$, $P \in Z$, $1 < P < N$, and $\delta = \lfloor \frac{N}{P} \rfloor + 1$. Then the mainlobe $\mathcal{M}$ of MF of an irregular P-periodc comb is given by

$$\mathcal{M} = z'(n = -S) = \delta. \quad (10)$$

Proof Follows directly from (1) and (7). □

Theorem 6 Take N a prime, $N \geq 3$, $P \in Z$, $1 < P < N$, and $\delta = \lfloor \frac{N}{P} \rfloor + 1$. Then the largest sidelobe $\mathcal{S}$ of MF of an irregular P-periodic comb is given by

$$\mathcal{S} = z'(n = P - S) = \delta - 1. \quad (11)$$

Proof Follows from application of theorem 4 to the difference $\mathcal{M} - \mathcal{S}$. Details of the proof will be given elsewhere. □

The performance of MF of an irregular comb is summarized by the following corollary.

Corollary 2 The ratio of mainlobe to the largest sidelobe of MF of an irregular P-periodic comb is given by

$$\frac{z'(-S)}{z'(P-S)} = \frac{\delta}{\delta - 1}. \quad (12)$$

The last result characterizes performance of SPOMF of irregular comb.

Theorem 7 Set $\delta = \lfloor \frac{N}{P} \rfloor + 1$. The SPOMF of an irregular P-periodic comb $x'_{N,P,S}$ is given by

$$w'(n) = \begin{cases} 1, & n = -S, \\ 0, & \text{otherwise}. \end{cases} \quad (13)$$

Proof Using (4) and (9) we have

$$\begin{aligned} w'(n) &= \frac{1}{N} \sum_{k=0}^{N-1} e^{-2\pi i k n/N} \frac{\mathbf{x}(k)\bar{\mathbf{y}}(k)}{|\mathbf{x}(k)\bar{\mathbf{y}}(k)|} \\ &= \frac{1}{N} + \frac{1}{N} \sum_{k=1}^{N-1} e^{-2\pi i k(n+S)/N} \frac{\mathrm{sign}(1-\cos(2\pi k P \delta/N))}{\mathrm{sign}(1-\cos(2\pi k P/N))} \\ &= \frac{1}{N} + \frac{1}{N} \sum_{k=1}^{N-1} e^{-2\pi i k(n+S)/N}, \end{aligned}$$

which leads to (13). □

As can be seen from theorems 4-7, when the DNA sequence is an irregular P-periodic comb, then SPOMF significantly outperforms MF. While SPOMF yields a perfect cross-correlation sequence, MF gives rise to sidelobes, the largest of which approaches the magnitude of the cross-correlation mainlobe, as N increases and P remains constant. In the penultimate section of this paper we will provide illustrations of performance of MF and SPOMF, and consider a scenario of DNA sequence alignment, where the sequences to be compared are contaminated by distinct random insertions.

5. EXPERIMENTS

In the first experiment we have compared the performance of MF with the performance of SPOMF, for a single symbol periodic sequence alignment ($x_{29,5}$). The prime length SPOMF yields a perfect cross-correlation sequence (Figure 1). MF of $x_{29,5}$ has a peak at $n = N - S$ equal to $\delta = \lfloor \frac{29}{5} \rfloor + 1 = 6$, and multiple sidelobes, the largest one occurring at $n = N - S - P$ and being equal to $\delta - 1 = 5$, as predicted by theorems 5 and 6.

Figure 2 illustrates processing of the same sequence, but with the inclusion of a multiple symbol pattern, 'acagt'. The symbols a, c, g and t are marked in the plot with stems equal to 1, 2, 3 and 4. The cross-correlations shown are the sums of cross-correlations performed on the individual symbol sequences. As before, MF contains sidelobes, while prime length SPOMF produces a perfect cross-correlation sequence.

The previous two experiments have shown examples of cross-correlations performed on purely periodic, perfectly matched data. In practice, the DNA fragments that need to be compared often contain sequence mismatches and random sequence insertions. Figure 3 shows an example of such a case (the MF and SPOMF sequences shown in the plot have been normalized to illustrate the relative difference in sidelobe magnitude). The two misaligned sequences, $x_{61,5}$ and $y_{61,5}$, contain a 46-base fragment, the first 16 bases of which are random, the remaining 30 bases constitute a repetitive 'acagt' pattern. The library sequence ($x_{61,5}$) is appended by a random 15-base fragment and the query sequence ($y_{61,5}$) is concatenated to a different random 15-base fragment. In effect, $x_{61,5}$ and $y_{61,5}$ are misaligned by 15 bases and mismatched (when aligned) at 15 bases. The cross-correlation peak in both cases occurs at $n = N - S = 46$. While the prime length SPOMF does not produce an ideal cross-correlation sequence in this case, its sidelobes are significantly smaller then the sidelobes of the MF (0.248 vs 0.740).

The last experiment involved sequences with unequal gaps between patterns ($x_{71,5}$ and $y_{71,5}$, Figure 4). The sequences are similar to the ones in the previous experiment, except that the repetitive pattern is 35 bases long (rather then 30) and that it is interrupted between the third and fourth repetition by 5 random bases in $x_{71,5}$, and by 3 random bases in $y_{71,5}$. In effect, a perfect alignment of the entire 51 base fragment is impossible, i.e., either only 'the 16 base random sequence + the 15 base pattern segment' can be matched, or only 'the second 20 base pattern segment' can be matched. As before, SPOMF outperforms MF in detecting the global misalignment (n=56), and is more sensitive than MF in detecting the local misalignment of the second pattern segment (n=58).

6. SUMMARY

We have proved that the prime length SPOMF greatly outperforms the standard MF when applied to alignment of DNA sequences containing repetitive patterns. Experiments (both on synthetic and real data, the later ones not included here due to lack of space) have shown that SPOMF is robust to random insertions, symbol mismatches and symbol deletions. These results combined with the well known computational efficiency of the cross-correlation methods (which can be still further improved via the use of binary SPOMF) lead us to believe that the prime number SPOMF could become one of the standard tools for DNA sequence analysis.

7. REFERENCES

[1] M. Abramowitz and I. Stegun (ed.), ”Handbook of mathematical functions”, Dover Publications, New York, 1972.

[2] D. Anastassiou, ”Genomic signal processing”, IEEE Trans. SP, Vol. 18, pp. 8-20, July 2001.

[3] A.K. Brodzik and O. Peters, ”Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences”, *ICASSP Proc.* pp 373-376, 2005.

[4] J. Butler, ”Forensic DNA typing: biology and technology behind STR markers”, Academic Press, 2003.

[5] F. Chan and E. Rabe, ”A non-linear phase-only algorithm for active sonar signal processing”, *OCEANS’97 Proceedings*, Vol. 1, pp 506-511, 1997.

[6] E.A. Cheever, G.C. Overton and D.B. Searls, ”Fast Fourier transform-based correlation of DNA sequences using complex plane encoding”, *Cabios*, Vol. 7, No. 2, pp 143-154, 1991.

[7] Q. Chen, M. Defrise and F. Deconinck, ”Symmetric phase-only matched filtering of Fourier-Mellin transforms for image registration and recognition”, *IEEE Trans. PAMI*, Vol. 16, No. 12, pp 1156-1168, 1994.

[8] A.L. Delcher *et al*, ”Alignment of whole genomes”, *Nucleic Acids Research*, Vol. 27, No. 11, pp. 2369-2376, 1999.

[9] D. Holste and I. Grosse, ”Repeats and correlations in human DNA sequences”, *Physical Review E*, 67, 2003.

[10] J.L. Horner and P.D. Gianino, ”Phase-only matched filtering”, *Applied Optics*, 23, 6, pp. 812-816, 1984.

[11] T. Kalker and A.J.E.M. Janssen, ”Analysis of watermark detection using SPOMF”, *ICIP Proc.*, Vol. 1, pp 316-319, 1999.

[12] E. S. Lander *et al*, ”Initial sequencing and analysis of the human genome”, *Nature*, Vol. 409, pp 860-921, February 2001.

[13] S.B. Needleman and C.D. Wunsch, ”A general method applicable to the search for similarities in the amino acid sequence of two proteins”, *J. Mol. Biol.*, Vol. 48, pp 443-453, 1970.

[14] S. Rajasekaran, X. Jin and J.I. Spouge, ”The efficient computation of position-specific match scores with the fast Fourier transform”, *J. Comp. Biol.*, Vol. 9, No. 1, pp 23-33, 2002.

[15] S. Tavare and B. W. Giddings, ”Some statistical aspects of the primary structure of nucleotide sequences”, in M. S. Waterman (ed.): Mathematical methods for DNA sequences (pp. 117-131), Boca Raton, CRC Press, 1989.

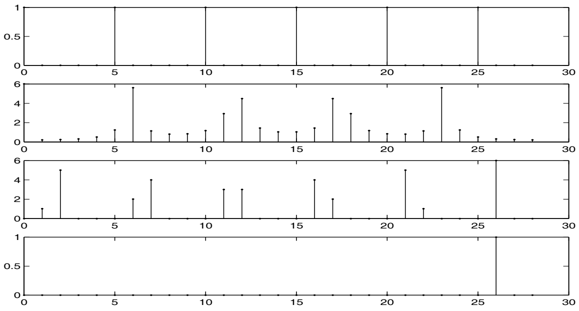

Figure 1: Single symbol periodic DNA sequence $x_{29,5,3}$, its DFT, MF, and SPOMF.

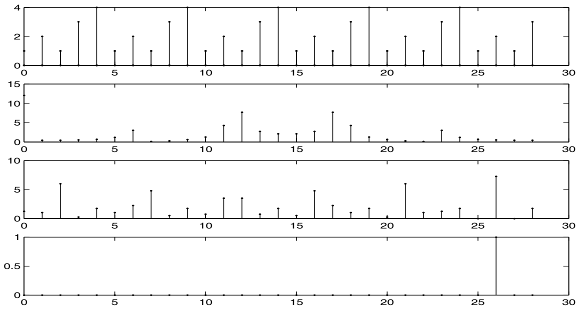

Figure 2: 4-symbol periodic DNA sequence $x_{29,5,3}$, DFT of the ’a’ sequence, MF, and SPOMF.

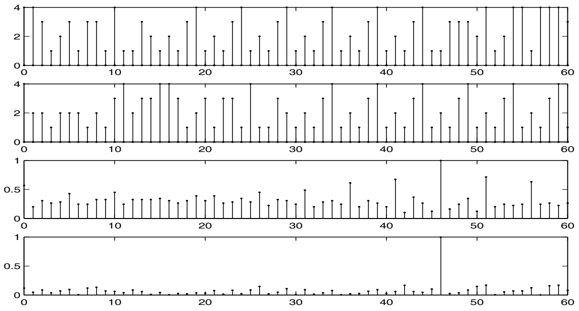

Figure 3: Semi-periodic DNA sequence $x_{61,5,15}$, misaligned DNA sequence, MF, and SPOMF.

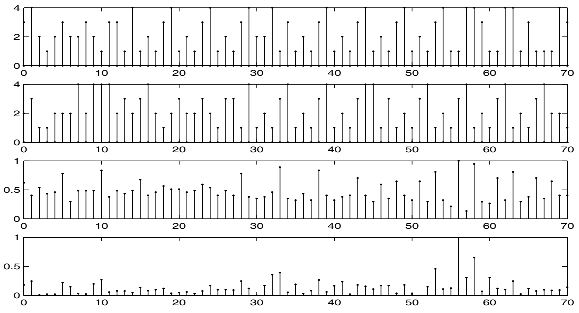

Figure 4: Semi-periodic DNA sequence $x_{71,5,15}$ with a gap, misaligned DNA sequence, MF, and SPOMF.