Phase-Only Filtering for the Masses (of DNA Data): a New Approach to Sequence Alignment

Andrzej K. Brodzik, *Senior Member, IEEE*

The MITRE Corporation
202 Burlington Rd, Bedford MA 01730
email: abrodzik@mitre.org
phone: 781.271.6992
fax: 781.271.2184 *

Abstract

Alignment of DNA segments containing repetitive nucleotide base patterns is an important task in several genomics applications, including DNA sequencing, DNA fingerprinting, pathogen detection, and gene finding. One of the most efficient procedures used for this task is the cross-correlation method. The main computations of the procedure are the discrete Fourier transform and a pointwise multiplication of two complex Fourier transform sequences. In this work the standard magnitude-and-phase cross-correlation technique is compared with the lesser known but closely related phase-only cross-correlation method. It is shown that for a periodic DNA sequence the standard approach leads to significant sidelobes in the cross-correlation, the magnitude of which increases with sequence length, while the phase-only approach yields a perfect cross-correlation with zero sidelobes. For a DNA sequence that contains both irregularly distributed symbols and periodic patterns the difference in performance is less pronounced, but still significant. Numerical experiments on synthesized and real data demonstrate that the phase-only approach is robust to isolated symbol insertions and deletions, and that it is capable of identifying positions of matching

[*] Approved for Public Release; Distribution Unlimited; Case : 06-0072

segments in the sequence.

Index Terms: DNA sequence alignment, DNA symbol repeat, cross-correlation, matched filter, phase-only filtering.

1 Introduction

DNA sequence alignment is one of the most important data processing tasks in computational genomics. The task appears in several applications, where efficient manipulation of large data records arranged in several different ways is required. In DNA fingerprinting [11], for example, an unknown collection of DNA fragments is acquired, typically few tens to few thousands of bases long. This unknown collection is then compared with one of several known collections of DNA fragments contained in a library. Either or both of these collections might be incomplete, unordered, or contain errors, including symbol insertions and symbol deletions. Finding a match between collections establishes genome identity.

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 often given. The pattern may be a part of pathogen signature or may indicate the start of a coding region. This relatively short sequence is then compared with a sequence that can be several millions of bases long. A match of the pattern with a specific region of the analyzed sequence confirms previous pathogen exposure or identifies an exon [45].

Related problems occur in comparative genomics and evolutionary tree reconstruction. The goal in these applications is to identify and align islands of similarity in two or more long DNA sequences [46], [36], [4]. Consensus between significant parts of sequences, including coding and conserved non-coding regions, indicates functional relationship or evolutionary proximity.

Many approaches to DNA sequence alignment have been proposed over the last two decades.

Among the best known are the Needleman-Wunsch (NW) algorithm [37], the Smith-Waterman (SW) algorithm [43], FASTA [33], BLAST [2], MUMmer [17], REPuter [29], and MAFFT [26]. NW, SW, and FASTA are based on dynamic programming; BLAST utilizes a heuristic search; MUMmer and REPuter rely on suffix trees; and MAFFT performs an FFT-based cross-correlation. Many other methods belong to one of these four groups. The methods vary in terms of the length of query sequence allowed, the degree of sequence similarity required, treatment of gaps, type of alignment (global or local), and speed and accuracy. While a detailed comparative study of the different procedures would be useful, this is beyond the scope of this work; for a glimpse at the state-of-the-art, the reader is referred to [38] and [35]. Since each of the methods has shortcomings, and since the amount of genomic data grows at a much faster rate than improvements in computing technology, investigation of new techniques that could deliver both computational efficiency and alignment accuracy continues to be an active area of research.

In this paper we investigate the cross-correlation approach to DNA sequence alignment [14], [39], [26], [44]. This choice is mainly motivated by the low computational complexity of the method, which is of the order of $M \log_2 N$ operations (where M and N are the lengths of library and query sequences). This is in contrast to $O(MN)$ operations required by most sequence alignment approaches, especially the ones relying on dynamic programming and suffix trees. The popular BLASTn and MegaBLAST algorithms (of the BLAST family) are reputed to be faster; however, complexity of these methods is difficult to evaluate since the number of computations required depends on many parameters, including the length of the seed sequence, the drop-off value for seed extension, treatment of gaps, sequence similarity, etc. [34], [31], [28]. Computational complexity of the cross-correlation approach, on the other hand, is easily quantifiable, it does not depend on the data, and it does not have inherent limitations on the sequence size. The approach is the basis of several existing algorithms, of which the best known is perhaps MAFFT [26]. Since efficiency of the cross-correlation approach is well established [14], [18], [26], [27], [39], we focus here on performance.

The two main defficiencies of the cross-correlation method are: (1) the inability to handle symbol insertions and deletions, and (2) poor performance when applied to the local alignment problem [26]. In this work we attempt to address these two problems. We focus our analyses on periodic sequences.

Although, in general, the distribution of symbols in a DNA sequence might appear to be irregular, many relevant genetic phenomena can be associated with occurrence of periodically-spaced DNA symbols. Well known examples include mutations and genetic diseases [45], [6], [41], start of a coding region [22], [42], 10.5-base repeats that are due to a 3.5 amino acid repeat in alpha-helical coiled-coil regions in proteins [47], transposon-derived Alu repeats [21], and interspersed repeats occurring in multi-species conserved sequences [46]. In fact, since DNA repeats are estimated to comprise more than one half of the human genome [30], most sequences of interest are likely to contain some periodic repetitions.

The main goal of this paper is to demonstrate the performance gain achieved in the analysis of periodic and semi-periodic DNA data by replacing the standard cross-correlation procedure with the lesser known phase-only cross-correlation approach. The first method is frequently referred to as a matched filter (MF), and the second one as a symmetric phase-only matched filter (SPOMF). We analyze the performance of both methods. First, using simple models of the DNA data, we show that the MF of a periodic sequence often does not produce a unique alignment, and that the computation of a SPOMF is unstable. We suggest a remedy in the form of a prime length cross-correlation. Subsequently, we prove that while this modified approach leads to unique alignment of identical sequences, MF produces large sidelobes arising from partial matches of shifts of the periodic sequence. Since magnitude of these sidelobes is proportional to the length of the sequence, MF of a long sequence produces sidelobes that obscure the main detection peak. Conversely, the use of a SPOMF yields a perfect cross-correlation with zero sidelobes in case of identical sequences, and a relatively artifact-free cross-correlation in case of sequences with moderate contaminations, thereby mitigating the aforementioned difficulties.

As an additional benefit of the analyses, removal of sidelobes allows identification of detection peaks corresponding to alignments of distinct sequence segments and enables construction of a new local alignment algorithm. A key feature of this algorithm is the ability to obtain positional information about the matching segments. It has been often observed that the cross-correlation sequence produced by the standard approach does not encode positional information about misaligned segments. In the case of a single matching segment, the alignment peak conveys information about how much the analyzed sequence needs to be shifted to produce segment alignment, but not about the position of the segment within the sequence. In case of multiple matching segments the difficulty is compounded, since sidelobes of the dominant segment compete with mainlobes of smaller segments. In [26], a remedy to this problem has been suggested in the form of the short time Fourier transform. Here we show that information about position of the segment can be easily extracted from the phase-only cross-correlation sequence in a single step (i.e., without splitting the analyzed sequence into smaller sections), by identifying consecutive symbol matches in the product of the aligned sequences.

The content of this paper is as follows: in Section II we introduce the SPOMF approach, in Section III we analyze periodic DNA sequences and motivate the use of prime length Fourier transform; in Section IV we state the main theoretical result of the paper that quantifies performance of MF and SPOMF, in Section V we perform numerical experiments on synthesized and real data, compare the two methods in robustness to symbol contaminations, and outline the new local alignment algorithm, and in Section VI we give a brief comparison of SPOMF with BLAST.

2 MF and SPOMF

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

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

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

$$\mathbf{x}(k) = \mathrm{DFT}\{x(n)\} = \sum_{n=0}^{N-1} x(n)e^{2\pi ink/N}, \quad 0 \leq k < N. \tag{2}$$

Since

$$\mathbf{z}(k) = \mathbf{x}(k)\bar{\mathbf{y}}(k), \tag{3}$$

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

$$z(n) = \mathrm{DFT}^{-1}\left\{\mathbf{x}(k)\bar{\mathbf{y}}(k)\right\}, \tag{4}$$

where computation of each DFT requires $N \log_2 N$ operations.

Recently, a different procedure, known as the symmetric phase-only filter (SPOMF), has been introduced and found useful in several applications. The SPOMF of two discrete sequences x and y is given by the formula

$$w(n) = \mathrm{DFT}^{-1}\left\{\frac{\mathbf{x}(k)\bar{\mathbf{y}}(k)}{|\mathbf{x}(k)\bar{\mathbf{y}}(k)|}\right\}, \quad \mathbf{x}(k) \text{ and } \bar{\mathbf{y}}(k) \neq 0. \tag{5}$$

SPOMF had been proposed two decades ago in optical signal processing [23], 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 [15], watermarking [25], and sonar [13], among others. Several flavors of the basic formulation (5) of SPOMF have been suggested in literature, including a version where the norm in the denominator is taken with a fractional power and the reduced complexity binary and ternary filters [24].

3 Regular periodic DNA sequences

DNA sequence is a symbolic string of characters 'a', 'c', 'g', and 't' denoting the four nucleotides that make up the genetic code: adenine, cytosine, guanine, and thymine. Various methods of mapping a symbolic DNA sequence to a numeric sequence have been proposed, including the use of complex and hypercomplex number systems [3], [9], [14]. For the sake of simplicity, but without a loss of generality, in the next two sections we will consider only single-symbol DNA sequences represented by binary numbers. Furthermore, we will only consider periodic sequences (in the sense specified below). In Section V we apply the developed formalism to the 4-symbol DNA data and discuss processing of semi-periodic and 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 shifted by S is an N-point sequence

$$x_{N,P,S}(n) \quad = \quad \begin{cases} 1, & (n-S) \equiv 0 \pmod{P}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

We have $x_{N,P,S} = x_{N,P,S \bmod P}$. 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 $\bar{P} = N/P \in Z,\quad 0 \leq S < P$. 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

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

We are now ready to compute the MF and SPOMF of regular combs.

Theorem 2 The MF of a regular P-periodic comb $x_{N,P,S}$ is a P-periodic comb

$$z(n) = \begin{cases} \bar{P}, & (S-n) \equiv 0 \pmod{P}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Proof Using theorem 1 and equation (4) we have

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

which leads to (7). □

Theorem 2 shows that the MF of a regular comb provides a measure of DNA sequence misalignment, provided $S < P$, which is of limited use. Due to theorem 1 the SPOMF of a regular comb is not defined, since insertion of $\mathbf{x}_{N,P}$ into (5) results in division by zero. One way to circumvent this problem is to assign to the 'divide by zero' points some small fixed value. Like the MF however, SPOMF may provide a measure of DNA sequence misalignment only when $S < P$. Moreover, there is a performance penalty associated with zeroes of the DFT of a comb, manifesting itself in sidelobes of the cross-correlation sequence. As will be shown in the next section, both obstructions will 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 shifted by S is an N-point sequence

$$x'_{N,P,S}(n) \quad = \quad \begin{cases} 1, & (n-S) \equiv 0 \pmod{P}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

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].[1] Moreover, as with the power-of-two length DFT, there are fast algorithms for computing a prime length DFT. We will 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

$$\mathbf{x'}_{N,P,S}(k) = \begin{cases} \lfloor \frac{N}{P} \rfloor + 1, & k = 0, \\ \frac{1-\exp(2\pi i k P(\lfloor N/P \rfloor + 1)/N)}{1-\exp(2\pi i k P/N)} e^{2\pi i k S/N}, & \text{otherwise,} \end{cases} \tag{9}$$

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

Proof We have

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

which leads to (9). □

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

[1] Distribution of primes does become sparser for very large numbers. However, it's the relative, not absolute, increase in the sequence size that matters (i.e., $\frac{N_0 - N}{N_0}$, where N_0 is the size of the data and $N \geq N_0$ is the nearest prime). Moreover, the analysis of DNA sequences is typically restricted to sequence segments of length $N < 10^6$.

Proof Follows directly from theorem 3. □

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

Theorem 4 Set $\delta = \lfloor \frac{N}{P} \rfloor + 1$. The MF z' of an irregular P-periodic comb $x'_{N,P,S}$ is given by[2]

$$z'(n) = \frac{\delta^2}{N} + \frac{1}{N} \sum_{k=1}^{N-1} e^{2\pi i k(S-n)/N} \frac{1-\cos(2\pi kP\delta/N)}{1-\cos(2\pi kP/N)} \tag{10}$$

The mainlobe of z' is given by

$$\mathcal{M} = z'(n=S) = \delta, \tag{11}$$

and the largest sidelobe of z' is given by

$$\mathcal{S} = z'(n=S+P) = \delta - 1. \tag{12}$$

Proof (10) follows directly from inserting $\mathbf{x'}_{N,P,S}$ and the conjugate of $\mathbf{x'}_{N,P}$ (theorem 3) into (4). (11) follows from (1) and (8). To obtain (12) take

$$\begin{aligned} \mathcal{M} - \mathcal{S} &= \frac{1}{N} \sum_{k=1}^{N-1} (1 - e^{-2\pi ikP/N}) \frac{1-\cos(2\pi k\delta P/N)}{1-\cos(2\pi kP/N)} \\ &= \frac{1}{N} \sum_{k=1}^{N-1} [\mathcal{R}(k) + i\mathcal{I}(k)], \end{aligned}$$

where

$$\mathcal{R}(k) = 1 - \cos(2\pi k\delta P/N)$$

[2] Incidentally, this result suggests several surprisingly non-trivial trigonometric sum evaluations that seem to be related to the Gaussian sum. The simplest of these sums is

$$\sum_{k=1}^{N-1} \frac{1}{\cos(2\pi k/N)} = \begin{cases} N-1, & N \equiv 1 \pmod 4, \\ -N-1, & N \equiv 3 \pmod 4, \end{cases}$$

where N is an odd integer, $N \geq 3$. These evaluations will be addressed in a separate work.

and

$$\mathcal{I}(k) = -\sin(2\pi kP/N)\frac{1-\cos(2\pi k\delta P/N)}{1-\cos(2\pi kP/N)}.$$

Since $\mathcal{I}(k) = -\mathcal{I}(N-k)$ for $1 \leq k \leq (N-1)/2$, then $\sum_{k=1}^{N-1}\mathcal{I}(k) = 0$. Moreover, we have

$$\sum_{k=1}^{N-1}\mathcal{R}(k) = \sum_{k=1}^{N-1}[1-\cos(2\pi k\delta P/N)] = N.$$

In effect $\mathcal{S} = \mathcal{M} - (\mathcal{M} - \mathcal{S}) = \delta - 1$. □

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

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

$$\frac{z'(S)}{z'(S+P)} = \frac{\delta}{\delta-1}. \tag{13}$$

The well-known next result characterizes the performance of SPOMF of an arbitrary non-trivial prime length binary sequence, which includes the case of an irregular P-periodic comb.

Theorem 5 Take N to be an odd prime number and S to be an integer, $0 \leq S < N$. The SPOMF of an N-point binary sequence, shifted by S, that is not all-zero or all-one, is given by

$$w'(n) = \begin{cases} 1, & n = S, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Proof The conditions $x(n)$ is not all-zero or all-one and N is an odd prime guarantee that $\mathbf{x}(k) \neq 0$ $\forall_k$. Then it follows from (5) and from the shift property of the Fourier transform that

$$w'(n) = \frac{1}{N}\sum_{k=0}^{N-1} e^{-2\pi ikn/N}\frac{\mathbf{x}(k)\bar{\mathbf{y}}(k)}{|\mathbf{x}(k)\bar{\mathbf{y}}(k)|} = \frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi ik(S-n)/N} = \begin{cases} 1, & n = S, \\ 0, & \text{otherwise. } \square \end{cases}$$

As can be seen from theorems 4 and 5, when the DNA sequence is an irregular P-periodic comb, then the SPOMF significantly outperforms the MF. While the SPOMF yields a perfect cross-correlation sequence, the MF gives rise to sidelobes, the largest of which approaches the magnitude of cross-correlation mainlobe as N increases and P remains constant. Conversely, decreasing the length of the P-periodic comb and maintaining a constant period reduces the sidelobes of the MF cross-correlation sequence. Since a real DNA sequence is never purely periodic, and many sequences contain a limited number of symbol repeats (although some repeats are always present in a sufficiently long sequence due to the roughly equal distribution of symbols a, c, g, and t, and the three-base encoding of amino acids in exons), a question arises as to how the performance of a MF algorithm scales with the decreasing content of periodic symbols. This question is not easy to answer, in part because of the difficulty of defining the opposite of a periodic sequence. Although many authors contrast occurrence of patterns with random symbol distribution, this is not a true dichotomy; moreover, it is questionable whether the distribution of symbols in a DNA sequence can ever be random. A more suitable definition of a non-periodic binary sequence might include a rule requiring the difference between positions of any two symbols to be unique. An additional constraint could be added that this set of differences be exhaustive.[3] We can state this more formally as follows.

Definition 3 An N-point binary sequence $x_0, ..., x_i, ..., x_j, ..., x_{N-1}$ having $2 \leq L < N$ non-zero elements is non-periodic, iff for all distinct pairs of non-zero elements x_i and x_j, $i \neq j$, the set of values $(i - j) \pmod N$ is identical with the set of integers $1 \leq k \leq N$.

Definition 3 is in fact used in the construction of special sequences having an ideal two-valued

[3] Less restrictive criteria of randomness of a genomic sequence are possible to state, including the well-known Lempel-Ziv complexity measure [32], however, the construction described here is unique in the sense that it removes *all* periodic behavior.

auto-correlation,

$$z'(n) = \begin{cases} L, & n = 0, \\ \\ 1, & \text{otherwise.} \end{cases} \tag{15}$$

These special sequences are known as modular Golomb rulers in communications [7], and as cyclic difference sets (CDS) in combinatorics [5].[4] Simple examples of CDS include the sequences '0110100' and '1101000001000'. In general CDS are not easy to obtain, and for many sequence sizes CDS do not exist. For example, the CDS given above are the only ones for a seven- and a thirteen-base sequence. The scarcity of CDS implies that perfectly irregular DNA sequences are relatively rare, and therefore most DNA sequences can be considered semi-periodic.

Equipped with the concept of CDS as a model for a non-periodic sequence, the two cross-correlation approaches can be more easily compared. From theorem 5, the SPOMF of any binary sequence (including periodic and non-periodic sequences), for which it can be defined, yields a perfect cross-correlation. Assuming a fixed frequency of DNA symbols, the MF of a periodic sequence yields a cross-correlation with the ratio $\mathcal{M}/\mathcal{S}$ tending to one, as N increases. In contrast, the MF of a non-periodic sequence yields a two-valued cross-correlation, with the ratio $\mathcal{M}/\mathcal{S}$ tending to infinity as N increases. When a sequence is neither periodic nor non-periodic, it is called semi-periodic. In general a semi-periodic sequence might contain multiple segments, some periodic and some irregular. In a simplified model a semi-periodic sequence will contain only two segments: a truncation of a periodic sequence and a cyclic difference set (when it exists). Since performance of the MF of a non-periodic sequence is optimal, cross-correlation of a semi-periodic sequence can be expected to depend mainly on the length of its periodic component. [5] This prediction has been confirmed by experiments (not included here due to lack of space); the results indicate that sidelobes of the MF decrease when sequences become more irregular, and that performance of the two methods becomes comparable when sequences have no obvious periodic component.

[4] The author is grateful to Joe Rushanan for bringing the concept of Golomb rulers to his attention.

[5] A rigorous determination of the MF of a semi-periodic sequence would require evaluation of the product of Fourier transforms of the two components. This evaluation will be given in a sequel.

5 Comparison of SPOMF and MF

In this section we give examples of SPOMF and MF of several synthesized and real DNA sequences of increasing complexity. To illustrate the theoretical results of the previous section, we begin with cyclicly shifted, purely periodic sequences. Later, we investigate cross-correlations of linearly shifted semi-periodic sequences, and test robustness of the SPOMF approach to single and multiple symbol insertions and deletions, and symbol mismatches.

Since the intention here is to provide a proof of concept for the SPOMF approach, and to illustrate the key aspects of the alignment problem in a manner that facilitates visual inspection, most examples are limited to relatively short sequences. The small size of the sequences does not limit the generality of the examples, since in practice DNA data is often processed in segments of a few hundred to a few tousand bases at a time. The examples include moderately homologous sequences (50%-100% homology), that are typical in many applications, including cross-species sequence comparisons [19]. In the later part of subsection 5.2 and in subsection 5.3, where the construction of a novel local alignment algorithm is discussed, two examples of more complex DNA sequences are given, the second one having no significant periodic component.

The implicit focus of this paper up to now has been on global sequence alignment. This section will begin with examples of a global alignment, and then it will gradually progress toward the more general case (in the sense explained below) of a local alignment. Since the terms ‘global’ and ‘local’ are often used rather casually in the literature, a brief discussion of the two concepts is included.

The goal of the global alignment is to maximize the total number of symbol matches in the two sequences, regardless of the relative position of symbols within the sequence. For example, when sequences are long and not closely related, then the matching symbols are likely to be distributed throughout the sequence. In contrast, the goal of the local alignment is to identify a smaller section of the sequence that contains a large number of matching symbols, even if the number of matching symbols in that section is less than the number of matching symbols in the entire sequence (in the global alignment). In general, there might be more than one region of this type, and the distance between these regions in two sequences might be not preserved. In such a case multiple local alignments need to be performed. Moreover, in both local and global alignments symbol insertions or deletions (the so-called indels) in the sequences might be allowed, thereby blurring the distinction between the two alignments.

For the purpose of this text we will adopt a narrow definition of the local alignment. We will assume that indels can occur between matching regions, but not within a region, and that matching symbols within a region are consecutive. We believe that this condition is not necessary for our approach to be effective; however, the assumption will significantly simplify analysis of the results, and we deem it is reasonable, since our focus is on sequences with a significant content of periodic motifs.

5.1 Synthesized data

Example 1

In the first experiment the two approaches were applied to a composite, four-symbol periodic sequence, consisting of five individual combs, $x_{29,5}$. The constituent a, c, g, and t sequences were chosen to create a repetitive pattern '*acagt*' (plot one in Figure 1). The query sequence was identical to the cyclicly shifted by 3 symbols library sequence, $y_{29,5} = x_{29,5,3}$ (plot two in Figure 1; the indices of x are as in definition 1). The symbols '*a*', '*c*', '*g*', and '*t*' were marked in the plots with stems equal to 1, 2, 3, and 4, respectively. The cross-correlations shown (plots three and four in Figure 1) are the sums of cross-correlations performed on individual symbol sequences.

The use of SPOMF produced a perfect cross-correlation sequence with no sidelobes. The use of MF produced a cross-correlation sequence with a peak at $n = S = 3$, equal to $\delta = \lfloor \frac{29}{5} \rfloor + 1 = 6$, and multiple sidelobes. The largest sidelobe was equal to $\delta - 1 = 5$, and occured at $n = S + P = 8$, as predicted by theorem 4. The magnitudes of MF and SPOMF sequences shown in Figure 1 (and subsequently) were normalized to facilitate visual comparison of results.

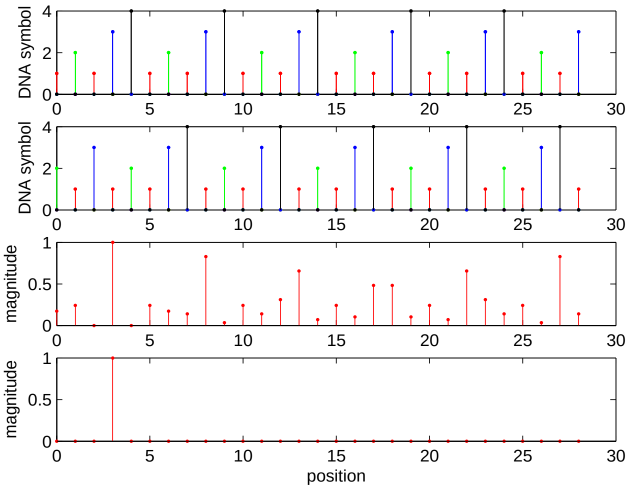

Figure 1: Efficacy of the two alignment approaches in application to a four-symbol periodic sequence. Top to bottom: the composite DNA sequence $x_{29,5}$, the misaligned by 3 symbols DNA sequence $y_{29,5}$, MF, and SPOMF.

Example 2

The first experiment have shown an example of sequence alignment performed on purely periodic, perfectly matched data. In practice, in addition to periodic motifs, DNA sequences often contain segments of irregularly distributed symbols and segments that do not match. Figure 2 shows an example of such a case. The two misaligned sequences, x_{61} and y_{61}, share a 46-base segment, the first 16 bases of which are random, while the remaining 30 bases contain a repetitive pattern '*acagt*'. The library sequence, x_{61}, is appended by a random 15-base segment, and the query sequence, y_{61}, is concatenated to a different random 15-base segment. In effect, x_{61} and y_{61} are misaligned by 15 bases, and mismatched (when aligned) at 15 bases. The cross-correlation peak in both MF and SPOMF occurs at $n = S = 15$. While due to the 15-base segment mismatch the SPOMF sequence does not produce an ideal cross-correlation sequence in this case, its sidelobes are significantly smaller than the sidelobes of the MF (the largest sidelobe of SPOMF equal to

0.238 occurs at $n = 39$, and the largest sidelobe of MF equal to 0.736 occurs at $n = 10$).

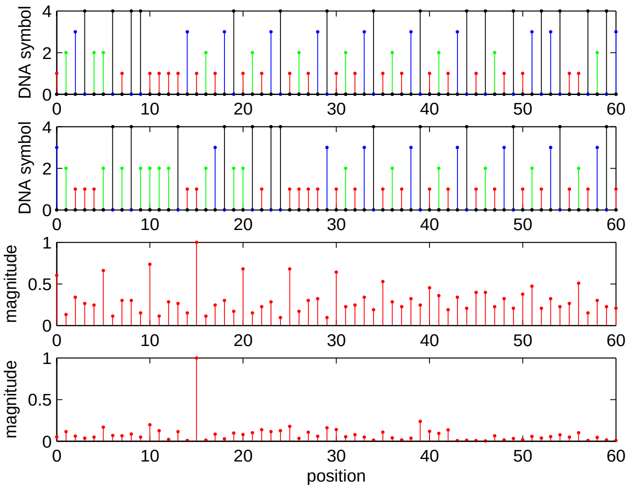

Figure 2: Alignment of semi-periodic DNA sequences containing a repetitive pattern, random symbols, and a mismatching segment. Top to bottom: the synthesized DNA sequence x_{61}, the misaligned by 15 symbols DNA sequence y_{61}, MF, and SPOMF.

5.2 Real data

In the second part of this section we describe experiments performed on two real DNA sequences selected from GenBank: *Mus musculus* BAC clone RP23-1I16, locus AC098708, bp 46101:46502 (x_{401} and y_{401}), and *Mus musculus*, chromosome 9, locus AC103610, bp 143101:144114 (x_{523} and y_{523}). Both DNA sequences contained about 200-base long segments of three to six bases long repetitive patterns of two (a and g) or three (a, c and g) symbols. Insertions and deletions were not part of the original data, but were induced artificially. MF and SPOMF computations in all experiments were performed using a linear cross-correlation of length that was twice the length of the analyzed sequence; however, to fascilitate visual comparisons only the relevant half of the cross-correlation sequence samples is shown in the plots.

Example 3

In the first experiment of this subsection we have analyzed two linearly shifted sequences, x_{401} and $y_{401} = x_{401,170}$, derived from the GenBank sequence AC098708. The sequences contained 231 contiguous matching symbols, including a 190-base repetitive pattern '*accagg*', and 170 mismatching symbols. No deletions or insertions were applied to the sequences. The MF approach produced a cross-correlation sequence having the ratio of magnitudes of the largest sidelobe (occuring due to a partial match of the shifted pattern) and the mainlobe equal to 0.69. The SPOMF approach produced a cross-correlation sequence having the ratio of magnitudes of the largest sidelobe (occuring due to a 170-base segment mismatch) and the mainlobe equal to 0.19 (Figure 3).

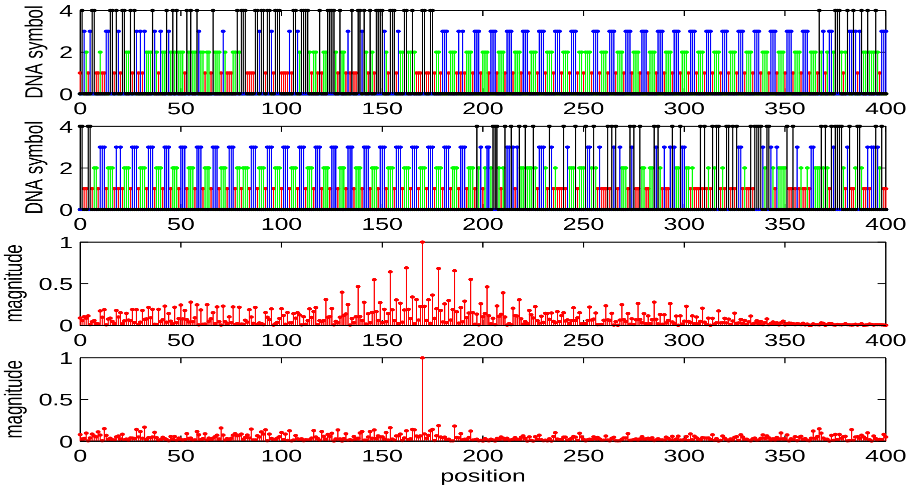

Figure 3: Alignment of semi-periodic contiguous DNA sequences containing a 190-base repetitive segment, GenBank AC098708. Top to bottom: the DNA sequence x_{401}, the misaligned by 170 symbols DNA sequence y_{401}, MF, and SPOMF.

Example 4

In the previous experiment the matching symbols in the two sequences were contiguous. In many practical applications the matching segments may contain contaminations, including symbol insertions, deletions and substitutions. This situation is addressed in this experiment.

The sequences, x_{401} and y_{401} (plots one and two in Figure 4), are similar to the ones used in the previous experiment, except for two modifications. First, the second sequence was shifted by 70 symbols rather than by 170 symbols with respect to the first sequence ($y_{401} = x_{401,70}$). This produced a matching segment containing a significant number of both repetitions and irregular symbols. Second, the sequence x_{401} was modified by a deletion of three symbols at bp 177:179 (the number of consecutive deletions is irrelevant here, except for providing a more convenient display of alignment results). The deletion effectively split the library sequence x_{401} into two segments. The shorter one (105 bases) was comprised of a largely random DNA symbol assembly. The longer one (220 bases) was comprised mostly of a complex repetitive motif with a predominance of symbols 'c' and 'g'.

In effect, a perfect (global) alignment of the entire 331 base segment was impossible, i.e., either only the sequence segment prior to the deletion could be matched (in one local alignment), or only the sequence segment following the deletion could be matched (in another local alignment). To identify both segments, the cross-correlation sequence needs to produce two distinct peaks. Both approaches do produce two such peaks, one at $n = 70$ and one at $n = 73$. However, while the two peaks can be clearly identified in the SPOMF plot (magnitudes 0.9 and 1.0, plot four in Figure 4), in the MF plot the peak corresponding to the second alignment is much smaller (magnitude 0.19, plot three in Figure 4), and is obscured by multiple large sidelobes.

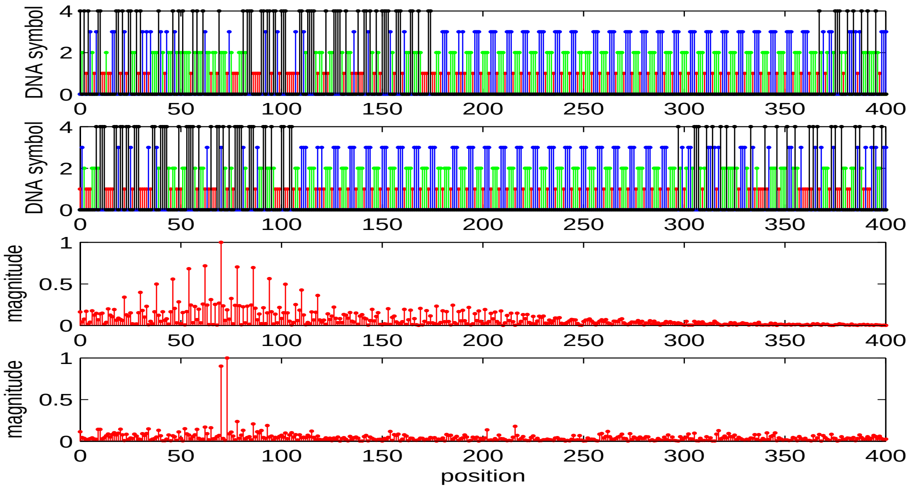

Figure 4: Alignment of semi-periodic DNA sequences containing a 190-base repetitive segment preceded by a symbol deletion, GenBank AC098708. Top to bottom: the DNA sequence x_{401}, the DNA sequence y_{401} with two segments misaligned with respect to x_{401} by 70 and 73 symbols, MF, and SPOMF.

Example 5

In the third experiment a more elaborate case of non-contiguous matching segments was considered. Two linearly shifted sequences, x_{523} and $y_{523} = x_{523,60}$, derived from the GenBank sequence AC103610, were used. The matching 463 bases included both repetitive and irregular symbols. While in example 4 only one of the sequences had an insertion, here each sequence contained a unique insertion: x_{523} had a 5-symbol insertion at $n = 21$, and y_{523} had a 5-symbol insertion at $n = 401$. As in experiment 4, the insertions had an effect of splitting the sequences into two matching segments. Unlike in experiment 4, however, one of the segments contained a gap. The case can be best explained symbolically. Take as the first sequence the string '$a_1xbbba_2a_2$' and as the second sequence the string '$a_1bbbxa_2a_2$', where x denotes deletion, and a_1, a_2 and b are arbitrary symbols. Both sequences contain two identical segments, i.e., 'bbb' and '$a_1....a_2a_2$'. However, the second segment is not contiguous. This is illustrated in more detail in Figure 6. The first segment (in the middle of the sequence, marked with a bar in the second plot), is comprised mostly of periodic repetitions of symbols 'a' and 'g'. The second, composite segment (having one component at the beginning of the sequence, and another component in the later part of the sequence, marked with bars in the third plot), is comprised of an assembly of mostly irregular DNA symbols.

Alignments of both segments can be easily discerned in the SPOMF sequence (detection peaks at $n = 55$ and at $n = 60$ of magnitude 1.0 and 0.843, respectively, in plot four of Figure 5). Note, that SPOMF detects both segments, even though the second segment is non-contiguous and non-periodic. In contrast, only one of the alignments can be identified in the MF sequence (a detection peak at $n = 55$ in plot three of Figure 5). The second alignment in the MF sequence (at $n = 60$, equal to 0.531) is obscured by sidelobes. Moreover, the MF produces an anomalous correlation peak (the second largest in the correlation sequence, at $n = 43$), that corresponds to a shift of the first segment (plot four of Figure 6).

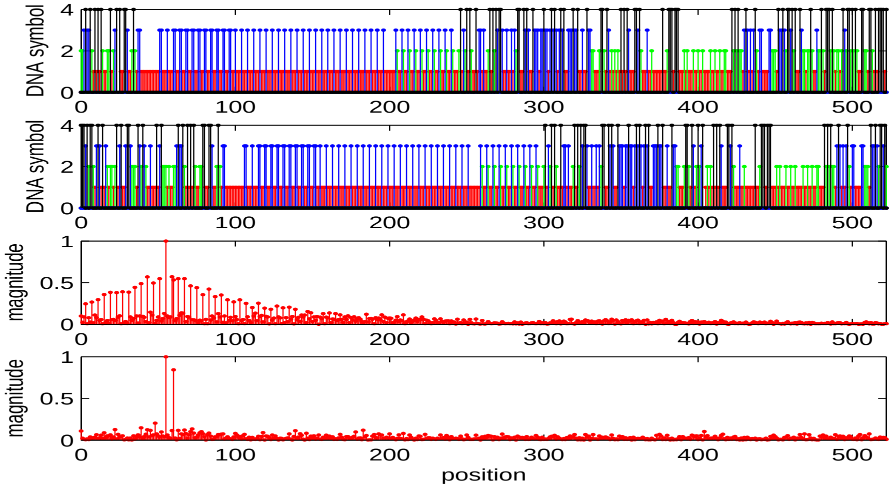

Figure 5: Alignment of semi-periodic DNA sequences with symbol insertions, GenBank AC103610. Top to bottom: the DNA sequence x_{523} with an insertion, the DNA sequence y_{523} with a distinct insertion, MF, and SPOMF. Detection peaks at $n = 55$ and $n = 60$ in the SPOMF plot identify the two matching segments.

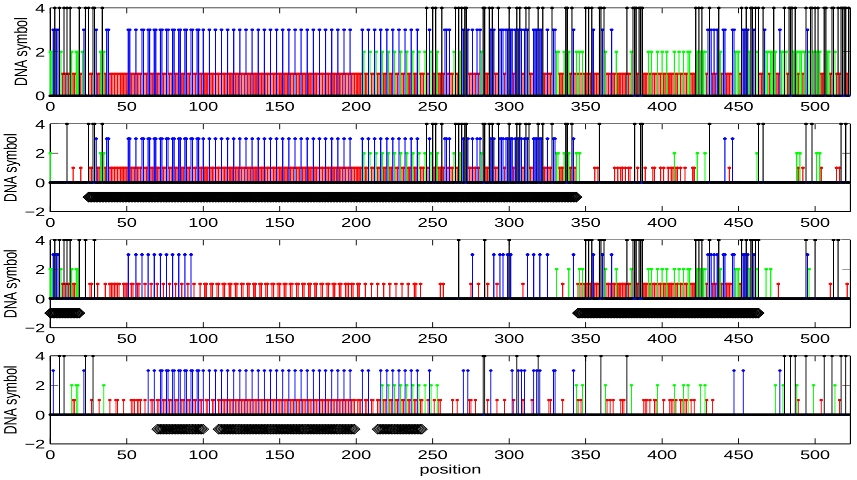

Figure 6: Alignment of semi-periodic DNA sequences with symbol insertions, GenBank AC103610. Top to bottom: the DNA sequence x_{523}, and the three matching subsequences of y_{523}. The subsequences correspond to the two largest peaks of SPOMF and the two largest peaks of MF (since the dominant peaks in the two cross-correlations coincide, and the second largest ones do not, hence three subsequences). The subsequences are shifted by $n = 55$ (MF and SPOMF dominant peaks, second plot of Figure 5), by $n = 60$ (SPOMF second peak, third plot of Figure 5), and by $n = 43$ (MF second, anomalous peak, fourth plot of Figure 5). Bars denote matching segments of x_{523}.

5.3 Local alignment

The results of the experiments warrant several observations. The robustness of the phase-only method to partial matches of shifts of periodic DNA segments results in the removal of sidelobes and an improved readability of the SPOMF cross-correlation plot. Removal of sidelobes is particularly important in the analyses of multi-component sequences. Multi-components occur, e.g., when a contiguous matching segment becomes contaminated by a symbol deletion or insertion. The segment may then be partitioned into several components, each requiring a separate local alignment. Since the SPOMF sequence does not produce anomalous partial match sidelobes, peaks of the SPOMF cross-correlation sequence can be associated with these local alignments.

The decomposition of y_{523} in the last experiment into two *almost* orthogonal subsequences (Figure 6, plots two and three) corresponding to the two distinct peaks of the SPOMF sequence (Figure 5, plot four) suggests that not only can local alignment of the individual segments be performed, but positional information about the individual segments within a sequence can be extracted as well. An outline of one such possible procedure tuned to the last example is given below.

Local alignment algorithm:

- Normalize sequences x and y, for each of the four symbols independently,
- Form auxiliary sequences $y_1(n) = y(n - S_1)$ and $y_2(n) = y(n - S_2)$, where S_1 and S_2 are sequence shifts corresponding to locations of the two distinct peaks in the SPOMF sequence,
- Compute the point-wise products $z_1 = xy_1$ and $z_2 = xy_2$,
- Identify location of individual segments with contiguous non-negative subsequences of z_1 and z_2,
- Align segments identified in the previous step by shifting them by S_1 and S_2.

Normalization replaces the assignment of binary values, marking symbol occurrence at a given position in a sequence, with an assignment of some positive/negative values (that depend on the total count of symbols in the sequence). If symbols match, then all four sequences, z_a, z_c, z_g, z_t, are positively valued. Conversely, if symbols mismatch, then two of the four sequences are negatively valued. In effect, detection of a symbol mismatch is equivalent to the identification of a negatively valued sequence. Note, that the normalization step is included here to facilitate visual evaluation of the alignment, but is not essential in the algorithm.

Results of the processing are illustrated in Figure 7. An important advantage of this procedure is that it does not rely on windowing, as does the method described in [26], and therefore resolution of the positional information is not limited by the window width. Further details of the implementation of the algorithm will be given elsewhere.

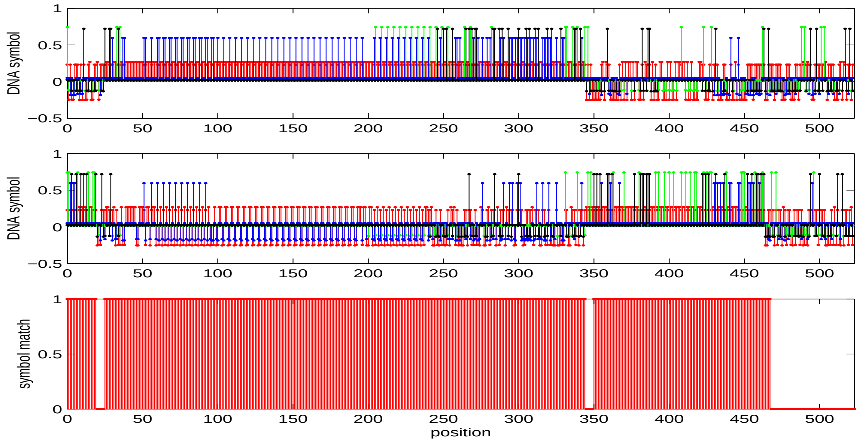

Figure 7: An illustration of the local alignment procedure. The first two plots show the auxiliary sequences z_1 and z_2. The non-negative values mark the locations of the two constituent segments of the DNA sequence x_{523}. Segment 1 (the top plot) corresponds to the SPOMF peak at $n = 55$ in Figure 5. Segment 2 (the middle plot) corresponds to the SPOMF peak at $n = 60$ in Figure 5. The bottom plot shows the composite alignment of the two sequences. The first gap in the plot is due to the symbol insertion in x_{523}, the second gap is due to the symbol insertion in y_{523}, and the segment of sixty zeros at the end of the sequence reflects sequence mismatch due to the shift by 60 symbols.

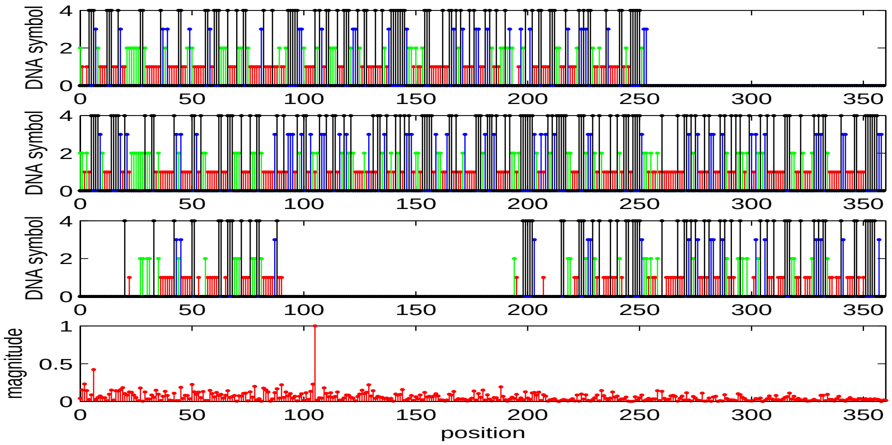

Figure 8: Local alignment of two homologous fruitfly sequences. From top to bottom: sequences J01125 and V00225, matching segments of the first sequence produced by the local alignment algorithm, and the SPOMF sequence. The SPOMF approach produces detection peaks at $n = 6$ and at $n = 105$, corresponding to the locations of matching segments in the second sequence.

Example 6

In the last experiment we have performed an alignment of two homologous sequences of satellite DNA molecules from the fruitfly. The data has been described in [12] and analyzed in [20]. The example demonstrates the efficacy of the phase-only algorithm in the alignment of two- and three-segment weakly semi-periodic DNA sequences, with 76% homology in the two matching segments. The first sequence, GenBank J01125, was 253 bp long, and contained two segments. The second sequence, GenBank V00225, was 359 bp long, and contained three segments. The first and last segments of both sequences were roughly matched. No significant periodic component was present in either sequence. Results of the alignment are shown in Figure 8. The SPOMF approach produced two peaks in the cross-correlation sequence corresponding to sequence shifts $S_1 = 6$ and $S_2 = 105$ (bottom plot in Figure 8). The third plot in Figure 8 shows (appropriately shifted by S_1 and $S_1 + S_2$) the matching segments of the first sequence, produced by the local alignment algorithm. The first segment starts at base 27 and ends at base 90 (relative to the second sequence). The second segment starts at base 215 and ends at base 357. Neither of the two segments of the first sequence matches exactly the segments of the second sequence; in both cases there are multiple single and double symbol mismatches. Moreover, the matching segments are preceded by 20- and 25-base long mini-segments that match the second sequence very poorly (the matching strings '*t.a*' and '*ca..tttttg...a*', respectively). Despite the relatively large number of mismatching symbols in the two segments, the SPOMF approach succeeds in producing a local alignment of the two sequences.

6 Comparison of SPOMF and BLAST

The validation part of this work focused on performance comparison of the proposed approach with the closely related matched filter technique. The later was previously considered by several researchers [14], [18], [26], [27], [39], and its advantages and disadvantages vis à vis the standard methods were discussed. It has been acknowledged that the principal advantage of the Fourier-based approach is its computational efficiency. Moreover, the approach: (1) is easily adoptable to include certain frequently performed sequence manipulation tasks, such as symbol repetition detection and filtering [14], (2) can re-use computations from the identical base search in the search of complementary bases [14], and (3) can be extended to an efficient multiple sequence alignment procedure [26], [39]. On the other hand, the utility of the Fourier-based approach is limited by a number of obstructions. These include the inabilities: to align sequences with gaps, to differentiate between contiguous and non-contiguous patterns, and to produce a local alignment. For a detailed discussion of these issues, the reader is directed to [14], [18], and [39].

While these obstructions are inherent in the direct implementation of MF, they can be mediated by various adaptations of the algorithm. Recently, Cheever *et al* [14], Rajasekaran *et al* [39], and Katoh *et al* [26] have shown that the last two of these problems can be partly overcome by windowing the query sequence. In this paper a different approach was proposed, based on the phase-only filter. It was shown that much of the difficulty in obtaining a local alignment derives from the occurrence of ambiguous sidelobes in the cross-correlation sequence. Subsequently, it was demonstrated that the phase-only filter substantially reduces these sidelobes, thus improving readability of the global alignment plot, and facilitating the local alignment procedure. The phase-only approach can be used alternatively or complementarily to the windowing approach of Cheever. While other problems associated with the Fourier-based approach, including the treatment of gaps, remain unsolved, it is hoped that this result will stimulate further research.

The differences between methods have been previously discussed in computational molecular biology literature; however, since the Fourier-based approach is not, in general, as well-known as some other sequence alignment techniques, especially the industry standard, BLAST, a limited comparison of the two methods can be useful. This is not an easy task. First, the two approaches are based on two very different philosophies. BLAST uses a heuristic search to identify local matches, evaluates symbol similarity according to statistical significance, and focuses on short strings, which are subsequently extended to longer segments. MF and SPOMF, on the other hand, exhaustively search for the best global match, do not, in general, use statistical criteria in ranking of alignment scores, and perform computations of all alignments at once. Second, both resolution and efficiency of BLAST, while considered to be well-balanced, when compared to other sequence alignment tools, depends on many parameters *and* on the data. Hence theoretical computational complexity bounds for BLAST are difficult to obtain, and results of experimental evaluations can be misleading. Third, while BLAST is a mature tool, the phase-only filter is still in a prototype stage. In effect, a comparison of the two approaches has to be limited and the results tentative.[6]

Nevertheless, we have attempted to draw some inferences about the relative advantages and disadvantages of the two methods. Since the performance issues were addressed in previous section, we decided to focus on computational efficiency.

We have chosen from the BLAST family the fastest program, called MegaBLAST, which has been designed for a rapid comparison of large, closely related sequences. Furthermore, also for efficiency, we have set the MegaBLAST parameters to perform unfiltered (-F F) and ungapped (-g F) search. Since BLAST, unlike the phase-only filter, does not compare two sequences directly in its entirety, but instead selects a short, adjustable-length fragment from the query sequence, called a seed, running BLAST requires specifying the seed length. We have chosen the two most often used seed sizes, 11 and 28 (the later being the longest allowed) [28]. Seed size 28 delivers a faster but less sensitive search than seed size 11, since the matching pattern in the library sequence

[6] A comparison of a Fourier-based approach with other standard tools (ClustalW and T-Coffee) have been given by Katoh [26] in the context of multiple sequence alignment.

needs to be at least 28 symbols long. For example, when applied to the sequences analyzed in experiment 6 of the previous section, MegaBLAST-11 identified both matching segments, while MegaBLAST-28 detected only the longer, second segment. We have included both versions in our comparison for completeness, however we found that sensitivity did not play a major role in the results of alignment of relatively long, highly homologous sequences.

Experiments were performed on 1.6 GHz Pentium PC with 1 GB of memory, running Microsoft Windows XP. MegaBLAST was run as part of the latest version of the Bl2seq code, obtained from the NCBI website on November 20, 2005. SPOMF was run in MATLAB 7.0.1 (R14). For the data we used subsets of sequences *Arabidopsis thaliana* (chromosome 2 and 4) and *Homo sapiens* (chromosome 21 and 22), analyzed in [34], of lengths ranging from 65×10^3 to 16×10^6 bp. The full sizes of the *Arabidopsis thaliana* sequences were 19.6 and 17.5×10^6 bp; the full sizes of the longest contigs of *Homo sapiens* sequences were 28.6 and 23.3×10^6 bp, respectively. The results of the experiment are summarized in Figure 9. Several conclusions can be drawn.

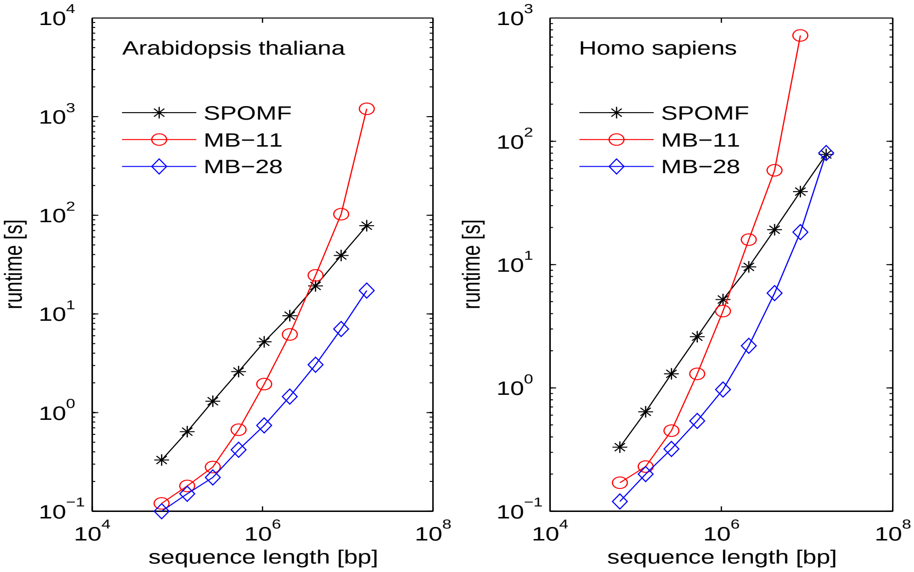

Figure 9: Sequence alignment runtimes (in seconds) of two versions of BLAST and SPOMF for subsets of *Arabidopsis thaliana* (chromosomes 2 and 4) and *Homo sapiens* (chromosomes 21 and 22).

First, the runtime and the runtime scalability with sequence size of both versions of MegaBLAST are not data-independent. For example, the runtime of MegaBLAST-28 for *Arabidopsis thaliana* increases almost 6-fold with a 4-fold increase in sequence length (from 4×10^6 to 16×10^6 bp). In contrast, the runtime of MegaBLAST-28 for *Homo sapiens* increases almost 14-fold for the same sequence lengths. For sequences of length 16×10^6 bp the runtimes of MegaBLAST-28 differ by a factor of 5. The dependence of BLAST performance on the data is well-known: in [16] an example of DNA sequences was given where the runtimes differed by more than a factor of 100. This is in contrast to SPOMF, whose performance depends solely on the data length. The reasons for the content dependence of BLAST performance are complex, and include, among others, the algorithm's sensitivity to sequence homology.

Second, the performance of MegaBLAST-11 scales very poorly with the sequence length. A 4-fold increase in sequence length (from 2×10^6 to 8×10^6 bp) leads to a 16-fold increase in runtime for the first sequence, and an almost 45-fold increase in runtime for the second sequence. In effect, while MegaBLAST-11 can be useful in the analysis of genes and smaller chromosomes, only MegaBLAST-28 and SPOMF are likely to perform well in large searches, such as genome-wide alignments. The results of the comparison of these two methods were mixed. For the *Homo sapiens* sequence SPOMF scaled better than MegaBLAST-28, and therefore it is likely to outperform MegaBLAST-28 for sequences longer than 16×10^6 bp. For the *Arabidopsis thaliana* sequence both SPOMF and MegaBLAST-28 scaled nearly linearly with the size of the data (although MegaBLAST-28 runtimes increase faster). As MegaBLAST-28 is roughly four times faster than SPOMF, it can be expected that the phase-only filter will not be competitive with BLAST for sequences of the *Arabidopsis thaliana* type, at least for lengths less than 10^8 bp.

The results of the experiment do not show a clear efficiency advantage of the SPOMF approach. This is in part due to the limited range of sequence lengths tested. As the rate of increase of computational complexity for both sequences grows faster for BLAST, it can be expected that

SPOMF will perform better for sequence lengths in the range of 10^8 to 10^9 bp. Moreover, the comparison made was not entirely fair. While BLAST is a mature product, developed and refined over fifteen years, SPOMF was implemented as a prototype MATLAB code, and incorporates a number of inefficiencies. For example, a future C-code version should improve computational efficiency of SPOMF by making a better use of computer memory. Reductions in the number of required operations due to complex encoding of DNA sequences, suggested in [14], could further bring at least a two-fold improvement in efficiency. Other possible encoding schemes that could be used alternatively or complementarily to the complex encoding include schemes based on gap coding [44], Huffman coding [40], and hypercomplex algebra [9]. Cumulatively, these refinements should make SPOMF competitive with BLAST over a wider range of sequence lengths. While this prediction will need to be validated using a more mature version of the current code and much more extensive testing on the data, it is congruent with observations made elsewhere. Perhaps more importantly, since FFT is one of the most common tasks in signal processing, PCI accelerator boards are commercially available, and can speed-up the computation of SPOMF significantly. In a 2003 paper [27], Kauer and Blöcker remark that the "Cheetah" card affords performing 5.2 million operations in about 1ms. Extrapolating this number to 384 million $\approx \mathrm{N} \log_2 \mathrm{N}$, where N=16 million is the length of one of the sequences used in our experiment, yields 0.074s per a 16000000-point FFT. This translates to about a 50-fold reduction in the execution time of the phase-only algorithm (in addition to the speed-up due to the coding improvements and algorithmic complexity reductions mentioned above). Finally, since the phase-only filter concept have been conceived in the field of optical signal processing, one has to consider a potential optical implementation of the SPOMF sequence aligner. Such an implementation was in fact proposed several years ago [10], and if technologically feasible, might deliver the most efficient solution to the sequence alignment problem.

7 Summary

We have shown that the SPOMF approach significantly outperforms the standard magnitude-and-phase MF method, when applied to periodic and semi-periodic DNA data, and performs similarly to MF with irregular data. Experiments on real DNA sequences indicate that the new approach is robust to isolated symbol insertions, symbol deletions, and symbol mismatches, and that local alignment of closely related DNA sequences (i.e., of 50%-100% homology) is feasible. In a limited experiment we have compared the current prototype version of the SPOMF code with the fastest version of BLAST. The results of the experiment appear to confirm the potential computational efficiency advantage of the FFT-based approach, especially when applied to sequences longer than 10^7 bp. A number of issues remain to be explored. Is the phase-only filter applicable to analysis of dissimilar sequences? Can the seed extension method [28] be adopted from BLAST to increase sensitivity of SPOMF to inexact matches? Can the approach be modified to allow a fully unconstrained gapped alignment? These issues will need to be answered in future research.

Acknowledgements

The author is grateful to Marnie Salisbury, Ann Jones and Olivia Peters for encouragement and support in pursuing this work, to John Dileo for discussions of repetitions in DNA sequences, to Marvin Drake for discussions of optical implementations of SPOMF, and to Ed Hanfling for software support.

References

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

[2] S. F. Altschul *et al.*, "Basic local alignment search tool", *J. Mol. Biol.*, Vol. 215, pp 403-410, 1990.

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

[4] A.K. Bansal, "An automated comparative analysis of 17 complete microbial genomes", *Bioinformatics*, Vol. 15, No. 11, pp 900-908, 1999.

[5] L. D. Baumert, "Cyclic difference sets", Springer, Berlin, 1971.

[6] G. Benson, "Tandem repeat finder: a program to analyse DNA sequences", *Nucleic Acids Research*, Vol. 27, No 2, pp 573-580, 1999.

[7] G. S. Bloom and S. W. Golomb, "Applications of numbered undirected graphs", *Proceedings of IEEE*, Vol. 65, No 4, pp 562-570, 1977.

[8] A.K. Brodzik, "A comparative study of cross-correlation methods for alignment of DNA sequences containing repetitive patterns", *Eusipco Proc.*, 2005.

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

[10] N. Brousseau *et al*, "Analysis of DNA sequences by an optical time-integrating correlator", *Applied Optics*, Vol. 31, No. 23, pp 4802-4815, 1992.

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

[12] M. Carlson and D. Brutlag, "Different regions of a complex satellite DNA vary in size and sequence of the repeating unit", *J. Mol. Biol.*, Vol. 135, pp 483-500, 1979.

[13] F. Chan and E. Rabe, "A non-linear phase-only algorithm for active sonar signal processing", *OCEANS Proc.*, Vol. 1, pp 506-511, 1997.

[14] 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.

[15] 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.

[16] A. Das *et al.*, "Performance of runtime optimization on BLAST", Technical Report, TR 04-038, University of Minnesota, 2004.

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

[18] J. Felsenstein, S. Sawyer and R. Kochin, "An efficient method for matching nucleic acid sequences", *Nucleic Acids Research*, Vol. 10, No. 1, pp. 133-139, 1982.

[19] K.A. Frazer *et al.*, "Cross-species sequence comparisons: a review of methods and available resources", *Genome Research*, No. 13, pp. 1-12, 2003.

[20] J. Gregor and M. G. Thomason, "Dynamic programming alignment of sequences representing cyclic patterns", *IEEE Trans. PAMI*, Vol. 15, No. 2, pp 129-135, 1993.

[21] D. Grover *et al.*, "Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition", *Bioinformatics*, Vol. 20, No. 6, pp. 813-817, 2004.

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

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

[24] J.L. Horner and P.D. Gianino, "Pattern recognition with binary phase-only filters", *Applied Optics*, 24, pp. 609-611, 1985.

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

[26] K. Katoh, K. Misawa, K. Kuma and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform", *Nucleic Acids Research*, Vol. 30, No 14, pp 3059-3066, 2002.

[27] G. Kauer and H. Blocker, "Applying signal theory to the analysis of biomolecules", *Bioinformatics*, Vol. 19, No. 16, pp 2016-2021, 2003.

[28] I. Korf, M. Yandell and J. Bedell, "BLAST", O'Reilly, Sebastopol, 2003.

[29] S. Kurtz and C. Schleiermacher, "REPuter - fast computation of maximal repeats in complete genomes", *Bioinformatics*, Vol. 15, pp 426-427, 1999.

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

[31] A. Lefebvre, T. Lecroq, H. Dauchel and J. Alexandre, "FORRepeats: detects repeats on entire chromosomes and between genomes", *Bioinformatics*, Vol. 19, No 3, pp 319-326, 2003.

[32] A. Lempel and J. Ziv, "On the complexity of finite sequences", *IEEE Trans. IT*, Vol. IT-22, No 1, pp 75-81, 1976.

[33] D. J. Lipman and W. R. Pearson, "Rapid and sensitive protein similarity search", *Science*, Vol. 227, pp 1435-1441, 1985.

[34] B. Ma, J. Tromp and M. Li, "PatternHunter: faster and more sensitive homology search", *Bioinformatics*, Vol. 18, No 3, pp 440-445, 2002.

[35] W. Miller, "Comparison of genomic DNA sequences: solved and unsolved problems", *Bioinformatics*, Vol. 17, No 5, pp 391-397, 2001.

[36] B. Morgenstern et al., "Exon discovery by genomic sequence alignment", *Bioinformatics*, Vol. 18, No. 6, pp 777-787, 2002.

[37] 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.

[38] C. Notredame, “Recent progress in multiple sequence alignment: a survey”, *Pharmacogenomics*, Vol. 3, No 1, pp 1-14, 2002.

[39] 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.

[40] S. Rajasekaran *et al*, “Efficient algorithms for local alignment search”, *J. Combinatorial Optimization*, No. 5, pp 117-124, 2001.

[41] D. Sidransky, “Nucleic acid-based methods for detection of cancer”, *Science*, Vol. 278, pp 1054-1058, 1997.

[42] D. Sharma *et al*, “Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation”, *Bioinformatics*, Vol. 20, No. 9, pp 1405-1412, 2004.

[43] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences”, *J. Mol. Biol.*, Vol. 147, pp 195-197, 1981.

[44] S.-C. Su, C. H. Yeh and C.-C. J. Kuo, “Structural analysis of genomic sequences with matched filtering”, *IEEE EMBS Proc.*, Vol. 3, pp 2893-2896, 2003.

[45] 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.

[46] J. W. Thomas *et al.*, “Comparative analyses of multi-species sequences from targeted genomic regions”, *Nature*, Vol. 424, pp 788-793, 2003.

[47] E. N. Trifonov, “3-, 10.5-, 200- and 400-base periodicities in genome sequences”, *Physica A*, Vol. 249, pp 511-516, 1998.