On the Structure of Small Motif Recognition Instances

Christina Boucher, Daniel G. Brown, and Stephane Durocher

D. R. Cheriton School of Computer Science,
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
{cabouche,browndg,sdurocher}@cs.uwaterloo.ca

Abstract. Given a set of sequences, S, and degeneracy parameter, d, the Consensus Sequence problem asks whether there exists a sequence that has Hamming distance at most d from each sequence in S. A *valid motif set* is a set of sequences for which such a consensus sequence exists, while a *decoy set* is a set of sequences that does not have a consensus sequence but whose pairwise Hamming distances are all at most $2d$. At present, no efficient solution is known to the Consensus Sequence problem when the number of sequences is greater than three. For instances of Consensus Sequence with binary sequences and cardinality four, we present a combinatorial characterization of decoy sets and a linear-time exact algorithm, resolving an open problem posed by Gramm *et al.* [7].

1 Introduction

Understanding the structure and function of genomic data remains an important biological and computational challenge. *Motifs* are short sequences of genomic DNA responsible for controlling biological processes, such as gene expression. Motifs with the same function may not entirely match, due to random mutations or chemical properties. The *motif consensus* of the instances is a short sequence representing their shared pattern. Given a number of DNA sequences, *motif recognition* is the task of discovering motif instances in sequences without prior knowledge of the consensus or their placement within the sequence.

Closely related to the motif recognition problem is the Consensus Sequence problem that asks, given a parameter d and a set of sequences $S = \{s_1, \ldots, s_n\}$ each of length l, whether there exists a sequence s^*, which we call a consensus, that is of distance at most d from each sequence in S. Note that the consensus sequence need not be contained in S. In this context, the distance metric is the Hamming distance, $H(s_i, s_j)$, between two sequences s_i and s_j. Consensus Sequence is NP-complete, even for the case where each sequence is binary; therefore, no polynomial-time solution is possible unless $P = NP$ [5]. Clearly, a set for which the distance between any pair of sequences exceeds $2d$ cannot have a consensus. We say a set of sequences S is *pairwise bounded* if for all sequences $a, b \in S$, $H(a, b) \leq 2d$. Thus, the Consensus Sequence problem essentially reduces to discerning between pairwise bounded sets that have a consensus, and if so, finding one such sequence s^*, and those that do not. A set of sequences S

A. Amir, A. Turpin, and A. Moffat (Eds.): SPIRE 2008, LNCS 5280, pp. 269–281, 2008.

is a *motif set* if there exists a consensus sequence, s^*. We say set S is a *decoy* set if S is pairwise bounded but does not have a consensus.

These problems – motif recognition and Consensus Sequence – have an extensive number of applications, due to the fact that many problems aim to determine if a set of sequences has a specific measure of similarity. For example, the Consensus Sequence problem arises in areas such as coding theory [3,5], data compression [6], and bioinformatics [7,8,10]. In the context of coding theory, a well-known problem related to Consensus Sequence asks if there exists a code that is not too far away from a given set of codes [3,5]. Given its applicability, the Consensus Sequence problem needs to be solved efficiently in practice. Li *et al.* [12] present a polynomial-time approximation scheme (PTAS) for Consensus Sequence. For a given value of r, all choices of r subsequences of length l are considered from the n sequences. The algorithm has $O(l(nm)^{r+1})$ run time, which is polynomial for any constant r. Many researchers have studied the algorithm due to Li *et al.* [12]; for a variant of Consensus Sequence there are known "weak" instances for which the approximation ratio is $1 + \Theta(1/\sqrt{r})$ [1], and "strong" instances for which the PTAS will be guaranteed to determine the correct answer in efficient time [2].

Another approach is to investigate the parameterized complexity of Consensus Sequence. A problem φ is said to be *fixed-parameter tractable* (FPT) with respect to parameter k if there exists an algorithm that solves φ in $f(k) \cdot n^{O(1)}$ time, where f is a function of k that is independent of n [8]. Gramm *et al.* [7] demonstrate that Consensus Sequence is FPT when the number of sequences remains fixed: the problem is polynomial-time solvable with a fixed number of sequences. This FPT result is based on an Integer Linear Programming (ILP) formulation with a constant number of variables (assuming n is fixed), and the application of the result of Lenstra [11], which states that ILP is polynomial-time solvable when the number of variables remains fixed. Unfortunately, such an ILP formulation is only of theoretical interest since the corresponding algorithms lead to very long running times even when the number of sequences is small (e.g., four sequences over a binary alphabet). Other parameterizations of the Consensus Sequence also exist; for example, when d is fixed, the problem can be solved in $O(nl + nd(d+1)^d)$ time [8].

Gramm *et al.* [7] and Sze *et al.* [14] give direct (non-ILP based) combinatorial algorithms for solving Consensus Sequence exactly for three sequences. The algorithm of the former authors considers the possible combinations of alphabet symbols that can occur for three sequences, then specifies conditions for which a consensus sequence can be constructed [7]. Sze *et al.* [14] give a counting argument to demonstrate a condition for which a set of three sequences has a consensus and when it does not. In fact, a stronger property applies to binary sequences: any three pairwise-bounded binary sequences have a consensus.

Gramm *et al.* state that the problem of finding an efficient polynomial-time algorithm for solving Consensus Sequence on a set of four sequences remains open "due to the enormous combinatorial complexity [of the ILP-based solution]" [8, p. 13]. We resolve this open problem for binary sequences; specifically, we give an exact combinatorial algorithm for four binary sequences. This result is inspired by the combinatorial decomposition theorem for decoy sets that is also presented, which demonstrates that each decoy set can be characterized by containing two specific subsequences. Our aim is that these results might be extended to resolve the more general Consensus Sequence problem, in particular, for the four-symbol DNA alphabet, or for more than four sequences.

2 Preliminaries

We begin with some definitions concerning general sequence analysis. Let l, $d \leq l$, and n be positive integers and σ_i be a function that returns the ith symbol in a sequence. For any symbol $\beta \in \Gamma$ let β^l denote the l-length sequence of all β's. Given a set of sequences $S = \{s_1, \ldots, s_n\}$, each of which has length l, the ith *column* refers to the column vector $c_i = [\sigma_i(s_1), \ldots, \sigma_i(s_n)]^T$ in the $n \times l$ matrix representation of S. A sequence s^* is an *optimal sequence* for S if and only if there is no sequence s^*_2 with $\max_{i=1,\ldots,n} H(s^*_2, s_i) < \max_{i=1,\ldots,n} H(s^*, s_i)$. Note that the optimal sequence for S is not unique; there may exist multiple. We formally define the Consensus Sequence problem as follows:

Consensus Sequence
Instance: a set of n sequences, $S = \{s_1, s_2, \ldots, s_n\}$ over an alphabet Γ, each of length l, and a positive integer d.
Find: a l-length sequence s^* over alphabet Γ where $H(s^*, s_i) \leq d$ for every s_i in S, or declare that no such s^* exists.

The difficulty of Consensus Sequence lies in distinguishing between decoy and valid motifs. In the context of coding theory, Frances and Litman show that Consensus Sequence remains NP-hard even when restricted to a binary alphabet; in this case, they refer to the corresponding problem as Radius Decision [5]. We will be interested in the cardinality of a decoy set, that is, the number of sequences contained in the set. We say set $\hat{S} \subseteq S$ is a decoy of *minimal cardinality* if $\hat{S}$ is a decoy set such that for all $S' \subseteq S$, if $|S'| < |\hat{S}|$, then S' has a consensus.

Gramm *et al.* [7] refer to the process of permuting the columns of S such that these are grouped by column type as "normalization". A normalized instance can be derived from the input set of sequences by a simple linear-time algorithm. Given an optimal sequence for the normalized set of sequences, the inverse of this same permutation returns an optimal sequence for the original input [7].

Definitions Specific to Sets of Cardinality Four: Given a set $S = \{s_1, \ldots, s_4\}$ of binary sequences, the symbols in each column have either two, three, or four matching symbols. Sixteen types of columns are possible in general. We say a column belongs to *group i* if it has exactly i matching symbols. To reduce the number of possible types to eight, suppose without loss of generality that $s_4 = \beta^l$. Equivalently, create a new set S' by performing a logical exclusive-or of each

Table 1. The values $\lambda_{\alpha\beta\beta}$ through $\lambda_{\alpha\alpha\beta}$ denote the number of columns of each type in groups three and two. The symbol "-" implies that the value is undefined at these columns.

Group	Four	Three				Two		
# of columns.	$\lambda_{\beta\beta\beta}$	$\lambda_{\alpha\beta\beta}$	$\lambda_{\beta\alpha\beta}$	$\lambda_{\beta\beta\alpha}$	$\lambda_{\alpha\alpha\alpha}$	$\lambda_{\beta\alpha\alpha}$	$\lambda_{\alpha\beta\alpha}$	$\lambda_{\alpha\alpha\beta}$
s_1	β	α	β	β	α	β	α	α
s_2	β	β	α	β	α	α	β	α
s_3	β	β	β	α	α	α	α	β
s_4	β	β	β	β	β	β	β	β
maj_i	β	β	β	β	α	-	-	-

sequence in S with s_4 (say α corresponds to boolean true). A consensus sequence for S is found by performing another exclusive-or on a consensus sequence for S'. Let λ_{abc} denote the number of instances of column $(a, b, c, \beta)^T$, where $a, b, c \in \{\alpha, \beta\}$. See Table 1. Note that only columns of group three and two need to be considered, since any optimal sequence will have the majority vote at each column of group four. A pair of columns are considered to be *identical* if a pair of sequences in one column mismatch if and only if the same sequences mismatch in the second column. For example the column $[\alpha\alpha\beta\beta]^T$ is identical to $[\beta\beta\alpha\alpha]^T$, but neither is identical to $[\alpha\beta\beta\alpha]^T$.

Let maj_i denote the majority of the four symbols in column i. That is, $\text{maj}_i = \alpha$ if symbol α occurs three or more times in column i and $\text{maj}_i = \beta$ if symbol β occurs three or more times; maj_i is undefined if α and β each occur twice. Assuming that $s_4 = \beta^l$, only the columns associated with $\lambda_{\alpha\alpha\alpha}$ are such that $\text{maj}_i = \alpha$.

3 Ubiquitousness or Rareness of Bounded Decoy Sets

In this section, we consider the relative frequency, or infrequency, of decoy sets that do not have a proper subset that is also a decoy. Our empirical results demonstrate that the relative frequency of such decoy sets is minimal, and that the majority of decoy sets contain a decoy subset of cardinality four. Still, the results of Gramm *et al.* [8] imply that we cannot characterize all decoy sets of arbitrary size n as having a proper subset that is a decoy. We refer to a set Q of decoys, each of cardinality n, as having decoys of *bounded cardinality* if every decoy in Q has a proper subset that is a decoy.

Proposition 1. *Let Γ denote an alphabet of arbitrary fixed size. If $P \neq NP$, then for any n_0 there exist a decoy set S such that every subset of S of cardinality n_0 has a consensus.*

Proof. Suppose otherwise. That is, there exists an n_0 such that every decoy S of size $n \geq n_0$ has a subset of size n_0 that is a decoy. By Gramm *et al.* [8], for any fixed n_0, there exists an algorithm that decides whether a set of n_0 sequences is a decoy in $f(l, d)$ time, where $f(l, d)$ is polynomial in l and d. Consequently, for any set of $n \geq n_0$ sequences S, we can check each of the $\binom{n}{n_0}$ subsets of S of size n_0 to determine whether any is a decoy in time $O(n^{n_0} f(l, d))$. That is, we can determine in polynomial time whether S is a decoy. Since Consensus Sequence is NP-complete, this is possible only if P = NP. □

It should be noted that this corollary does not preclude the fact that there may exist values of n where all decoys of cardinality n have the cardinality of the minimal decoy also as n. What the result implies is that there does not exist a threshold γ such that for all values of n greater than γ the minimal decoy sets have bounded cardinality below γ, if $P \neq NP$. Although Proposition 1 implies that no fixed n_0 exists, we conjecture that most decoys have a subset of size four that is a decoy. We provide evidence toward this property with an empirical study on random sets of binary sequences which we now describe. In turn, these results motivate the need for an efficient algorithm for determining whether a set of four sequences is a consensus; we describe such an algorithm in Section 4.2.

We empirically investigate the rarity of the occurrence of decoy sets of cardinality n for which the cardinality of a minimal decoy set is large relative to n. We sampled without replacement 1000 times from the set of all possible pairwise-bounded sets of binary, l-length sequences; each set sampled has exactly n sequences taken from the binary alphabet. We varied the values for n, l and d. For each sample set, we determined whether the set is a decoy or a valid motif, with respect to the value of d, and determined the cardinality of the minimum decoy set. We repeated this experiment 10 times and calculated the mean values obtained. Table 2 outlines this data. One significant empirical trend demonstrates that as the number of sequences increased, the number of decoys that do not contain a minimal decoy of cardinality four became exponentially smaller; when

Table 2. Data obtained from calculating the average of 10 experiments that obtain a random sample, without replacement, of 1000 sequence sets and determine the size of the minimal decoy contained in each set decoy obtained in the sample. The first column is the cardinality of the minimum decoy set.

No. of	$l = 8$, $d = 3$			$l = 10$, $d = 3$			$l = 15$, $d = 4$		
sequences	$n = 6$	$n = 10$	$n = 12$	$n = 6$	$n = 10$	$n = 12$	$n = 6$	$n = 10$	$n = 12$
No. of valid motif	443.4	4.6	88.7	394.4	9.3	3.6	101.4	3.5	2.6
4	542.2	995.4	991.3	605.6	990.7	996.4	898.6	996.5	997.4
5	12.4	0	0	7	0	0	4.1	0	0
6	2	0	0	0.2	0	0	0.8	0	0
7	-	0	0	-	0	0	-	0	0
8	-	0	0	-	0	0	-	0	0
9	-	0	0	-	0	0	-	0	0
10	-	0	0	-	0	0	-	0	0
11	-	-	0	-	-	0	-	-	0
12	-	-	0	-	-	0	-	-	0

n was 10 and 12 the number of minimal decoys of size larger than four was 0. The only value of n for which decoys of size n were seen was 6. Further, the total number of decoys in 900,000 set of sequences sampled were approximately 1,500 in total. In summary, the empirical results appear to indicate that a large percentage of binary decoys can be characterized by containing a minimal decoy set of size four, the smallest size possible, further motivating the main results in this paper.

4 Investigating Binary Decoy Sets of Cardinality Four

Gramm *et al.* [7] suggest that a direct combinatorial approach to solve Consensus Sequence where n is fixed would be of practical and theoretical interest. Here we focus on partially resolving this open problem. We restrict interest to binary decoy sets of cardinality four, give a decomposition theorem and a linear-time, exact algorithm for these instances. We first prove that all binary sets of cardinality four can be decomposed into subsequences that have a specific characterization. The linear-time exact algorithm considers all possible combinations of symbols from the binary alphabet, and sequentially constructs a consensus or returns that no consensus exists.

4.1 A Decomposition Theorem

We will prove that each decoy of cardinality four can be decomposed into two subsequences that have a specific characterization. We begin by presenting the terminology and notation used to define these two subsequences. We define an $\alpha\beta$-set for an alphabet $\{\alpha, \beta\}$ as the set of all possible sequences of length two, that is, the set $\{\alpha\alpha, \alpha\beta, \beta\alpha, \beta\beta\}$. Given a set $S = \{s_1, s_2, s_3, s_4\}$ of cardinality four, we refer to S as containing an $\alpha\beta$-set if there are distinct indices i and j where the set of subsequences defined by columns i and j of S is an $\alpha\beta$-set. For example, the set of four sequences $\{\alpha\alpha\beta, \alpha\beta\beta, \beta\alpha\alpha, \beta\beta\beta\}$ contains an $\alpha\beta$-set at the first two columns. Next, we refer to a sequence s as *adequately far* if there exists a sequence, say $s_1 \in S$, such that $H(s, s_1) = d - 1$, and for all $s_i \in S$ the distance $H(s_i, s)$ is equal to either $d - 1$ or d. We refer to a sequence s^2 as *too close* if there exists some $s_i \in S$ with $H(s^2, s_i) \leq d - 2$ and $H(s^2, s_i) \leq d - 1$, for all $s_i \in S$. Putting these definitions together, we obtain the following property:

Definition 1. (Characterization of decoys of cardinality four) *A set of binary sequences S has property D if the following conditions hold:*

1. *S has an $\alpha\beta$-set realised at indices i and j, and*
2. *each optimal sequence for S', the set of sequences obtained from S by removing the columns i and j, is adequately far.*

We require Lemma 1 to prove our combinatorial decomposition theorem. The proof is omitted due to space constraints. We illustrate an example of property D in Figure 1: a decoy set where all sequences that have distance to the closest sequence equal to $d - 1$ and an $\alpha\beta$-set at the last two columns.

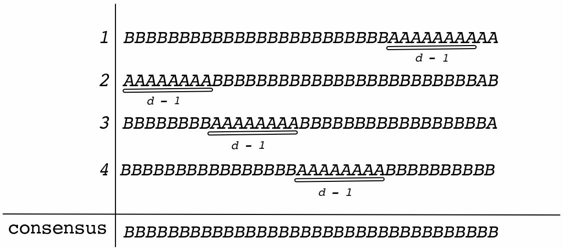
1 BBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAA
d - 1
2 AAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBAB
d - 1
3 BBBBBBBBAAAAAAAABBBBBBBBBBBBBBBBBA
d - 1
4 BBBBBBBBBBBBBBBBAAAAAAAABBBBBBBBBB
d - 1
consensus BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

Fig. 1. Illustrates a decoy set such that optimal sequence to the set of sequences obtained from S by removing the last two columns that have the $\alpha\beta$-set, is adequately far, and more specifically each has distance equal to $d-1$ to the optimal sequence

Lemma 1. *Assume $d \geq 2$, $l \geq 2$ and $\Gamma = \{\alpha, \beta\}$. Every decoy set S of cardinality four contains an $\alpha\beta$-set.*

Theorem 1. (Decomposition theorem for cardinality four) *Assume $d \geq 2$, $l \geq 2$ and $\Gamma = \{\alpha, \beta\}$. A set S of cardinality four is a decoy if and only if there exists a set of subsequences contained in S such that property D holds.*

Lastly, we demonstrate the following result about the existence of an unique consensus for sets of arbitrary cardinality. It seems natural that as the cardinality of a valid motif set S increases relative to d, the number of consensuses for S decreases. As we observe in Proposition 2, a valid motif set of maximal cardinality has an unique consensus when $d < l$.

Proposition 2. *If a set S has a consensus but no superset of S has a consensus, then either:*

1. *$d < l$ and S has a unique consensus, or*
2. *$d \geq l$ and S is the set of all possible binary sequences of length l, each of which is a consensus for S.*

Proof. Case 1. Assume $d < l$. Assume S has two distinct consensuses, denoted by a and b. Since no superset of S has a consensus, all sequences c such that $H(a,c) \leq d$ must be in S. The same holds for all sequences e such that $H(b,e) \leq d$. Furthermore, sequences a and b must be in S. Consequently, $H(a,b) \leq d$. Let $\delta = H(a,b)$. Without loss of generality, assume that a and b differ in the first δ bits. Let f denote a binary sequence that agrees with a in the first δ bits, differs from a in the next d bits, and agrees with a in any remaining bits. Observe that $H(f,b) = H(f,a) + H(a,b) = d + \delta > d$. Consequently, f is not in S. Since $H(a,f) \leq d$, sequence a is a consensus of $S \cup \{f\}$. This derives a contradiction; our assumption must be false and the consensus of S must be unique.

Case 2. Assume $d \geq l$. Any two binary sequences of length l differ in at most l bits. Since no superset of S has a consensus, $S = \Gamma^l$. □

In some cases, a set of sequences S that has a consensus does not have a decoy as a superset whereas in other cases, every pairwise bounded superset of S is a decoy. For example, let $d = 1$, let $S = \{\beta\beta\beta\beta, \beta\beta\alpha\alpha, \beta\alpha\beta\alpha, \alpha\beta\beta\alpha\}$, and let $S' = \{\beta\beta\beta\beta, \beta\beta\beta\alpha, \beta\beta\alpha\alpha\}$.

4.2 Finding a Consensus for a Set of Four Sequences

As motivated in Section 1, the only previous polynomial-time solution for finding a consensus of a set of four sequences [8] was intended more to demonstrate the fixed-parameterized tractability of the problem rather than to provide an efficient solution. As acknowledged by its authors, the corresponding description (for which many details are omitted) results in an algorithm with extremely high (although theoretically linear) run time and, furthermore, does not lend itself well to simple or practical implementation. In this section, we present a simple linear-time algorithm for finding a consensus of a set of four binary sequences or determining that the set is a decoy. After describing the algorithm, we prove its correctness and show its worst-case run time is $O(l)$ for any arbitrary d.

Given a set of binary sequences $S = \{s_1, \ldots, s_4\}$, algorithm BinaryConsensus4 identifies a consensus sequence s^* for S if one exists. Again, to simplify the algorithm's, description suppose $s_4 = \beta^l$. The algorithm greedily assigns symbols to s^*, one symbol at a time. Each column c_i is initially considered to be *free*; that is, no symbol has been assigned to $\sigma_i(s^*)$. Once it is assigned a symbol, we say column c_i is *fixed* and its value is not modified again. The algorithm has three phases in which columns of groups four, three, and two are fixed, respectively.

Phase One. Fix symbols of s^* in all columns of group four such that these agree with the symbol of the corresponding column.

Phase Two. The symbols of s^* in columns of group three are fixed sequentially. Say the first $i - 1$ columns of group three have been fixed and consider consider the ith such column. Let s_j denote the sequence of S that disagrees with the remaining three sequences in this column. Let s^+ denote the sequence given by the symbols of s^* in the fixed columns and the symbols of s_j in the free columns. If s^+ is a consensus for S, then let $s^* = s^+$ and return s^*. Otherwise, fix the current column of s^* to agree with the majority and continue to the next column of group three.

Phase Three. If phase three is reached, then only columns of group two remain, of which at most three types may be present. The free columns are fixed by selecting the number of columns of each type that will be assigned symbol α versus β. That is, a solution for columns of group two corresponds to a triple of integers (x, y, z), where $x \in [0, \lambda_{\beta\alpha\alpha}]$ denotes the number of columns of type $\lambda_{\beta\alpha\alpha}$ that will be assigned the symbol α and $\lambda_{\beta\alpha\alpha} - x$ represents the number that will be assigned the symbol β. The variables y and z are defined analogously. See Table 3. We denote the corresponding sequence by $s^*_{x,y,z}$. Therefore, the problem reduces to identifying an integer triple (x, y, z) selected from the region $R = [0, \lambda_{\beta\alpha\alpha}] \times [0, \lambda_{\alpha\beta\alpha}] \times [0, \lambda_{\alpha\alpha\beta}]$ that minimizes

$$f(x, y, z) = \max_{s_i \in S} H(s_i, s^*_{x,y,z}), \tag{1}$$

where

$$H(s_1, s^*) = \lambda_{\alpha\beta\beta} + x + \lambda_{\alpha\beta\alpha} - y + \lambda_{\alpha\alpha\beta} - z, \tag{2a}$$

$$H(s_2, s^*) = \lambda_{\beta\alpha\beta} + \lambda_{\beta\alpha\alpha} - x + y + \lambda_{\alpha\alpha\beta} - z, \tag{2b}$$

$$H(s_3, s^*) = \lambda_{\beta\beta\alpha} + \lambda_{\beta\alpha\alpha} - x + \lambda_{\alpha\beta\alpha} - y + z, \tag{2c}$$

$$H(s_4, s^*) = \lambda_{\alpha\alpha\alpha} + x + y + z. \tag{2d}$$

The sequence $s^*_{x,y,z}$ does not actually need to be constructed since the corresponding value of (1) is obtained in constant time upon fixing values for x, y, and z.

Table 3. The consensus s^* found by algorithm BinaryConsensus4 (if one exists) is displayed in the last row. The values $\lambda_{\beta\beta\beta}$ through $\lambda_{\alpha\alpha\beta}$ denote the number of columns of each type and functions x, y, and z denote the number of occurrences of symbol α in the corresponding column as derived by the algorithm.

Column Group		Four	Three				Two		
Algorithm Phase		1	2				3		
Number of Columns		$\lambda_{\beta\beta\beta}$	$\lambda_{\alpha\beta\beta}$	$\lambda_{\beta\alpha\beta}$	$\lambda_{\beta\beta\alpha}$	$\lambda_{\alpha\alpha\alpha}$	$\lambda_{\beta\alpha\alpha}$	$\lambda_{\alpha\beta\alpha}$	$\lambda_{\alpha\alpha\beta}$
Set S	s_1	β	α	β	β	α	β	α	α
	s_2	β	β	α	β	α	α	β	α
	s_3	β	β	β	α	α	α	α	β
	s_4	β	β	β	β	β	β	β	β
Consensus	s^*	β	β	β	β	α	x	y	z

Instead of evaluating all integer combinations for (x, y, z) (requiring $O(l^3)$ time), we identify a set $T \subseteq \mathbb{Q}^3 \cap R$ containing a constant number of triples such that the optimal (possibly non-integer) solution to (1) is a triple in T. Interpreted geometrically, (2a) through (2d) correspond to four respective hyperplanes in $\mathbb{R}^4$ whose maximum, $f(x, y, z)$,, defines a surface. Let

$$\begin{aligned} x_0 &= \frac{1}{4}\left(-\lambda_{\alpha\beta\beta} + \lambda_{\beta\alpha\beta} + \lambda_{\beta\beta\alpha} - \lambda_{\alpha\alpha\alpha} + 2\lambda_{\beta\alpha\alpha}\right), \\ y_0 &= \frac{1}{4}\left(\lambda_{\alpha\beta\beta} - \lambda_{\beta\alpha\beta} + \lambda_{\beta\beta\alpha} - \lambda_{\alpha\alpha\alpha} + 2\lambda_{\alpha\beta\alpha}\right), \\ z_0 &= \frac{1}{4}\left(\lambda_{\alpha\beta\beta} + \lambda_{\beta\alpha\beta} - \lambda_{\beta\beta\alpha} - \lambda_{\alpha\alpha\alpha} + 2\lambda_{\alpha\alpha\beta}\right). \end{aligned} \tag{3}$$

If $(x_0, y_0, z_0) \in R$, then let $T = \{(x_0, y_0, z_0)\}$. Otherwise, let T denote the set of triples that correspond to x-, y-, and z-coordinates of vertices of the intersection of the surface defined by (1) with the boundary of R. If this intersection is empty, then it follows that no consensus exists.

For each triple $(x, y, z) \in T$, evaluate the integer triples within unit ℓ_∞ distance of (x, y, z) in region R. That is, for every $(x, y, z) \in T$, consider the integer triples in $[\max(0, x-1), \min(x+1, \lambda_{\beta\alpha\alpha})] \times [\max(0, y-1), \min(y+1, \lambda_{\alpha\beta\alpha})] \times [\max(0, z-1), \min(z+1, \lambda_{\alpha\alpha\beta})]$, of which there are at most eight. Compute

(1) for each such integer triple (x, y, z) and store the corresponding minimizing sequence $s^*_{x,y,z}$. Let $s^* = s^*_{x,y,z}$.

Termination. Consider the maximum distance between s^* and a sequence in S, i.e., the minimum (integer) value of (1). If this value is at most d, then s^* is returned as a consensus sequence for S. Otherwise, S is a decoy set and no consensus sequence exists.

We now demonstrate that algorithm BinaryConsensus4 correctly returns a consensus s^* for every set S that is a valid motif set. Furthermore, this is achieved in $O(l)$ time, independently of d. The proof of Theorem 2 refers to Lemmas 2 and 3 which follow.

Theorem 2. *Given any $d \in \mathbb{Z}^+$, any $l \in \mathbb{Z}^+$, and any set S of four binary sequences of length l, algorithm* BinaryConsensus4 *returns a consensus for S with degeneracy parameter d if one exists or returns that S is a decoy in $O(l)$ time.*

Proof. The correctness of Phase 1 is straightforward. The correctness of Phase 2 follows by induction on i using Lemma 2. Consequently, if S has a consensus, then either a consensus has been found by the end of Phase 2 (i.e., $s^* = s^+$), or there exists a consensus s^* such that $\sigma_x(s^*) = \mathrm{maj}_x$ for all columns of groups three and four. The optimal solution for the remaining free columns is found in Phase 3. The correctness of Phase 3 follows by Lemma 3. Therefore, algorithm BinaryConsensus4 returns a consensus s^* if one exists, and returns that no consensus exists otherwise.

Each phase requires a single pass through the columns of S. Phase 1 simply requires counting the number of columns of each type. Phase 2 also requires maintaining the twelve distances $H(s_i, s_j^+)$ for each $\{i, j\} \subseteq \{1, \ldots, 4\}$, where s_j^+ denotes s^+ for which the free columns are defined according to s_j (as described in Phase 2 of algorithm BinaryConsensus4). Every time a column of group three is fixed, each of these twelve values can be updated in constant time. Phase 3 simply requires counting the number of columns of each type. Since (1) is defined by the maximum of four hyperplanes and region R is bounded by three pairs of parallel planes, the number of triples in T is constant and, furthermore, the coordinates of these triples are straightforward to compute in constant time. Finally, since any point in $\mathbb{R}^3$ has at most eight integer points within unit ℓ_∞ distance from it, the set of integer triples evaluated is also computed in constant time and space. Therefore, algorithm BinaryConsensus4 terminates in $O(l)$ time. □

Definition 2 (Majority Rule Property). *We say property $P(i)$ holds for a set S of four binary sequences if and only if either*

1. *S is a decoy, or*
2. *there exists a consensus of S for which the first i columns of group three have value* maj_i.

Lemma 2. *Let $S = \{s_1, \ldots, s_4\}$ denote a set of four binary sequences that has m columns of group three. If $P(i)$ holds for some $i \in \{0, \ldots, m-1\}$ and $s_j \in S$ denotes the sequence that mismatches in the $(i+1)$st column of group three, then either*

1. *$P(i+1)$ holds for S, or*
2. *s^+ is a consensus for S,*

where $\sigma_x(s^+) = \mathrm{maj}_x$ in the first i columns of group three and $\sigma_x(s^+) = \sigma_x(s_j)$ in the remaining columns.

Proof. If s^+ is a consensus for S then the claim holds. Similarly, if S is a decoy then $P(i+1)$ is true and the claim holds. Therefore, suppose S is not a decoy and s^+ is not a consensus for S. By $P(i)$, S has a consensus s^* for which the first i columns of group three have value maj_x. Let k denote the index of the $(i+1)$st column of group three.

Case 1. Suppose $\sigma_k(s^*) = \mathrm{maj}_k$. Therefore, $P(i+1)$ is true, and the claim holds.

Case 2. Suppose $\sigma_k(s^*) \neq \mathrm{maj}_k$. That is, $\sigma_k(s^*) = \sigma_k(s_j)$. Since s^+ is not a consensus, s^* and s^+ must differ in at least one column; let k' denote the index of such a column. Let s^{**} denote a sequence of length l such that $\sigma_x(s^{**}) \neq \sigma_x(s^*)$ for $x \in \{k, k'\}$ and $\sigma_x(s^{**}) = \sigma_x(s^*)$ otherwise. That is, $\sigma_k(s^{**}) = \mathrm{maj}_k$. Thus, $H(s_j, s^{**}) = H(s_j, s^*)$ and, furthermore,

$$\forall x \in \{1, \ldots, 4\},\ H(s_x, s^{**}) \leq H(s_x, s^*).$$

Therefore, s^{**} is a consensus for S, $P(i+1)$ is true, and the claim holds. □

Lemma 3. *There exists an integer triple $(x, y, z) \in [0, \lambda_{\beta\alpha\alpha}] \times [0, \lambda_{\alpha\beta\alpha}] \times [0, \lambda_{\alpha\alpha\beta}]$ that minimizes (1) and is within unit ℓ_∞ distance from a triple in T, where set T contains either (3) or the set of triples that correspond to vertices of the intersection of the surface defined by (1) with the boundary of R.*

Proof. Since no two of the hyperplanes induced by (2a) through (2d) are parallel, $f(x, y, z)$ is a convex function whose surface includes a unique simplicial vertex located at the point of intersection of these four hyperplanes. Furthermore, this point minimizes $f(x, y, z)$ since f is increasing as it tends to infinity in any direction. Thus, $f(x, y, z)$ is minimized at a unique (possibly non-integer) point found by solving for x, y, and z in

$$H(s_1, s^*) = H(s_2, s^*) = H(s_3, s^*) = H(s_4, s^*). \tag{4}$$

The constraints of (4) corresponds to system of three linear equations with the unique solution (3).

Since the coefficients of x in (2a) through (2d) are all ±1, function $f(x, y, z)$ has slope ±1 along the x-axis for any fixed y and z. The same holds for any fixed x and y or any fixed x and z. Consequently, since $f(x, y, z)$ is convex, a minimum integer solution to (1) lies within unit ℓ_∞ distance of its non-integer solution. The set T contains either the unique minimum (3) (if it lies within region R) or the set of triples that correspond to vertices of the intersection of the surface defined by $f(x, y, z)$ with the boundary of R, one of which must minimize $f(x, y, z)$ over R. Therefore, the claim holds. □

5 Conclusion

Motif recognition, in which the objective is to identify meaningful patterns in biological data, is a fundamental problem of computational biology. We have obtained a combinatorial characterization of the consensus problem for instances of four binary sequences, and a linear-time algorithm for obtaining a consensus for this restricted set of instances. Our results generalize previous work and answer some open problems concerning Consensus Sequence [7]. We aim to generalize our current results to identify a combinatorial characterization of decoy sets over larger alphabets. Such a generalization would invite many open problems in motif recognition to be revisited, as their tractability might be determined more concretely, opening the possibility for more efficient algorithmic solutions.

Acknowledgements

We gratefully acknowledge research support of the National Sciences and Engineering Research Council of Canada.

References

1. Brejová, B., Brown, D., Harrower, I., López–Ortiz, A., Vinař, T.: Sharper upper and lower bounds for an approximation scheme for Consensus–Pattern. In: Apostolico, A., Crochemore, M., Park, K. (eds.) CPM 2005, vol. 3537, pp. 1–10. Springer, Heidelberg (2005)
2. Brejová, B., Brown, D., Harrower, I., Vinař, T.: New bounds for motif-finding in strong instances. In: Lewenstein, M., Valiente, G. (eds.) CPM 2006. LNCS, vol. 4009, pp. 94–105. Springer, Heidelberg (2006)
3. Cohen, G., Honkala, I., Litsyn, S., Sole, P.: Long packing and covering codes. IEEE Trans. Inf. Theory 43(5), 1617–1619 (1997)
4. Fellows, M., Gramm, J., Niedermeier, R.: On the parameterized intractability of motif search problems. Combinatorica 26(2), 141–167 (2006)
5. Frances, M., Litman, A.: On covering problems of codes. Th. Comp. Sys. 30, 113–119 (1997)
6. Graham, R.L., Sloane, N.J.A.: On the covering radius of codes. Trans. Inf. Theory 31, 385–401 (1985)
7. Gramm, J., Niedermeier, R., Rossmanith, P.: Exact solutions for Closest String and related problems. In: Eades, P., Takaoka, T. (eds.) ISAAC 2001, vol. 2223, pp. 441–453. Springer, Heidelberg (2001)
8. Gramm, J., Niedermeier, R., Rossmanith, P.: Fixed-parameter algorithms for closest string and related problems. Algorithmica 37(1), 25–42 (2003)

9. Gramm, J., Guo, J., Niedermeier, R.: Parameterized intractability of distinguishing substring selection. Th. Comp. Sys. 39(4), 545–560 (2006)
10. Lanctot, J.K., Li, M., Ma, B., Wang, S., Zhang, L.: Distinguishing string selection problems. In: Proc. SODA 1999, pp. 633–642 (1999)
11. Lenstra, W.H.: Integer programming with a fixed number of variables. Math. of OR 8, 538–548 (1983)
12. Li, M., Ma, B., Wang, L.: Finding similar regions in many strings. J. Comp. and Sys. Sci. 65(1), 73–96 (2002)
13. Pevzner, P., Sze, S.: Combinatorial approaches to finding subtle signals in DNA sequences. In: Proc. ISMB 2000, pp. 344–354 (2000)
14. Sze, S., Lu, S., Chen, J.: Integrating sample-driven and patter-driven approaches in motif finding. In: Jonassen, I., Kim, J. (eds.) WABI 2004. LNCS (LNBI), vol. 3240, pp. 438–449. Springer, Heidelberg (2004)