Extracting Best Consensus Motifs from Positive and Negative Examples

Tateishi, Erika
Department of Information Systems, Kyushu University

Maruyama, Osamu
Department of Information Systems, Kyushu University

Miyano, Satoru
Research Institute of Fundamental Information Science Kyushu University

https://hdl.handle.net/2324/3217

出版情報：RIFIS Technical Report. 115, 1995-06-08. 九州大学理学部附属基礎情報学研究施設
バージョン：
権利関係：

RIFIS-TR-CS-115

RIFIS Technical Report

Extracting Best Consensus Motifs from Positive and Negative Examples

Erika Tateishi
Osamu Maruyama
Satoru Miyano

June 8, 1995
Revised: August 21, 1995

Research Institute of Fundamental Information Science
Kyushu University 33
Fukuoka 812, Japan
E-mail: tateishi@rifis.kyushu-u.ac.jp
Phone: 092-641-1101 ex. 4459

Extracting Best Consensus Motifs from Positive and Negative Examples

Erika Tateishi[*], Osamu Maruyama[*]and Satoru Miyano[†]

[*]Department of Information Systems, Kyushu University 39, Kasuga 816, Japan

[†]Research Institute of Fundamental Information Science
Kyushu University 33, Fukuoka 812, Japan

Abstract

We define the best consensus motif (BCM) problem motivated by the problem of extracting motifs from nucleic acid and amino acid sequences. A type over an alphabet Σ is a family Ω of subsets of Σ^*. A motif π of type Ω is a string $\pi = \pi_1 \cdots \pi_n$ of motif components, each of which stands for an element in Ω. The BCM problem for Ω is, given a yes-no sample $S = \{(\alpha^{(1)}, \beta^{(1)}), \ldots, (\alpha^{(m)}, \beta^{(m)})\}$ of pairs of strings in Σ^* with $\alpha^{(i)} \neq \beta^{(i)}$ for $1 \leq i \leq m$, to find a motif π of type Ω that maximizes the number of good pairs in S, where $(\alpha^{(i)}, \beta^{(i)})$ is good for π if π accepts $\alpha^{(i)}$ and rejects $\beta^{(i)}$. We prove that the BCM problem is NP-complete even for a very simple type $\Omega_1 = \{z \mid \emptyset \neq z \subseteq \Sigma\}$, which is used, in practice, for describing protein motifs in the PROSITE database. We also show that the NP-completeness of the problem does not change for the type $\Omega_\infty = \Omega_1 \cup \{\Sigma^+\} \cup \{\Sigma^{[i,j]} \mid 1 \leq i \leq j\}$, where $\Sigma^{[i,j]}$ is the set of strings over Σ of length between i and j. Furthermore, for the BCM problem for Ω_1, we provide a polynomial-time greedy algorithm based on the probabilistic method. Its performance analysis shows an explicit approximation ratio of the algorithm.

Keywords: algorithms and computational complexity, genome informatics

1 Introduction

PROSITE Database [3] compiles various important "motifs" of amino acid sequences of proteins. Most motifs have been extracted by biological experiments and alignment techniques tuned up with knowledge on molecular biology. Such motifs are expressed in a systematic way in PROSITE. For example, a motif of some family of DNA-binding proteins is expressed as a pattern W-$x(2)$-[LI]-[SAG]-$x(4,5)$-R-$x(8)$-[YW]-$x(3)$-[LIVM]. Here each component in the motif is separated from its neighbor by '-'. The component of the form [*symbols*] represents that any of the symbols between the brakets is accepted for the position. For example [SAG] stands for "S or A or G." The component $x(i,j)$ ($x(l)$) represents any sequences of length between i and j (of length l). Another example of a motif is [AC]-x-V-$x(4)$-{ED}, where x is used for a position where any symbol is accepted and {ED} stands for any symbol except E and D. Thus this can be translated as [A or C]-any-V-any-any-any-any-{any but E or D}.

Finding such motifs from sequence data is a very crucial problem in Genome Informatics/Molecular Bioinformatics since motifs provide us biologically important knowledge in terms of amino acid residues. Traditionally, various alignment techniques [5], which shall find a "reasonable" common subsequence for a family of amino acid sequences or nucleic acid sequences, have played a significant role in discovering these motifs from sequence data.

In this paper we consider the problem of finding such "motifs" from a collection of positive and negative examples while alignment techniques deal with only positive examples. We call a string $\pi = \pi_1 \cdots \pi_n$ a *motif*, where $\pi_1, \ldots, \pi_n$ stand for such components as V, [SAG], {ED}, $x(i,j)$, $x(l)$ and x. A *yes-no sample* is a set S of pairs $(\alpha^{(1)}, \beta^{(1)}), \ldots, (\alpha^{(m)}, \beta^{(m)})$ of positive

[1] Corresponding author: Erika Tateishi, Research Institute of Fundamental Information Science, Kyushu University 33, Fukuoka 812, Japan. Email: tateishi@rifis.kyushu-u.ac.jp.

and negative examples. A pair $(\alpha^{(i)}, \beta^{(i)})$ in S is said to be *good* for a motif π if π accepts $\alpha^{(i)}$ and rejects $\beta^{(i)}$. The *best consensus motif* (BCM) problem is, given a yes-no sample S, to find a motif that maximizes the number of good pairs in S. Some interesting results have been shown in [6, 7] for the consistency problem of regular patterns [1] from the point of computational learning theory. But regular patterns do not allow ambiguity in their expressions. An important motivation for defining the BCM problem is that sequences in the biological databases such as PIR, PDB, GenBank, etc., involve diversity and uncertainty.

The problem of finding, if any, a motif π such that *all* yes-no examples in S are good for π can be solved in polynomial time if the motif is restricted to use only components of the form [*symbols*]. In contrast, we show that the BCM problem for this special case turns NP-complete as in the case of MAX2SAT [4]. More generally, the BCM problem for motifs with [*symbols*], $x(i, j)$, $x(l)$, x and x^+ is also shown NP-complete, where x^+ represents the set of all nonempty strings.

For such obstacles to knowledge discovery from sequences, a big challenge is to devise efficient heuristic algorithms for approximating the BCM problem. We do not know whether it is possible to obtain any polynomial-time algorithm with any explicit performance guarantee. This paper answers a special case of the problem by showing a simple greedy algorithm approximating the BCM problem for motifs with components [*symbols*]. We prove a lower bound of its performance by a rigorous analysis of the algorithm with a probabilistic method in [14].

This paper is organized as follows: In Section 2, we define the best consensus motif problem. The NP-completeness of the problem is proved in Section 3. In Section 4, we give a greedy algorithm for the special case of the BCM problem and its performance analysis.

2 Best Consensus Motif Problem

For an alphabet Δ, we denote by Δ^* the set of all strings over Δ. The length of a string w in Δ^* is denoted by $|w|$. We denote $\Delta^+ = \Delta^* - \{\varepsilon\}$ (ε is the empty string). For $1 \leq l$ and $1 \leq i < j$, let $\Delta^l = \{w \in \Delta^* \mid |w| = l\}$ and $\Delta^{[i,j]} = \{w \in \Delta^* \mid i \leq |w| \leq j\}$. For a set S, the number of elements in S is also denoted by $|S|$.

Definition 1 Let Σ be a finite alphabet and Ω be a family of subsets of Σ^* called a *type over* Σ. Let V be a set whose elements are called *motif components* and τ be a mapping from V to Ω. For each motif component $x \in V$, the *range* of x is the set $\tau(x)$ in Ω which specifies the set of strings in Σ^* that the motif component x can get. A *motif* π of type Ω is an expression of the form

$$\pi = \pi_1 \cdots \pi_n,$$

where $\pi_1, \ldots, \pi_n$ are mutually distinct motif components. We denote by $L(\pi)$ the set of strings defined by $\{w_1 \cdots w_n \mid w_1 \in \tau(\pi_1), \ldots, w_n \in \tau(\pi_n)\}$. For a string w and a motif π, we say that π *accepts* (*rejects*) w if $w \in L(\pi)$ ($w \notin L(\pi)$).

A motif component x whose range $\tau(x)$ consists of a single element, say $w \in \Sigma^*$, is called a *constant* and denoted by w if no confusion occurs. On the other hand, if the range $\tau(x)$ consists of at least two elements, x is called a *variable*. We often confuse a motif component with its range without any notice.

Example 1 A regular pattern [1] is an expression of the form $\pi = w_0 x_1 w_1 \cdots w_{n-1} x_n w_n$ consisting of constants $w_0, \ldots, w_n \in \Sigma^*$ and distinct variables $x_1, \ldots, x_n$ with range Σ^+. The pattern π defines a set $L(\pi)$ of strings in Σ^* obtained by substituting any strings in Σ^+ to the variables $x_1, \ldots, x_n$. Thus any regular pattern is regarded as a motif of type $\Omega = \{\Sigma^+\} \cup \{\{w\} \mid w \in \Sigma^*\}$.

Example 2 Let Σ be the set of 20 symbols representing the amino acid residues. Let $\Omega_1 = \{z \mid \emptyset \neq z \subseteq \Sigma\}$ be the family of all nonempty subsets of Σ. Then the leucine zipper L-$x(6)$-L-$x(6)$-L-$x(6)$-L, a 'helix-turn-helix' motif of some family of DNA binding proteins [LIVM]-$x(1)$-[DE]-[LIVM]-A-$x(2)$-[STAG]-x-V-[SP]-$x(2)$-[STAG]-[LIVMA]-$x(2)$-[LIVMA]-[LIVM] and a motif such as [AC]-$x(1)$-V-$x(4)$-{ED} in PROSITE [3] can be regarded as motifs of type Ω_1, where $x(i)$ represents any strings of length i and the expressions like [LIVM] represents "L or I or V or M" and and $\{$ED$\}$ represents "any but E or D." Note that $x(1)$ corresponds to $\Sigma \in \Omega_1$ and $x(i)$ is expressed as $\overbrace{x(1) \cdots x(1)}^{i}$. In order to express the zinc finger motif C-$x(2,4)$-C-$x(12)$-H-$x(3,5)$-H, we need $\Omega = \{\Sigma^{[i,j]} \mid 1 \leq i \leq j\} \cup \Omega_1$ as its type.

Since these motifs are embedded in sequences, we should add motif components with range Σ^+ at the left and right ends of the motifs when we consider the whole sequences containing these motifs. Thus it suffices to consider the type $\Omega_\infty = \{\Sigma^+\} \cup \{\Sigma^{[i,j]} \mid 1 \leq i \leq j\} \cup \{z \mid \emptyset \neq z \subseteq \Sigma\}$ for expressing motifs.

A *yes-no example* is a pair (α, β) of strings in Σ^* with $\alpha \neq \beta$. For a motif π and a yes-no example (α, β), we say that (α, β) is *good* for π if π accepts α but rejects β. A *yes-no sample* is a set $S = \{(\alpha^{(1)}, \beta^{(1)}), \ldots, (\alpha^{(m)}, \beta^{(m)})\}$ of yes-no examples. We call strings $\alpha^{(1)}, \ldots, \alpha^{(m)}$ the *positive examples* of S and strings $\beta^{(1)}, \ldots, \beta^{(m)}$ the *negative examples* of S. For a motif π and a yes-no sample S, we define $\mathrm{cost}(S, \pi)$ to be the number of yes-no examples in S which are good for π. Note that $\mathrm{cost}(S, \pi) = |L(\pi) \cap P| \times |(\Sigma^* - L(\pi) \cap N|$ if a yes-no sample S is provided as $P \times N$ with two disjoint sets P and N of strings.

Let Ω be a type over Σ. The *best consensus motif problem for type* Ω is, given a yes-no sample S, to find a motif π of type Ω that maximizes $\mathrm{cost}(S, \pi)$.

3 Best Consensus Motif Problem is Intractable

This section investigates the complexity of finding a best consensus motif from a yes-no sample by considering two kinds of types. The one is

$$\Omega_1 = \{z \mid \emptyset \neq z \subseteq \Sigma\}$$

which allows only simple motif components and the other is

$$\Omega_\infty = \{\Sigma^+\} \cup \{\Sigma^{[i,j]} \mid 1 \leq i \leq j\} \cup \{z \mid \emptyset \neq z \subseteq \Sigma\}$$

which involves a variety of expressions. We consider the decision version of the best consensus motif problem for Ω_1 (Ω_∞), which is, given a yes-no sample S and an integer $K \geq 0$, to decide if there is a motif π of type Ω_1 (Ω_∞) such that $\mathrm{cost}(S, \pi) \geq K$.

We start with an observation on the type Ω_1. Let $S = \{(\alpha^{(1)}, \beta^{(1)}), \ldots, (\alpha^{(m)}, \beta^{(m)})\}$ be a yes-no sample. We can assume that $|\alpha^{(1)}| = \cdots = |\alpha^{(m)}|$ since the length of a motif of type Ω_1 must be the same as the length of $\alpha^{(i)}$ if $(\alpha^{(i)}, \beta^{(i)})$ is good for μ. The problem of finding, if any, a motif μ^* of type Ω_1 such that *all* yes-no examples in S are good for μ^* can be solved in polynomial time. This is because μ^* must be of the form $\mu^* = \tilde{\mu}_1 \cdots \tilde{\mu}_n$ with $\tilde{\mu}_k = \{\alpha_k^{(i)} \mid 1 \leq i \leq m\}$ for $1 \leq k \leq n$ and μ^* must reject all negative examples from S, where $\alpha^{(i)} = \alpha_1^{(i)} \cdots \alpha_n^{(i)}$ with $\alpha_1^{(i)}, \ldots, \alpha_n^{(i)} \in \Sigma$. We do not know wheter the same observation holds for the type Ω_∞.

The main result in this section asserts that the best consensus motif problem is computationally intractable even if the type is simple (Ω_1) or ample (Ω_∞).

Theorem 1

(1) The best consensus motif problem for $\Omega_1 = \{z \mid \emptyset \neq z \subseteq \Sigma\}$ is NP-complete.

(2) The best consensus motif problem for Ω_∞ is NP-complete.

The results also hold even if $\Sigma = \{0, 1\}$ and a yes-no sample S is provided as $P \times N$ with two disjoint sets P and N.

Proof. For a technical reason, we shall prove (2) first. The proof of (1) shall be given as a modified version of the proof of (2).

Obviously the problem is in NP. We shall show that there is a polynomial-time reduction from MAX2SAT [4], which is, given a formula in 2-CNF and a positive integer K, to decide if there is a truth assignment satisfying

$$
\begin{array}{lclllll}
s_1 & = & \overbrace{1000\cdots00}^{\overset{(1)}{2n}} & \overbrace{1000\cdots00}^{\overset{(2)}{2n}} & \cdots & \overbrace{1000\cdots00}^{\overset{(M)}{2n}} & \\
s_2 & = & 0100\cdots00 & 0100\cdots00 & \cdots & 0100\cdots00 & \\
\vdots & & \ddots & \ddots & \vdots & \ddots & \\
s_{2n} & = & 0000\cdots01 & 0000\cdots01 & \cdots & 0000\cdots01 & \\
e_1 & = & 0000\cdots00 & 0000\cdots00 & \cdots & 0000\cdots00 & \\
e_2 & = & 0000\cdots00 & 0000\cdots00 & \cdots & 0000\cdots00 & 0 \\
e_3 & = & 0000\cdots00 & 0000\cdots00 & \cdots & 0000\cdots00 & 00 \\
\vdots & & \vdots & \vdots & \vdots & \vdots & \ddots \\
e_L & = & \overbrace{0000\cdots00}^{2n} & \overbrace{0000\cdots00}^{2n} & \cdots & \overbrace{0000\cdots00}^{2n} & \overbrace{0000000\cdots0000}^{L-1}
\end{array}
$$

Figure 1: Positive examples.

at least K clauses. Let $F = C_1 \cdots C_m$ be a formula in 2-CNF with variables $u_1, u_2, \ldots, u_n$. We may assume without loss of generality that any clause does not contain both u_k and $\overline{u_k}$ for any $1 \leq k \leq n$. We shall define P of positive examples and N of negative examples such that at least K clauses of F are satisfiable if and only if there is a motif π with $\text{cost}(S, \pi) \geq (n+L)(3nM+K)$, where $S = P \times N$, $L = n^2 - n + 2$ and $M = (2n+L)m+1$. Let $K' = (n+L)(3nM+K)$. The set P consists of the following $2n+L$ positive examples $s_1, \ldots, s_{2n}, e_1, \ldots, e_L$ and the set N consists of the following $3nM+m$ negative examples $t_1^1, \ldots, t_M^n, h_1^1, \ldots, h_M^{2n}, d_1, \ldots, d_m$ (see Fig. 1 and Fig. 2):

$$
\begin{array}{ll}
s_i = (0^{i-1}10^{2n-i})^M & \text{for } 1 \leq i \leq 2n, \\
e_i = 0^{2nM+i-1} & \text{for } 1 \leq i \leq L, \\
t_j^k = (0^{2n})^{j-1} \cdot (00)^{k-1}11(00)^{n-k} \cdot (0^{2n})^{M-i} & \\
 & \text{for } 1 \leq k \leq n,\ 1 \leq j \leq M, \\
h_j^i = (0^{2n})^{j-1} \cdot (0^{i-1} \cdot 1 \cdot 0^{2n-i}) \cdot (0^{2n})^{M-j-1} \cdot 0^{2n-1} & \\
 & \text{for } 1 \leq i \leq 2n,\ 1 \leq j < M, \\
h_M^i = (0^{2n})^{M-1} \cdot 0^{i-1} \cdot 1 \cdot 0^{2n-i-1} & \text{for } 1 \leq i < 2n, \\
h_M^{2n} = 0^{2nM-1}, & \\
d_i = (r_1^i r_2^i \ldots r_n^i)^M & \text{for } 1 \leq i \leq m, \text{ where} \\
 & r_k^i = \begin{cases} 10 & \text{if literal } u_k \text{ appears in } C_i \\ 01 & \text{if literal } \overline{u_k} \text{ appears in } C_i \\ 0 & \text{otherwise.} \end{cases}
\end{array}
$$

Note that positive and negative examples are strings over $\Sigma = \{0, 1\}$. Therefore Ω_∞ consists of $\{0\}$, $\{1\}$, $\{0,1\}^+$ and $\{0,1\}^{[i,j]}$ for $j \geq i \geq 1$. In the following argument, ω^+ represents a variable whose range is $\{0,1\}^+$ and $\omega^{[i,j]}$ represents a variable whose range is $\{0,1\}^{[i,j]}$ for $j \geq i \geq 1$. The constants are 0 and 1. Let X denote the set of variables. For convenience, we denote by $X \cup 0$ the set consisting of all variables and constants 0.

Suppose that at least K clauses of F are satisfied by a truth assignment $\hat{u}_1, \ldots, \hat{u}_n$. Then we define a motif $\pi = (\tau_1 \ldots \tau_n)^M$ by putting $\tau_k = 0\omega^+$ if $\hat{u}_k = \mathit{true}$ and $\tau_k = \omega^+ 0$ if $\hat{u}_k = \mathit{false}$ for $1 \leq k \leq n$. Obviously, either s_{2k-1} or s_{2k} is accepted by π for $1 \leq k \leq n$, and all $e_1, \ldots, e_L$ are accepted by π. On the other hand, π rejects t_j^k for all $1 \leq k \leq n$, $1 \leq j \leq M$ and all h_j^i for all $1 \leq i \leq 2n$, $1 \leq j \leq M$. By the definition of r_k^i, a clause C_i of F is satisfied by the truth assignment $\hat{u}_1, \ldots, \hat{u}_n$ if and only if d_j is rejected by π. Thus the number of clauses satisfied by the truth assignment is equal to the number of d_j's rejected by π. Then $\text{cost}(S, \pi) \geq K'$.

Conversely, we suppose that there exists a motif π with $\text{cost}(S, \pi) \geq K'$.

First, we show Claim 1 below. For Claim 1, we introduce the following conventions. A motif is denoted as $\pi = \theta_0 y_1 \theta_1 \cdots \theta_{n-1} y_n \theta_n$, where y_i is a variable in X for $1 \leq i \leq n$ and θ_j is a finite (possibly empty) sequence of constants 0 and 1 for $0 \leq j \leq n$. For a variable $y_k \in X$, let $\|y_k\| = \min\{|w| \mid w \in y_k\}$ for $1 \leq k \leq n$. Then we denote by $\|\pi\| = |\theta_0| + \|y_1\| + |\theta_1| + \cdots |\theta_{n-1}| + \|y_n\| + |\theta_n|$. For $\pi = \theta_0 y_1 \theta_1 \cdots \theta_{n-1} y_n \theta_n$, a tuple $\varphi = [y_1 \leftarrow w_1, \ldots, y_n \leftarrow w_n]$ of assignments of strings $w_i \in y_i$ to the variable occurrences y_i for $1 \leq i \leq n$ is called a *fill-in of* π. Let $w = \theta_0 w_1 \theta_1 \cdots \theta_{n-1} w_n \theta_n$ be the string, denoted by $\pi\varphi$, obtained by the fill-in φ. We say that an occurrence of a substring w' in w is *covered* by the variable occurrence y_i under φ if w' is contained in w_i.

CLAIM 1. If a motif π satisfies $\text{cost}(S, \pi) \geq K'$, then $\pi \in (X \cup 0)^+$ and $\|\pi\| = 2nM$.

$$\left[\begin{array}{lclllll}
t_1^1 & = & \overbrace{1100\cdots00}^{\substack{(1)\\2n}} & \overbrace{0000\cdots00}^{\substack{(2)\\2n}} & \cdots & \cdots & \overbrace{0000\cdots00}^{\substack{(M)\\2n}}\\
t_2^1 & = & 0000\cdots00 & 1100\cdots00 & \cdots & \cdots & 0000\cdots00\\
\vdots & & \vdots & \vdots & \vdots & & \vdots\\
t_M^1 & = & 0000\cdots00 & 0000\cdots00 & \cdots & \cdots & 1100\cdots00
\end{array}\right.$$

$$\left[\begin{array}{lclllll}
t_1^2 & = & 0011\cdots00 & 0000\cdots00 & \cdots & \cdots & 0000\cdots00\\
t_2^2 & = & 0000\cdots00 & 0011\cdots00 & \cdots & \cdots & 0000\cdots00\\
\vdots & & \vdots & \vdots & \vdots & & \vdots\\
t_M^2 & = & 00\cdots0011 & 0000\cdots00 & \cdots & \cdots & 0011\cdots00
\end{array}\right.$$

$$\vdots$$

$$\left[\begin{array}{lclllll}
t_1^n & = & 00\cdots0011 & 00\cdots0000 & \cdots & \cdots & 00\cdots0000\\
t_2^n & = & 00\cdots0000 & 00\cdots0011 & \cdots & \cdots & 00\cdots0000\\
\vdots & & \vdots & \vdots & \vdots & & \vdots\\
t_M^n & = & 00\cdots0000 & 00\cdots0000 & \cdots & \cdots & 00\cdots0011
\end{array}\right.$$

$$\left[\begin{array}{lclllll}
d_1 & = & r_1^1r_2^1\cdots r_n^1 & r_1^1r_2^1\cdots r_n^1 & \cdots & & r_1^1r_2^1\cdots r_n^1\\
d_2 & = & r_1^2r_2^2\cdots r_n^2 & r_1^2r_2^2\cdots r_n^2 & \cdots & & r_1^2r_2^2\cdots r_n^2\\
\vdots & & \vdots & \vdots & \vdots & & \vdots\\
d_m & = & r_1^mr_2^m\cdots r_n^m & r_1^mr_2^m\cdots r_n^m & \cdots & & r_1^mr_2^m\cdots r_n^m
\end{array}\right.$$

$$\left[\begin{array}{lclllll}
h_1^1 & = & \overbrace{100\cdots000}^{2n} & \overbrace{000\cdots000}^{2n} & \cdots & \overbrace{000\cdots000}^{2n} & \overbrace{000\cdots00}^{2n-1}\\
h_2^1 & = & 000\cdots000 & 100\cdots000 & \cdots & 000\cdots000 & 000\cdots00\\
\vdots & & \vdots & \vdots & \vdots & & \vdots\\
h_M^1 & = & 000\cdots000 & 000\cdots000 & \cdots & 000\cdots000 & 100\cdots00
\end{array}\right.$$

$$\vdots$$

$$\left[\begin{array}{lclllll}
h_1^{2n-1} & = & 000\cdots010 & 000\cdots000 & \cdots & 000\cdots000 & 000\cdots00\\
h_2^{2n-1} & = & 000\cdots000 & 000\cdots010 & \cdots & 000\cdots000 & 000\cdots00\\
\vdots & & \vdots & \vdots & \vdots & & \vdots\\
h_{M-1}^{2n-1} & = & 000\cdots000 & 000\cdots000 & \cdots & 000\cdots010 & 000\cdots00\\
h_M^{2n-1} & = & 000\cdots000 & 000\cdots000 & \cdots & 000\cdots000 & 000\cdots01
\end{array}\right.$$

$$\left[\begin{array}{lclllll}
h_1^{2n} & = & 000\cdots001 & 000\cdots000 & \cdots & 000\cdots000 & 000\cdots00\\
h_2^{2n} & = & 000\cdots000 & 000\cdots001 & \cdots & 000\cdots000 & 000\cdots00\\
\vdots & & \vdots & \vdots & \vdots & & \vdots\\
h_{M-1}^{2n} & = & 000\cdots000 & 000\cdots000 & \cdots & 000\cdots001 & 000\cdots00\\
h_M^{2n} & = & 000\cdots000 & 000\cdots000 & \cdots & 000\cdots000 & 000\cdots00
\end{array}\right.$$

Figure 2: Negative examples.

Proof. If 1 occurs in π, then π rejects all e_i for $1 \leq i \leq L$. Then $\mathrm{cost}(S, \pi) \leq 2n(3nM + m) < K'$, by the definition of M and K'. Thus $\pi \in (X \cup 0)^+$. If π rejects all s_i for $1 \leq i \leq 2n$, then $\mathrm{cost}(S, \pi) \leq L(3nM+m) < K'$. Therefore, $\|\pi\| \leq 2nM$ and s_i is in $L(\pi)$ for some $1 \leq i \leq 2n$.

Suppose now that $\|\pi\| \leq 2nM - 1$. Let $\pi = \theta_0 y_1 \theta_1 \cdots y_p \theta_p$. Since s_i is accepted by π, there is a fill-in $\varphi = [y_1 \leftarrow \xi_1, \ldots, y_p \leftarrow \xi_p]$ such that $\pi\varphi = s_i$.

Fact 1. If π accepts s_i, then π accepts h_j^i or h_j^{i-1} for each $1 \leq j \leq M$, where $h_j^0 = h_{j-1}^{2n}$.

Proof. Note that s_i contains exactly M occurrences of 1. For each $1 \leq j \leq M$, we consider the jth occurrence of 1 in s_i. Let y_q be the variable occurrence covering the jth occurrence of 1 in s_i under φ. There are two cases according to the location of the jth occurrence of 1 in ξ_q.

(1) $\xi_q = 1$: Since $\|\pi\| \leq 2nM - 1$ and $|s_i| = 2nM$, there is some r with $|\xi_r| \geq 2$ such that $y_r = \omega^+$ or $y_r = \omega^{[l_1, l_2]}$ with $l_1 < |\xi_r| \leq l_2$. When $q < r$, let $\xi'_q = 1$, $\xi'_r = 0^{|\xi_r|-1}$, and $\xi'_l = 0^{|\xi_l|}$ $(l \neq q, r)$. Then for $\varphi' = [y_1 \leftarrow \xi'_1, \ldots, y_p \leftarrow \xi'_p]$ we have $\pi\varphi' = h_j^i$. When $r < q$, let $\xi'_q = 1$, $\xi'_r = 0^{|\xi_r|-1}$, and $\xi'_l = 0^{|\xi_l|}$ $(l \neq q, r)$. Then for $\varphi' = [y_1 \leftarrow \xi'_1, \ldots, y_p \leftarrow \xi'_p]$ we have $\pi\varphi' = h_j^{i-1}$.

(2) $\xi_q = \gamma 10 \gamma'$ or $\xi_q = \gamma 01 \gamma'$: If $y_q = \omega^+$ or $y_q = \omega^{[l_1, l_2]}$ with $l_1 < |\xi_q|$, then let $\xi'_q = 0^{|\gamma|}10^{|\gamma'|}$ and $\xi'_l = 0^{|\xi_l|}$ $(l \neq q)$ and let $\varphi' = [y_1 \leftarrow \xi'_1, \ldots, y_p \leftarrow \xi'_p]$. Then if $\xi_q = \gamma 10 \gamma'$, then $\pi\varphi' = h_j^i$. If $\xi_q = \gamma 01 \gamma'$, then $\pi\varphi' = h_j^{i-1}$. If $y_q = \omega^{[l_1, l_2]}$ with $l_1 = |\xi_q|$, then by the same argument as (1) we can conclude that π accepts h_j^i or h_j^{i-1}. □

Fact 2. If π accepts b positive examples from $s_1, \ldots, s_{2n}$, then π also accepts at least $\lfloor \frac{b}{2} \rfloor \cdot M$ negative examples from $h_1^1, \ldots, h_{M-1}^{2n}$.

Proof. Let $s_{i_1}, \ldots, s_{i_b}$ be positive examples accepted by π, where $i_1 < i_2 < \cdots < i_b$. By Fact 1, for each s_{i_k}, π accepts M distinct negative examples $h_j^{i_k-1}$ or $h_j^{i_k}$ for $1 \leq j \leq M$. Let H_k be the set of these M negative examples $h_j^{i_k-1}$ or $h_j^{i_k}$ $(1 \leq j \leq M)$ accepted by π for s_{i_k}.

Then $H_1, H_3, \ldots, H_{2 \cdot \lfloor \frac{b}{2} \rfloor + 1}$ are pairwise disjoint. Therefore $|H_1 \cup H_3 \cup \cdots \cup H_{2 \cdot \lfloor \frac{b}{2} \rfloor + 1}| = \lfloor \frac{b}{2} \rfloor \cdot M$. □

By Fact 2, if π accepts b positive examples from $s_1, \ldots, s_{2n}$, then

$$\mathrm{cost}(S, \pi) \leq (L + b)((3n - \lfloor \tfrac{b}{2} \rfloor)M + m)$$

since at least $\lfloor \frac{b}{2} \rfloor \cdot M$ negative examples from $h_1^1, \ldots, h_{M-1}^{2n}$ are accepted by π. By easy but tedious calculations, we can show that $(L + b)((3n - \lfloor \frac{b}{2} \rfloor)M + m) < K'$ for all $1 \leq b \leq 2n$. This is again a contradiction. Thus $\|\pi\|$ must be exactly $2nM$. □

CLAIM 2. If $\mathrm{cost}(S, \pi) \geq K'$, π accepts at least n positive examples from $s_1, \ldots, s_{2n}$.

Proof. Otherwise, $\mathrm{cost}(S, \pi) \leq (L + n - 1)(3nM + m) < K'$. □

CLAIM 3. Let π be a motif that maximizes $\mathrm{cost}(S, \pi)$. If $\mathrm{cost}(S, \pi) \geq K'$, then π can be described as a motif of type $\{\{0, 1\}^+\}$ having the form

$$\pi = (\pi_1 \pi_2 \cdots \pi_{2n})^M,$$

where $\pi_{2k-1}\pi_{2k} = 0\omega^+$ or $\pi_{2k-1}\pi_{2k} = \omega^+0$ for $1 \leq k \leq n$.

Proof. By Claim 1, the motif π satisfies $\|\pi\| = 2nM$ but may contain variables and $\omega^{[l_1, l_2]}$. However, by replacing all occurrences $\omega^{[l_1]}$ $(\omega^{[l_1, l_2]})$ in π with $\overbrace{\omega^+ \cdots \omega^+}^{l_1}$, we obtain a motif $\tilde{\pi}$ of type $\{\{0, 1\}^+\}$ such that $|\tilde{\pi}| = 2nM$ and $\tilde{\pi}$ accepts the same positive examples in P as π and $\tilde{\pi}$ rejects the same negative examples in N as π. Let π be this $\tilde{\pi}$. Since $|\pi| = 2nM$, π is written as $\pi = \pi_1\pi_2 \cdots \pi_{2nM}$ with $\pi_i \in X$ or $\pi_i = 0$ for $1 \leq i \leq 2nM$. For each $1 \leq k \leq n$, we consider positive examples s_{2k-1} and s_{2k} and segments $\pi_{2n(i-1)+2k-1}\pi_{2n(i-1)+2k}$ for $1 \leq i \leq M$.

Case 1. $s_{2k-1} \in L(\pi)$ and $s_{2k} \in L(\pi)$: Then $\pi_{2n(i-1)+2k-1}\pi_{2n(i-1)+2k} = \omega^+\omega^+$ for all $1 \leq i \leq M$. If we replace $\pi_{2n(i-1)+2k-1}\pi_{2n(i-1)+2k}$ with $0\omega^+$ (or ω^+0) in π for all $1 \leq i \leq M$. Let π' be the resulting motif obtained from π by this replacement. Then π' rejects s_{2k-1} (or s_{2k}) and M negative examples $t_1^k, \ldots, t_M^k$ in addition to all positive and negative examples rejected by π. On the other hand, π accepts all e_i for $1 \leq i \leq L$ since $\pi \in (X \cup 0)^+$. Moreover, by Claim 2, π must accepts at least n positive examples from $s_1, \ldots, s_{2n}$. By assumption, π accepts s_{2k-1}, s_{2k} and $t_{(k-1)M+1}, \ldots, t_{kM}$. Therefore $\mathrm{cost}(S, \pi') - \mathrm{cost}(S, \pi) \geq (n + L)M - 1 \cdot ((3n - 1)M + m) > 0$, by the definition of L and M. This contradicts that π maximizes $\mathrm{cost}(S, \pi)$.

Case 2. $s_{2k-1} \notin L(\pi)$ and $s_{2k} \notin L(\pi)$: By Claim 2, π must accept at least n positive examples from $s_1, \ldots, s_{2n}$. Therefore there is some $1 \leq j \leq n$ such that $s_{2j-1} \in L(\pi)$ and $s_{2j} \in L(\pi)$. But this does not hold from Case 1.

Thus either $s_{2k-1} \in L(\pi)$ or $s_{2k} \in L(\pi)$ for $1 \leq k \leq n$.

Case 3. $s_{2k-1} \in L(\pi)$ and $s_{2k} \notin L(\pi)$: Then $\pi_{2n(i-1)+2k-1}\pi_{2n(i-1)+2k}$ is either $\omega^+\omega^+$ or ω^+0 for $1 \leq i \leq M$. If $\pi_{2n(i-1)+2k-1}\pi_{2n(i-1)+2k} = \omega^+\omega^+$ for some $1 \leq i \leq M$, then we replace $\pi_{2n(i-1)+2k-1}\pi_{2n(i-1)+2k}$ with ω^+0 in π. Let π' be the resulting motif. Then π' does not reject any positive examples which are accepted by π and π' does not accept any negative examples which are rejected by π. In addition, π' rejects t_i^k. Therefore $\mathrm{cost}(S,\pi') > \mathrm{cost}(S,\pi)$. This also contradicts the hypothesis that $\mathrm{cost}(S,\pi)$ is maximum. Thus $\pi_{2n(i-1)+2k-1}\pi_{2n(i-1)+2k} = \omega^+0$ for all $1 \le i \le M$.

Case 4. $s_{2k-1} \notin L(\pi)$ and $s_{2k} \in L(\pi)$: In the same way as Case 3, we can see that $\pi_{2n(i-1)+2k-1}\pi_{2n(i-1)+2k} = 0\omega^+$ for all $1 \le i \le M$.

By Cases 1–4, we see that π is written as $(\pi_1\pi_2\cdots\pi_{2n})^M$ with $\pi_{2k-1}\pi_{2k} = 0\omega^+$ or $\pi_{2k-1}\pi_{2k} = \omega^+0$ for $1 \le k \le n$. □

By Claim 3, there is a motif $\pi = (\pi_1\cdots\pi_{2n})^M$ such that $\mathrm{cost}(S,\pi) \ge K'$ and $\pi_{2k-1}\pi_{2k}$ is $0\omega^+$ or ω^+0 for $1 \le k \le n$. The motif π accepts either s_{2k-1} or s_{2k} for each $1 \le k \le n$ and accepts all other positive examples while π rejects at least all negative examples except $d_1,\ldots,d_m$. Since $\mathrm{cost}(S,\pi) \ge K' = (n+L)(3nM+K)$, at least K negative examples from $d_1,\ldots,d_m$ must be rejected by π. Then we define a truth assignment $\hat{u}_k = \mathit{true}$ if $\pi_{2k-1}\pi_{2k} = 0\omega^+$ and $\hat{u}_k = \mathit{false}$ if $\pi_{2k-1}\pi_{2k} = \omega^+0$ for $1 \le k \le n$. Clearly, at least K clauses C_i are satisfied by this truth assignment. This completes the proof of (2).

We now give a proof of (1). Since the basic idea of the proof is similar to that of (2), we shall give only the sketch. Since the type is Ω_1, any variable occurrence in a motif gets a symbol in Σ. For a 2-CNF formula $F = C_1\cdots C_m$ with n variables and an integer K, let $M' = (2n-1)m+1$. Then let P' be the set of positive examples consisting of $e'_1, s'_1, \ldots, s'_{2n}$, where $e'_1, s'_1, \ldots, s'_{2n}$ are the strings in $\{0,1\}^*$ obtained from $e_1, s_1, \ldots, s_{2n}$ in Fig. 1 by replacing M with M'. Similarly, let N' be the set of negative examples consisting of $t'^1_1, \ldots, t'^n_{M'}$ and $d'_1, \ldots, d'_m$ obtained from $t^1_1, \ldots, t^n_M$ and $d_1, \ldots, d_m$ in Fig. 2 by replacing M with M'. Let $K'' = (n+1)(nM'+K)$. Assume that there is a motif π of type Ω_1 such that $\mathrm{cost}(S',\pi) \ge K''$. Let π be a motif that maximizing $\mathrm{cost}(S',\pi)$. If π contains constant 1, then at most one positive example is accepted by π. In this case, $\mathrm{cost}(S',\pi) \le nM' + m < K''$, a contradiction. Then by an argument similar to Claim 3, we can show that π must be of the form $\pi = (\eta_1\cdots\eta_n)^{M'}$ with $\eta_j = 0\omega^+$ or $\eta_j = \omega^+0$ for $1 \le j \le n$. The rest of the proof is similar and left to the reader. This completes the proof of (1). □

4 Approximating the Best Consensus Motif Problem for Ω_1

From Theorem 1, the best consensus motif problem for $\Omega_1 = \{z \mid \emptyset \ne z \subseteq \Sigma\}$ is hardly solvable in polynomial time. This section presents a polynomial time greedy algorithm for this problem and its performance analysis by exploiting the probabilistic method due to Yannakakis [14] which was first applied to the maximum satisfiability problem. Since we consider only sets in Ω_1, we assume that a yes-no sample $S = \{(\alpha^{(1)}, \beta^{(1)}), \ldots, (\alpha^{(m)}, \beta^{(m)})\}$ satisfies $|\alpha^{(1)}| = |\beta^{(1)}| = \cdots = |\alpha^{(m)}| = |\beta^{(m)}|$. For convenience, we identify the range of a motif component with the motif component itself.

For $1 \le k \le n$, we assume probabilities $p_k(z)$ for $z \in \Omega_1$, i.e., $0 \le p_k(z) \le 1$ and $\sum_{z\in\Omega_1} p_k(z) = 1$. Let π_k be a random variable taking a value z in Ω_1 with probability $p_k(z)$. We call an expression $\rho_k = z_1\cdots z_{k-1}\pi_k\cdots\pi_n$ a *random motif* of length n with fixed motif components $z_1, \ldots, z_{k-1} \in \Omega_1$ and random variables $\pi_k, \ldots, \pi_n$. For a random motif $z_1\cdots z_{k-1}\pi_k\cdots\pi_n$, we denote by $P((\alpha,\beta), z_1\cdots z_{k-1}\pi_k\cdots\pi_n)$ the probability that a yes-no example (α,β) is good for $z_1\cdots z_{k-1}\pi_k\cdots\pi_n$. Formally, let

$$H((\alpha,\beta), z_1\cdots z_{k-1}\pi_k\cdots\pi_n) = \{(y_k,\ldots,y_n) \in \Omega_1^{n-k+1} \mid (\alpha,\beta) \text{ is good for } z_1\cdots z_{k-1}y_k\cdots y_n\}.$$

Then $P((\alpha,\beta), z_1\cdots z_{k-1}\pi_k\cdots\pi_n)$ is given by

$$\sum_{(y_1,\ldots,y_n)\in H((\alpha,\beta),z_1\cdots z_{k-1}\pi_k\cdots\pi_n)} p_k(y_k)\cdots p_n(y_n).$$

It requires exponential time with respect to n if we simply calculate the above formula. However, it is polynomial time computable as shown below.

For $\sigma \in \Sigma$, let $S_\sigma = \{z \in \Omega_1 \mid \sigma \in z\}$ and $p_k(S_\sigma) = \sum_{z\in S_\sigma} p_k(z)$. For a string $\gamma = \gamma_1\cdots\gamma_n$, let $\delta(\gamma, z_1\cdots z_{k-1}) = 1$ if $\gamma_1\cdots\gamma_{k-1} \in z_1\cdots z_{k-1}$ else 0. Then for a yes-no example (α,β) $(\alpha = \alpha_1\cdots\alpha_n,\ \beta = \beta_1\cdots\beta_n)$, the probability that (α,β) is good for $z_1\cdots z_{k-1}\pi_k\cdots\pi_n$ is

```
Input: a yes-no sample $S$ and propabilities $p_1(z), \dots, p_n(z)$ for $z \in \Omega_1$.
Output: a motif $\hat{\pi} = z_1 \cdots z_n$.
/* $\pi_1, \dots, \pi_n$ are random variables with $p_1, \dots, p_n$, respectively. */
for $k \leftarrow 1$ to $n$
  begin
    Find $z \in \Omega_1$ maximizing $E(S, z_1 \cdots z_{k-1} z \pi_{k+1} \cdots \pi_n)$;
    $z_k \leftarrow z$
  end
```

Figure 3: Greedy algorithm.

expressed as:

$$
\begin{aligned}
& P((\alpha,\beta), z_1 \cdots z_{k-1}\pi_k \cdots \pi_n) \\
= \;& \delta(\alpha, z_1 \cdots z_{k-1}) \cdot \prod_{j \in \{j \geq k \mid \alpha_j = \beta_j\}} p_j(S_{\alpha_j}) \cdot \left(\prod_{j \in \{j \geq k \mid \alpha_j \neq \beta_j\}} p_j(S_{\alpha_j}) - \delta(\beta, z_1 \cdots z_{k-1}) \cdot \prod_{j \in \{j \geq k \mid \alpha_j \neq \beta_j\}} p_j(S_{\alpha_j} \cap S_{\beta_j}) \right) \\
= \;& \delta(\alpha, z_1 \cdots z_{k-1}) \cdot \left(\prod_{j=k}^{n} p_j(S_{\alpha_j}) - \delta(\beta, z_1 \cdots z_{k-1}) \cdot \prod_{j=k}^{n} p_j(S_{\alpha_j} \cap S_{\beta_j}) \right).
\end{aligned}
$$

This can be computed in polynomial time with respect to n. Therefore, for a yes-no sample S, the expectation

$$E(S, z_1 \cdots z_{k-1}\pi_k \cdots \pi_n) = \sum_{(\alpha,\beta) \in S} P((\alpha,\beta), z_1 \cdots z_{k-1}\pi_k \cdots \pi_n)$$

is also computable in polynomial time with respect to n and $m = |S|$.

Our algorithm for the best consensus motif problem for Ω_1 is very simple as shown in Fig. 3. We shall give a rigorous analysis of the algorithm and prove a lower bound of its performance.

Theorem 2 *Let $s = |\Sigma|$. If the probability distributions $p_1, \dots, p_n$ on Ω_1 satisfy the following conditions for $1 \leq k \leq n$,*

1. *$p_k(S_\sigma) \leq \frac{s+1}{2s}$ for all σ in Σ,*
2. *$p_k(S_\sigma \cap S_\tau) \geq \frac{s+2}{4s}$ for all σ, τ in Σ,*

then

$$\mathrm{cost}(S, \hat{\pi}) \geq E(S, \pi_1 \cdots \pi_n),$$

where $\hat{\pi}$ is a motif of type Ω_1 produced by the algorithm in Fig. 3 for a yes-no sample S and π_k is a random variable taking a value z with probablity $p_k(z)$ for $z \in \Omega_1$ $(1 \leq k \leq n)$.

Proof. Let $S = \{(\alpha^{(1)}, \beta^{(1)}), \dots, (\alpha^{(m)}, \beta^{(m)})\}$ be a yes-no sample. We denote $\alpha^{(i)}$ and $\beta^{(i)}$ as $\alpha^{(i)} = \alpha_1^{(i)} \cdots \alpha_n^{(i)}$ and $\beta^{(i)} = \beta_1^{(i)} \cdots \beta_n^{(i)}$, where $\alpha_k^{(i)}$ and $\beta_k^{(i)}$ are in Σ for $1 \leq k \leq n$.

Let

$$\mathrm{gain}(S, k, z) = E(S, z_1 \cdots z_{k-1} z \pi_{k+1} \cdots \pi_n) - E(S, z_1 \cdots z_{k-1}\pi_k\pi_{k+1} \cdots \pi_n)$$

for $z \in \Omega_1$ and $1 \leq k \leq n$. For the proof, it suffices to prove that for each $1 \leq k \leq n$ there is some z in Ω_1 such that $\mathrm{gain}(S, k, z) \geq 0$. In order to describe $P((\alpha^{(i)}, \beta^{(i)}), z_1 \cdots z_{k-1}\pi_k \cdots \pi_n)$ conveniently, let

$$
\begin{aligned}
A_k^{(i)} &= \prod_{j \in \{j \geq k \mid \alpha_j^{(i)} = \beta_j^{(i)}\}} p_j(S_{\alpha_j^{(i)}}), \\
B_k^{(i)} &= \prod_{j \in \{j \geq k \mid \alpha_j^{(i)} \neq \beta_j^{(i)}\}} p_j(S_{\alpha_j^{(i)}}), \\
C_k^{(i)} &= \prod_{j \in \{j \geq k \mid \alpha_j^{(i)} \neq \beta_j^{(i)}\}} p_j(S_{\alpha_j^{(i)}} \cap S_{\beta_j^{(i)}}).
\end{aligned}
$$

We first classify the set $I = \{1, \ldots, m\}$ of indices into the following three sets:

$$\begin{aligned}
I_1 &= \{i \mid \forall j_1(1 \leq j_1 < k)[\alpha_{j_1}^{(i)}, \beta_{j_1}^{(i)} \in z_{j_1}]\} \\
I_2 &= \{i \mid (\forall j_1(1 \leq j_1 < k)[\alpha_{j_1}^{(i)} \in z_{j_1}]) \wedge (\exists j_2(1 \leq j_2 < k)[\beta_{j_2}^{(i)} \notin z_{j_2}])\} \\
I_3 &= \{i \mid \exists j_1(1 \leq j_1 < k)[\alpha_{j_1}^{(i)} \notin z_{j_1}]\}
\end{aligned}$$

Then

$$P((\alpha^{(i)}, \beta^{(i)}), z_1 \cdots z_{k-1}\pi_k \cdots \pi_n) = \begin{cases} A_k^{(i)} \cdot (B_k^{(i)} - C_k^{(i)}) & \text{if } i \in I_1, \\ A_k^{(i)} \cdot B_k^{(i)} & \text{if } i \in I_2, \\ 0 & \text{if } i \in I_3. \end{cases}$$

By an easy calculation, for z in Ω_1 we have

$$\begin{aligned}
\mathrm{gain}(S, k, z) &= \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} = \beta_k^{(i)} \\ \alpha_k^{(i)} \in z}} (1 - p_k(S_{\alpha_k^{(i)}})) \cdot A_{k+1}^{(i)} \cdot (B_{k+1}^{(i)} - C_{k+1}^{(i)}) && (1) \\
&+ \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} \neq \beta_k^{(i)} \\ \alpha_k^{(i)}, \beta_k^{(i)} \in z}} A_{k+1}^{(i)} \cdot ((1 - p_k(S_{\alpha_k^{(i)}})) \cdot B_{k+1}^{(i)} - (1 - p_k(S_{\alpha_k^{(i)}} \cap S_{\beta_k^{(i)}})) \cdot C_{k+1}^{(i)}) && (2) \\
&+ \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} \neq \beta_k^{(i)} \\ \alpha_k^{(i)} \in z,\ \beta_k^{(i)} \notin z}} A_{k+1}^{(i)} \cdot ((1 - p_k(S_{\alpha_k^{(i)}})) \cdot B_{k+1}^{(i)} + C_k^{(i)}) && (3) \\
&- \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} \notin z}} A_k^{(i)} \cdot (B_k^{(i)} - C_k^{(i)}) && (4) \\
&+ \sum_{\substack{i \in I_2 \\ \alpha_k^{(i)} \in z}} (1 - p_k(S_{\alpha_k^{(i)}})) \cdot A_{k+1}^{(i)} \cdot B_{k+1}^{(i)} && (5) \\
&- \sum_{\substack{i \in I_2 \\ \alpha_k^{(i)} \notin z}} A_k^{(i)} \cdot B_k^{(i)}. && (6)
\end{aligned}$$

We shall show that

$$\sum_{z \subseteq \Sigma} |z| \cdot \mathrm{gain}(S, k, z) \geq 0$$

if $p_1, \ldots, p_n$ satisfy the conditions of the theorem. Then this implies that $\mathrm{gain}(S, k, z) \geq 0$ for some z in Ω_1.

We evaluate the above sum in the six parts (1)–(6) constituting $\mathrm{gain}(S, k, z)$.

Part (1): Let $D(i, k) = (1 - p_k(S_{\alpha_k^{(i)}})) \cdot A_{k+1}^{(i)} \cdot (B_{k+1}^{(i)} - C_{k+1}^{(i)})$.

$$\begin{aligned}
\sum_{r=1}^{s} r \sum_{\substack{z \subseteq \Sigma \\ |z| = r}} \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} = \beta_k^{(i)} \\ \alpha_k^{(i)} \in z}} D(i, k) &= \sum_{r=1}^{s} r \sum_{\sigma \in \Sigma} \sum_{\substack{z \subseteq \Sigma \\ |z| = r}} \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} = \beta_k^{(i)} \\ \alpha_k^{(i)} = \sigma \in z}} D(i, k) = \sum_{r=1}^{s} r \binom{s-1}{r-1} \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} = \beta_k^{(i)}}} D(i, k) \\
&= (s+1) \cdot 2^{s-2} \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} = \beta_k^{(i)}}} D(i, k) && (7)
\end{aligned}$$

Part (2): Let $E(i, k) = A_{k+1}^{(i)} \cdot ((1 - p_k(S_{\alpha_k^{(i)}})) \cdot B_{k+1}^{(i)} - (1 - p_k(S_{\alpha_k^{(i)}} \cap S_{\beta_k^{(i)}})) \cdot C_{k+1}^{(i)})$.

$$\sum_{r=1}^{s} r \sum_{\substack{z \subseteq \Sigma \\ |z| = r}} \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} \neq \beta_k^{(i)} \\ \alpha_k^{(i)} \in z \\ \beta_k^{(i)} \in z}} E(i, k) = \sum_{r=1}^{s} r \sum_{\sigma, \tau \in \Sigma} \sum_{\substack{z \subseteq \Sigma \\ |z| = r}} \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} \neq \beta_k^{(i)} \\ \alpha_k^{(i)} = \sigma \in z \\ \beta_k^{(i)} = \tau \in z}} E(i, k) = \sum_{r=1}^{s} r \binom{s-2}{r-2} \sum_{\substack{i \in I_1 \\ \alpha_k^{(i)} \neq \beta_k^{(i)}}} E(i, k)$$

$$= \quad (s+2)\cdot 2^{s-3} \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}\neq\beta_k^{(i)}}} E(i,k) \qquad (8)$$

Part (3): Let $F(i,k) = A_{k+1}^{(i)} \cdot ((1-p_k(S_{\alpha_k^{(i)}}))\cdot B_{k+1}^{(i)} + C_k^{(i)})$.

$$\sum_{r=1}^{s} r \sum_{\substack{z\subseteq\Sigma \\ |z|=r}} \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}\neq\beta_k^{(i)} \\ \alpha_k^{(i)}\in z \\ \beta_k^{(i)}\notin z}} F(i,k) \quad = \quad \sum_{r=1}^{s} r \sum_{\sigma,\tau\in\Sigma} \sum_{\substack{z\subseteq\Sigma \\ |z|=r}} \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}\neq\beta_k^{(i)} \\ \alpha_k^{(i)}=\sigma\in z \\ \beta_k^{(i)}\notin z}} F(i,k) = \sum_{r=1}^{s} r \binom{s-2}{r-1} \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}\neq\beta_k^{(i)}}} F(i,k)$$

$$= \quad s\cdot 2^{s-3} \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}\neq\beta_k^{(i)}}} F(i,k) \qquad (9)$$

Part (4): Let $G(i,k) = A_k^{(i)} \cdot (B_k^{(i)} - C_k^{(i)})$.

$$\sum_{r=1}^{s} r \sum_{\substack{z\subseteq\Sigma \\ |z|=r}} \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}\notin z}} G(i,k) \quad = \quad \sum_{r=1}^{s} r \sum_{\sigma\in\Sigma} \sum_{\substack{z\subseteq\Sigma \\ |z|=r}} \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}=\sigma\notin z}} G(i,k) = \sum_{r=1}^{s} r \binom{s-1}{r} \sum_{i\in I_1} G(i,k)$$

$$= \quad (s-1)\cdot 2^{s-2} \sum_{i\in I_1} G(i,k) \qquad (10)$$

(5) and (6) are similar to Part (1) and (4), respectively.
Part (5): Let $H(i,k) = (1-p_k(S_{\alpha_k^{(i)}}))\cdot A_{k+1}^{(i)} \cdot B_{k+1}^{(i)}$.

$$\sum_{r=1}^{s} r \sum_{\substack{z\subseteq\Sigma \\ |z|=r}} \sum_{\substack{i\in I_2 \\ \alpha_k^{(i)}\in z}} H(i,k) \quad = \quad \sum_{r=1}^{s} r \binom{s-1}{r-1} \sum_{i\in I_2} H(i,k) = (s+1)\cdot 2^{s-2} \sum_{i\in I_2} H(i,k) \qquad (11)$$

Part (6): Let $I(i,k) = A_k^{(i)} \cdot B_k^{(i)}$.

$$\sum_{r=1}^{s} r \sum_{\substack{z\subseteq\Sigma \\ |z|=r}} \sum_{\substack{i\in I_2 \\ \alpha_k^{(i)}\notin z}} H(i,k) \quad = \quad \sum_{r=1}^{s} r \binom{s-1}{r} \sum_{i\in I_2} H(i,k) = (s-1)\cdot 2^{s-2} \sum_{i\in I_2} H(i,k) \qquad (12)$$

By definition, if $\alpha_k^{(i)} = \beta_k^{(i)}$, then $A_k^{(i)} = p_k(S_{\alpha_k^{(i)}})\cdot A_{k+1}^{(i)}$, $B_k^{(i)} = B_{k+1}^{(i)}$, and $C_k^{(i)} = C_{k+1}^{(i)}$. If $\alpha_k^{(i)} \neq \beta_k^{(i)}$, then $A_k^{(i)} = A_{k+1}^{(i)}$, $B_k^{(i)} = p_k(S_{\alpha_k^{(i)}})\cdot B_{k+1}^{(i)}$, and $C_k^{(i)} = p_k(S_{\alpha_k^{(i)}} \cap S_{\beta_k^{(i)}})\cdot C_{k+1}^{(i)}$. Then, by using these relations, we can show the following equation:

$$\begin{aligned}
\sum_{z\subseteq\Sigma} |z|\cdot \mathrm{gain}(S,k,z) \quad &= \quad (7)+(8)+(9)-(10)+(11)-(12) \\
&= \quad \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}=\beta_k^{(i)}}} ((s+1)\cdot 2^{s-2} - s\cdot 2^{s-1}\cdot p_k(S_{\alpha_k^{(i)}}))\cdot A_{k+1}^{(i)}\cdot(B_{k+1}^{(i)} - C_{k+1}^{(i)}) \\
&+ \quad \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}\neq\beta_k^{(i)}}} ((s+1)\cdot 2^{s-2} - s\cdot 2^{s-1}\cdot p_k(S_{\alpha_k^{(i)}}))\cdot A_{k+1}^{(i)}\cdot B_{k+1}^{(i)} \\
&+ \quad \sum_{\substack{i\in I_1 \\ \alpha_k^{(i)}\neq\beta_k^{(i)}}} (s\cdot 2^{s-1}\cdot p_k(S_{\alpha_k^{(i)}} \cap S_{\beta_k^{(i)}}) - (s+2)\cdot 2^{s-3})\cdot A_{k+1}^{(i)}\cdot C_{k+1}^{(i)} \\
&+ \quad \sum_{i\in I_2} ((s+1)\cdot 2^{s-2} - s\cdot 2^{s-1}\cdot p_k(S_{\alpha_k^{(i)}}))\cdot A_{k+1}^{(i)}\cdot B_{k+1}^{(i)}.
\end{aligned}$$

Therefore, if p_k satisfies $p_k(S_\sigma) \leq \frac{s+1}{2s}$ for all σ in Σ and $p_k(S_\sigma \cap S_\tau) \geq \frac{s+2}{4s}$ for all σ, τ in Σ, then $\sum_{z\subseteq\Sigma} |z| \cdot \mathrm{gain}(S,k,z) \geq 0$. □

Theorem 2 has an advantage that it allows variations for motifs by specifying the probabilities for z in Ω_1 as long as they satisfy the two conditions in Theorem 2. As a corollary of Theorem 2, we can prove the following lower bounds of the expectation:

Corollary 1 *Let $m = |S|$, $s = |\Sigma|$ and let n be the length of a motif.*

(1) If $p_k(Z) = \frac{|Z|}{s \cdot 2^{s-1}}$ for all Z in Ω_1 and for all $1 \leq k \leq n$, then

$$E(S, \pi_1 \cdots \pi_n) \geq \frac{m}{4} \cdot \left(\frac{s+1}{2s}\right)^{n-1}.$$

This is the case that any $z \in \Omega_1$ is allowed for a motif.

(2) If $p_k(\{\sigma\}) = \frac{1}{2s}$ for all σ in Σ, $p_k(\Sigma) = \frac{1}{2}$ and $p_k(Z) = 0$ for other z in Ω_1 for all $1 \leq k \leq n$, then

$$E(S, \pi_1 \cdots \pi_n) \geq \frac{m}{2s} \cdot \left(\frac{s+1}{2s}\right)^{n-1}.$$

This is the case that only Σ and $\{\sigma\}$ for $\sigma \in \Sigma$ are allowed for a motif.

From Theorem 2, the greedy algorithm in Fig. 3 produces a motif $\hat{\pi} = z_1 \cdots z_n$ with $\mathrm{cost}(S, \hat{\pi})$ at least as large as $E(S, \pi_1 \cdots \pi_n)$. The lower bounds of $E(S, \pi_1 \cdots \pi_n)$ in Corollary 1 are not good when n and s are larger. In [13], we have implemented this greedy algorithm and applied to finding motifs for exon/intron splicing sites and *E. coli* promoters. The experimental results in [13] reveal that this algorithm can achieve much better performance in practice.

It is yet another method to apply the approximation algorithm for the maximum generalized satisfiability problem (MAXGSAT) (see Theorem 13. 2 in [8]) by reformulating the problem as MAXGSAT. Given a set $\Psi = \{\psi_1, \ldots, \psi_m\}$ of boolean expressions in n variables, this approximation algorithm produces a truth assignment satisfying at least $\frac{m}{2^k}$ expressions in Ψ, where k is the maximum number of distinct variables in ψ_i for $1 \leq i \leq m$. For this purpose, we transform a yes-no sample $S = \{(\alpha^{(1)}, \beta^{(1)}), \ldots, (\alpha^{(m)}, \beta^{(m)})\}$ to a collection $\Psi = \{\psi_1, \ldots, \psi_m\}$ of boolean expressions as follows: Let $s = |\Sigma|$. We use boolean variables $\pi_k[z]$ for $1 \leq k \leq n$ and $z \in \Omega_1$ in a way that $\pi_k[z] = true \Leftrightarrow \pi_k = z$. Let

$$\varphi = \bigwedge_{k=1}^{n}((\bigvee_{z \in \Omega_1} \pi_k[z]) \wedge (\bigwedge_{z \neq z' \in \Omega_1} (\overline{\pi_k}[z] + \overline{\pi_k}[z']))).$$

Obviously, a truth assignment satisfying μ determines a motif $\pi = z_1 \cdots z_n$. Then for $(\alpha^{(i)}, \beta^{(i)})$, let ψ_i be a boolean expression with these $n \cdot (2^s - 1)$ variables defined by

$$\psi_i = \varphi \wedge ((\bigwedge_{k=1}^{n} \bigvee_{\alpha_k^{(i)} \in z \in \Omega_1} \pi_k[z]) \wedge (\bigvee_{k=1}^{n} \bigvee_{\beta_k^{(i)} \in z \in \Omega_1} \overline{\pi_k}[z]))$$

Let $\pi = z_1 \cdots z_n$ be a motif determined by a truth assignment. Then ψ_i is satisfied if and only if $(\alpha^{(i)}, \beta^{(i)})$ is good for π for each $1 \leq i \leq m$. By applying the approximation algorithm for MAXGSAT to the instance $\Psi = \{\psi_1, \ldots, \psi_m\}$, we obtain a motif π with

$$\mathrm{cost}(S, \pi) \geq \frac{m}{2^{n \cdot (2^s - 1)}}.$$

This lower bound becomes drastically worse if the size s of Σ gets larger.

5 Conclusion

First, we defined the best consensus motif problem (BCM) motivated by the problem of extracting motifs from positive and negative sequences. Second, we proved the NP-completeness of the BCM problem for type $\Omega_1 = \{z \mid \emptyset \neq z \subseteq \Sigma\}$. Furthermore, for this problem, we devised a polynomial-time greedy algorithm and showed a lower bound of its performance. This problem has a special importance when data are provided as tuples of values of attributes $A_1, \ldots, A_n$

as in [9]. We do not know whether there is any polynomial time algorithm with better performance. We also showed that the BCM problem for type $\Omega_\infty = \Omega_1 \cup \{\Sigma^+\} \cup \{\Sigma^{[i,j]} \mid 1 \leq i \leq j\}$ is NP-complete. From both theory and practice, it is an challenging open problem to devise any efficient approximation algorithm for finding such general motifs together with a good performance guarantee.

Aiming at knowledge discovery from amino acid sequences, the third author's research group has developed a system called BONSAI Garden [2, 10, 12] that produces, from positive and negative examples, a mapping ψ called an alphabet indexing which classifies twenty amino acid residues into a smaller categories and a decision tree whose internal nodes are labeled with regular patterns. Since regular patterns are used for making decisions at nodes, only exact pattern matching is allowed. We are planing to implement the algorithm developed in this paper in a forthcoming version of BONSAI Garden so that it can cope with sequences with ambiguity.

Acknowledgments

The authors would like to thank Satoru Kuhara for his guidance in Genome Informatics. This work is partly supported by Grant-in-Aid for Scientific Research on Priority Area "Genome Informatics", No. 06558047 and No. 06680326 from Ministry of Education, Science and Culture, Japan.

References

[1] Angluin, D., Finding patterns common to a set of strings, *J. Comput. System Sci.* **21** (1980) 46–62.

[2] Arikawa, S., Miyano, S., Shinohara, A., Kuhara, S., Mukouchi, Y., and Shinohara, T., A machine discovery from amino acid sequences by decision trees over regular patterns, *New Generation Computing* **11** (1993) 361–375.

[3] Bairoch, A., PROSITE: a dictionary of sites and patterns in proteins, *Nucleic Acids Res.* **19** (1991) 2241–2245.

[4] Garey, M.R., Johnson, D.S. and Stockmeyer, L., Some simplified NP-complete problems, *Theoret. Comput. Sci.* **1** (1976) 237-267.

[5] Gribskov, M. and Devereux, J., *Sequence Analysis Primer*, Stockholm Press, 1991.

[6] Jiang, T. and Li, M., On the complexity of learning strings and sequences, *Proc. 4th Workshop on Computational Learning Theory*, 1991, 367–371.

[7] Miyano, S., Shinohara, A. and Shinohara, T., Which classes of elementary formal systems are polynomial-time learnable?, *Proc. Second Workshop on Algorithmic Learning Theory*, 1991, 139–150.

[8] Papadimitriou, C.H., *Computational Complexity*, Addison-Wesley, 1994.

[9] Quinlan, J.R., Induction on decision trees, *Machine Learning* **1** (1986) 81–106.

[10] Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., and Arikawa, S., Knowledge acquisition from amino acid sequences by machine learning system BONSAI, *Transactions of Information Processing Society of Japan* **35** (1994) 2009–2018.

[11] Shinohara, T., Polynomial time inference of extended regular pattern languages, *Lecture Notes in Computer Science* **147** (1983) 115–127.

[12] Shoudai, T., Lappe, M., Miyano, S., Shinohara, A., Okazaki, T., Arikawa, S., Uchida, T., Shimozono, S., Shinohara, T., and Kuhara, S., BONSAI Garden: parallel knowledge discovery system for amino acid sequences, *Proc. Third International Conference on Intelligent Systems for Molecular Biology* (AAAI Press), 1995, 359–366.

[13] Tateishi, E. and Miyano, S., A greedy strategy for finding motifs from positive and negative examples, RIFIS-TR-CS-118, Research Institute of Fundamental Information Science, Kyushu University, 1995.

[14] Yannakakis, M., On the approximation of maximum satisfiability, *J. Algorithms* **17** (1994) 475–502.