DNA Motif Representation with Nucleotide Dependency*

Francis Chin[1] and Henry Leung[1]
[1] Department of Computer Science,
The University of Hong Kong, Pokfulam, Hong Kong
{chin, cmleung2}@cs.hku.hk

Abstract. The problem of discovering novel motifs of binding sites is important to the understanding of gene regulatory networks. Motifs are generally represented by matrices (PWM or PSSM) or strings. However, these representations cannot model biological binding sites well because they fail to capture nucleotide interdependence. It has been pointed out by many researchers that the nucleotides of the DNA binding site cannot be treated independently, e.g. the binding sites of zinc finger in proteins. In this paper, a new representation called Scored Position Specific Pattern (SPSP), which is a generalization of the matrix and string representations, is introduced which takes into consideration the dependent occurrences of neighboring nucleotides. Even though the problem of discovering the optimal motif in SPSP representation is proved to be NP-hard, we introduce a heuristic algorithm called SPSP-Finder, which can effectively find optimal motifs in most simulated cases and some real cases for which existing popular motif-finding software, such as Weeder, MEME and AlignACE, fail.

Index Term: I. Computing Methodologies; 5. Pattern Recognition; 2. Design Methodology; c. Pattern analysis

1 Introduction

A *gene* is a segment of DNA that is the blueprint for protein. In most cases, genes seldom work alone; rather, they cooperate to produce different proteins for a particular function. In order to start the protein decoding process (*gene expression*), a molecule called *transcription factor* will bind to a short region (*binding site*) preceding the gene. One kind of transcription factor can bind to the binding sites of several genes to cause these genes to co-express. These binding sites have similar patterns called *motifs*. Discovering novel motifs of unknown transcription factors and the binding sites from a set of DNA sequences is a critical step for understanding the *gene regulatory network*.

In order to discover motifs of unknown transcription factors, we must first have a model to represent motifs. There are two popular models: string representation [4,6-8,13,14,17,20,22,23,25-30,31-33] and matrix representation [1,2,9,11,15,16,18,19,21]. String representation is the most basic representation which uses a length-l string of symbols (or nucleotides) 'A', 'C', 'G' and 'T' to describe a motif. To improve the representation's descriptive power, wildcard symbols [6,26,31] can be introduced into the string to represent choice from a subset of symbols at a particular position (e.g. 'K' can denote 'G' or 'T'). Matrix representation further improves descriptive power. In the matrix representation, motifs of length l are represented by *position weight matrices* (PWMs) or *position specific scoring matrices* (PSSMs) of size $4 \times l$ with the four entries

* The research was supported in parts by the RGC grant HKU 7120/06E

in the *j*-th column of the matrix, effectively giving the occurrence probabilities of the four nucleotides at position *j*. While matrix representation appears superior, the solution space for PWMs and PSSMs, which consists of 4*l* real numbers is infinite in size and there are many local optimal matrices, thus, algorithms generally either produce a sub-optimal motif matrix [1,2,9,15,16,21] or take too long to run when the motif is longer than 10 bp [19].

As it turns out, the string and the matrix representations share an important common weakness: they assume the occurrence of each nucleotide at a particular position of a binding site is independent of the occurrence of nucleotides at other positions. This assumption does not represent the true picture. According to Bulyk et al [5], analysis of wild-type and mutant Zif268 (Egr1) zinc fingers gives compelling evidence that nucleotides of transcription factor binding sites should not be treated independently, and a more realistic motif representation should be able to describe nucleotide interdependence. Man and Stormo [24] have arrived at a similar conclusion in their analysis of *Salmonella* bacteriophage repressor Mnt: they found that interactions of Mnt with nucleotides at positions 16 and 17 of the 21 bp binding site are in fact not independent.

When the positions of binding sites are known, we may represent the motif by hidden Markov model (HMM) [36], Bayesian network [3] or enhanced PWM [10] which can overcome the above weakness. However, these models cannot be easily extended to discover novel motifs especially when the number of co-expressed genes is small (say less than 10). It is because the input data does not contain enough information for deriving the hidden motif and the above models usually overfit the input data. Hence, they are far less popular representations.

In this paper, we introduce a new motif representation called *Scored Position Specific Pattern* (SPSP) which has the following advantages:

(a) *Better representation*. SPSP can describe the interdependence between neighboring nucleotides with similar number of parameters as string and matrix representations.

(b) *Generalization of string and matrix representations.* These two commonly-used representations are special cases of the SPSP representation. Thus SPSP representation can model more motifs than these two representations.

(c) *Computationally feasibility*. Finding the optimal motif in SPSP representation, for some restricted cases, is more feasible than finding the optimal PWM or PSSM.

This paper tackles a "restricted" motif discovering problem based on the SPSP representation. Although this is a restricted problem, it can model all motifs in string representation and most motifs in matrix representation. Because this restricted problem is NP-complete (proof shown in the Appendix), we introduce a heuristic algorithm called *SPSP-Finder* which can find the optimal SPSP motifs in most simulated cases and some real cases, for which Weeder [25], MEME [16] and AlignACE [12] fail.

This paper is organized as follows. In Section 2, we describe the SPSP representation, the corresponding motif problem and its restricted version in detail. In Section 3, we introduce the heuristic algorithm SPSP-Finder. Experimental results on simulated data and real biological data comparing SPSP-Finder with some popular software are given in Section 4, followed by concluding remarks in Section 5.

2 Scored Position Specific Pattern (SPSP)

Consider the wildcard-augmented string representation with 15 symbols representing all combinations of the four nucleotides ‘A’, ‘C’, ‘G’ and ‘T’. For example, the wildcard symbol ‘Y’ represents ‘C’ or ‘T’ and wildcard symbol ‘S’ represents ‘C’ or ‘G’. Consider the motif for the transcription factor HAP2 [34] which exists as a heterotrimeric complex with the HAP3 and HAP4 proteins. The HAP2/3/4 complex binds to the patterns “CCAATCA”, “CCAATGA” or “CCAACCA”. We can represent the motif by “CCAAYSA” with two wildcard symbols. In fact, we may also represent “CCAAYSA” as follows:

$$(\mathrm{C})(\mathrm{C})(\mathrm{A})(\mathrm{A})\begin{pmatrix}\mathrm{C}\\ \mathrm{T}\end{pmatrix}\begin{pmatrix}\mathrm{C}\\ \mathrm{G}\end{pmatrix}(\mathrm{A})$$

However, this representation has the problem that the pattern “CCAACGA” is also considered as a binding site (false positive). In order to prevent the inclusion of false positive patterns, we replace the substring “YS” by a set of length-2 patterns: i.e.,

$$(\mathrm{C})(\mathrm{C})(\mathrm{A})(\mathrm{A})\begin{pmatrix}\mathrm{TC}\\ \mathrm{TG}\\ \mathrm{CC}\end{pmatrix}(\mathrm{A}) \text{ or } (\mathrm{CCAA})\begin{pmatrix}\mathrm{TC}\\ \mathrm{TG}\\ \mathrm{CC}\end{pmatrix}(\mathrm{A})$$

The Scored Position Specific Pattern (SPSP) representation uses such an idea to represent motifs. Based on this SPSP representation, our algorithm can find the motif and binding sites of HAP2 while the other software fails to do so. The formal definition of SPSP is described in the following section.

2.1 Formal Definition of Pattern Sets Representation

A set of length-l binding site patterns can be described by a *Scored Position Specific Pattern* (SPSP) representation P which contains c ($c \le l$) sets of patterns P_i, $1 \le i \le c$, where each set of patterns P_i contains length-l_i patterns $P_{i,j}$ of symbols ‘A’, ‘C’, ‘G’ and ‘T’, and $\sum_i l_i = l$. Each length-l_i pattern $P_{i,j}$ is associated with a score $s_{i,j}$ which represents the “closeness” of a pattern to be a binding site, i.e. the lower the score, the more likely that the pattern is a binding site. The score of a length-l string $\sigma = \sigma_1\sigma_2\ldots\sigma_c$ where $|\sigma_i| = l_i$, $1 \le i \le c$ with respect to P can be defined as follows:

$$\mathrm{score}(\sigma, P) = \sum_{i=0}^{c} \begin{cases} s_{i,j} & \exists j, P_{i,j} = \sigma_i \\ \infty & \text{otherwise} \end{cases}$$

A string σ is a *binding site* with respect to an SPSP motif P if and only if score(σ,P) is no more than some predefined threshold α.

For example, consider the following SPSP representation for the length-11 binding sites of the transcription factor CSRE [37] which activates the gluconeogenic structural genes.

$$P = (\mathrm{CGGA})\begin{pmatrix}\mathrm{TGA}\\ \mathrm{TAA}\\ \mathrm{CGG}\end{pmatrix}(\mathrm{A})\begin{pmatrix}\mathrm{A}\\ \mathrm{T}\end{pmatrix}(\mathrm{GG}) \text{ and}$$

$$\{s_{i,j}\} = (-\log(1))\begin{pmatrix}-\log(0.5)\\ -\log(0.3)\\ -\log(0.2)\end{pmatrix}(-\log(1))\begin{pmatrix}-\log(0.7)\\ -\log(0.3)\end{pmatrix}(-\log(1))$$

Note that the score $s_{i,j}$ is the negative of the logarithm of the occurrence probability of the corresponding pattern $P_{i,j}$. The score of the length-11 string σ = "CGGATAAAAGG" with σ_1 = "CGGA", σ_2 = "TAA", σ_3 = "A", σ_4 = "A" and σ_5 = "GG" can be calculated as -log(1) – log(0.3) – log(1) – log(0.7) – log(1) = -log(0.21). On the other hand, the score of σ = "CTGATAAAAGG" is ∞ as σ_1 = "CTGA" $\notin P_1$. The scores of these strings represent the negative log likelihood of these strings being binding sites of P. A string with smaller score is more likely to be a binding site of P.

Based on the SPSP representation, we can define the Motif Discovering Problem as follows:

> **Motif Discovering (MD) Problem:** Given t length-n DNA sequences T, we want to find a motif M in SPSP representation (P and score $\{s_{i,j}\}$ satisfying certain properties) to maximize/minimize some target function calculated based on the scores of the binding sites of M in T.

The following will show that SPSP representation is a generalization of the string and matrix representations. By applying different target functions, we can discover motifs with different properties under a certain score scheme $\{s_{i,j}\}$.

(a) Restricting $c = l$ (that means $l_i = 1$, $1 \le i \le c = l$), the SPSP representation P is equivalent to a position weight matrix (PWM) or position specific scoring matrix (PSSM) [1, 15, 16, 19]. Using the following probability matrix for transcription factor CSRE with threshold 0.04 as an example.

$$\begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array}\begin{pmatrix} 0 & 0 & 0 & 1.0 & 0 & 0.3 & 0.7 & 1.0 & 0.7 & 0 & 0 \\ 1.0 & 0 & 0 & 0 & 0.3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.0 & 1.0 & 0 & 0 & 0.7 & 0.3 & 0 & 0 & 1.0 & 1.0 \\ 0 & 0 & 0 & 0 & 0.7 & 0 & 0 & 0 & 0.3 & 0 & 0 \end{pmatrix}$$

It is equivalent to the following SPSP representation:

$$P = (\text{C})(\text{G})(\text{G})(\text{A})\begin{pmatrix}\text{C}\\\text{T}\end{pmatrix}\begin{pmatrix}\text{A}\\\text{G}\end{pmatrix}\begin{pmatrix}\text{A}\\\text{G}\end{pmatrix}(\text{A})\begin{pmatrix}\text{A}\\\text{T}\end{pmatrix}(\text{G})(\text{G}) \text{ and}$$

$$\{s_{i,j}\} = (0)(0)(0)(0)\begin{pmatrix}-\log(0.3)\\-\log(0.7)\end{pmatrix}\begin{pmatrix}-\log(0.3)\\-\log(0.7)\end{pmatrix}\begin{pmatrix}-\log(0.7)\\-\log(0.3)\end{pmatrix}(0)\begin{pmatrix}-\log(0.7)\\-\log(0.3)\end{pmatrix}(0)(0)$$

with threshold α = -log(0.04). Note that –log(1.0) = 0.

In order to find a set of binding sites with the minimum negative log likelihood, the MD problem is to find P and $\{s_{i,j}\}$ such that for $1 \le i \le c = l$, $s_{i,j} = -\log(p_{i,j})$ with $\sum_j p_{i,j} = 1$ so as to minimize the target function $\sum_\sigma[\text{score}(\sigma, P) + l\log(0.25)]$ for all binding site σ (i.e. with score(σ, P) $\le \alpha$ (threshold)).

(b) Restricting $c = l$, $s_{i,j} = 0$ or 1, $\sum_j s_{i,j} = 3$ and $\alpha = d$, the SPSP representation P is equivalent to a string representation [4,8,22,23,27] for the planted (l,d)-motif problem. For example, the HAP2 motif "CCAATTA" for the planted (7,d)-motif problem is equivalent to the following SPSP representation:

$$P = \begin{pmatrix}\text{A}\\\text{C}\\\text{G}\\\text{T}\end{pmatrix}\begin{pmatrix}\text{A}\\\text{C}\\\text{G}\\\text{T}\end{pmatrix}\begin{pmatrix}\text{A}\\\text{C}\\\text{G}\\\text{T}\end{pmatrix}\begin{pmatrix}\text{A}\\\text{C}\\\text{G}\\\text{T}\end{pmatrix}\begin{pmatrix}\text{A}\\\text{C}\\\text{G}\\\text{T}\end{pmatrix}\begin{pmatrix}\text{A}\\\text{C}\\\text{G}\\\text{T}\end{pmatrix}\begin{pmatrix}\text{A}\\\text{C}\\\text{G}\\\text{T}\end{pmatrix} \text{ and } \{s_{i,j}\} = \begin{pmatrix}1\\0\\1\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\\1\end{pmatrix}\begin{pmatrix}0\\1\\1\\1\end{pmatrix}\begin{pmatrix}0\\1\\1\\1\end{pmatrix}\begin{pmatrix}1\\1\\1\\0\end{pmatrix}\begin{pmatrix}1\\1\\1\\0\end{pmatrix}\begin{pmatrix}0\\1\\1\\1\end{pmatrix}$$

with threshold $\alpha = d$.

In order to find the maximum number of binding sites with at most d substitutions from a string motif, the MD problem is to find $\{s_{i,j}\}$ such that for $1 \le i \le c = l$, $s_{i,j} = 0$ for a particular j and = 1 for all other j, so as to maximize the number of binding sites as its target function. Note that the SPSP representation P is already fixed as shown above.

(c) Restricting $c = l$, $s_{i,j} = 0$ and $\alpha = 0$, the SPSP representation P is equivalent to a length-l string with wildcard symbols [20,31]. For example, the BAS2 [37] motif "TAATRA" in string representation with wildcard symbols is equivalent to the following SPSP representation

$$P = (\mathrm{T})(\mathrm{A})(\mathrm{A})(\mathrm{T})\begin{pmatrix}\mathrm{A}\\ \mathrm{G}\end{pmatrix}(\mathrm{A}) \text{ and } \{s_{i,j}\} = (0)(0)(0)(0)\begin{pmatrix}0\\ 0\end{pmatrix}(0)$$

with threshold $\alpha = 0$.

In order to find a set of binding sites with a minimum z-score [31] or p-value [20], the MD problem is to find the SPSP representation P such that for all i, j, $s_{i,j} = 0$, so as to minimize the z-score or p-value of the binding sites as its target function. Note that the z-score or p-value decreases with the inverse of the number of binding sites and the number of conserved symbols.

2.2 Restricted Motif Discovering Problem

In the real biological situation, transcription factors bind to binding sites by some components called DNA-binding domains (e.g. zinc finger). Each domain of the transcription factor usually binds to 3-4 bp consecutive regions of the binding sites [24, 35]. Therefore, we may assume the length l_i of each pattern $P_{i,j}$ is not larger than 4. Besides, the background occurrence probability of each length-l pattern in the input sequence is not the same. This uneven probability can be estimated by an order 0 to 3 Hidden Markov Model (HMM) [36].

Instead of solving the general Motif Discovering Problem described in Section 2.1, this paper tackles a "restricted" version of the motif problem based on the assumption that l_i is small, i.e. $l_i \le l_{\max}$ for a predefined value $l_{\max}$. Besides, the overall binding site patterns should be similar, i.e. the score $s_{i,j}$ of each length-l_i pattern $P_{i,j}$ must be equal to its Hamming distance with some representative length-l_i string R_i.

$$P_i = \begin{pmatrix}\mathrm{ACG}\\ \mathrm{ACT}\\ \mathrm{AGT}\\ \mathrm{CCG}\end{pmatrix} \text{ and } S_{ij} = \begin{pmatrix}1\\ 0\\ 1\\ 2\end{pmatrix}$$

if R_i = "ACT". Similar if R_i = "ACG",

$$S_{ij} = \begin{pmatrix}0\\ 1\\ 2\\ 1\end{pmatrix}$$

A length-l string σ is a binding site of M if and only if score(σ,P) $\le d$, i.e. σ should be within Hamming distance d from a particular motif pattern.

Intuitively the *Restricted Motif Discovering* (RMD) Problem is finding an SPSP representation P such that the number of possible string patterns for binding sites $\prod_i |P_i| = w$ is minimized and at the same time P can cover the maximum number of binding sites b.

For example, assume the occurrence probability of each length-l pattern is the same. Given the following binding sites $\{s_i\}$ and motif P_1 and P_2:

s_1	G	T	A	T	T	A	A
s_2	G	T	A	T	T	A	G
s_3	G	T	A	T	A	A	C
s_4	G	T	A	T	A	A	G
s_5	G	T	A	T	G	A	G
s_6	G	T	A	T	C	A	G
s_7	C	T	A	T	G	A	C
s_8	C	T	A	T	C	A	G
s_9	C	T	A	T	A	A	G
s_{10}	C	T	A	T	G	A	C

$$P_1 = (\text{GTAT})\begin{pmatrix}\text{A}\\\text{C}\\\text{G}\\\text{T}\end{pmatrix}(\text{A})\begin{pmatrix}\text{A}\\\text{C}\\\text{G}\end{pmatrix} \quad P_2 = \begin{pmatrix}\text{G}\\\text{C}\end{pmatrix}(\text{TAT})\begin{pmatrix}\text{A}\\\text{C}\\\text{G}\end{pmatrix}(\text{A})\begin{pmatrix}\text{C}\\\text{G}\end{pmatrix}$$

$$\{s_{i,j}\}_1 = (0)\begin{pmatrix}0\\1\\1\\1\end{pmatrix}(0)\begin{pmatrix}1\\0\\1\end{pmatrix} \qquad \{s_{i,j}\}_2 = \begin{pmatrix}0\\1\end{pmatrix}(0)\begin{pmatrix}0\\1\\1\end{pmatrix}(0)\begin{pmatrix}0\\1\end{pmatrix}$$

Score $\{s_{i,j}\}$ are defined such that score(σ,P) = Hamming distance between σ and "GTATAAC". Since the number of possible patterns of binding sites for P_1 and P_2 are the same, i.e. $w = 4 \times 3$ and $2 \times 3 \times 2$ respectively, P_2 is more likely to be a correct motif than P_1 as P_2 covers more binding sites (s_3 to s_{10}) than P_1 (s_1 to s_6).

Usually, it might not be so obvious which motif is more likely to be correct, e.g. when $w_1 < w_2$ and $b_1 < b_2$. In such case, we compare two motifs by the occurrence probabilities (p-values) of their corresponding binding sites in T with the assumption that T is a set of random sequences generated according to a Markov model. Given a motif with $\prod_i |P_i| = w$ (the w possible binding sites patterns are $\{B_k, k = 1, \ldots, w\}$) and having b binding sites in T, the occurrence probability of $\geq b$ binding sites in a set of random sequences can be calculated as

$$p - \text{value} = 1 - \sum_{i=0}^{b-1} \binom{t(n-l+1)}{i} \left(\sum_{k=1}^{w} P(B_k)\right)^i \left(1 - \sum_{k=1}^{w} P(B_k)\right)^{t(n-l+1)-i}$$

where $P(B_k)$ is the probability that B_k occurs at a particular position given that the sequence is generated according to the Markov model. A motif with low p-value means that it is likely to be an answer. Note that p-value increases as w and $P(B_k)$, the number and the occurrence probabilities of the possible binding site patterns B_k, increases and decreases as b, the number of binding sites, increases. Thus, we define the Restricted Motif Discovering Problem formally as follow.

> **Restricted Motif Discovering (RMD) Problem:** Given the Markov model for the background sequences, t length-n DNA sequences T, the threshold value d and $l_{\max}$, we want to find a length-l motif P and a set of score $\{s_{i,j}\}$ such that $s_{i,j}$ equals to the Hamming distance between $P_{i,j}$ and some representative length-l_i string R_i and having the minimum p-value of the corresponding binding sites.

Although we have imposed restriction to the MD problem, this restricted SPSP representation is still more descriptive than the string representation in the sense that all

string representations are special cases of this restricted representation. This restricted representation takes into account the dependence of the occurrence of the nucleotides in a binding site and some nucleotides in binding sites are conserved. Under the RMD problem, all possible binding sites have equal occurrence probability given that they are generated according to the motif. Note that it is not the same as the occurrence probabilities of the binding sites generated according to the Markov model (background). Thus, we cannot determine whether this restricted representation is more descriptive than the matrix representation. Besides, as we assume the transcription factor binds to the binding sites by DNA-binding domains at short consecutive regions, the RMD problem does not model binding sites with dependency over a long region, e.g. GAL4, functioned as a homodimer with binding pattern $CGGN_{11}CCG$ (the prefix CGG is a reverse complement of the suffix CCG), dependency over a long region.

We shall show in the next section that there is an efficient heuristic to solve the RMD problem with which we can successfully find motifs in some cases for which popular motif-finding software fail.

3 Algorithm SPSP-Finder

In this section, we describe a heuristic algorithm, *SPSP-Finder*, to solve the Restricted Motif Discovering Problem. This algorithm starts with a set of "good" string patterns and, based on local search, finds some local optimal SPSP representations and their corresponding binding sites. This algorithm has two main steps. The first step, served as seed searching, is to find a set of length-l string motifs with many binding sites in the input sequences. In the second step, we start with each length-l string R as a seed SPSP representation and merge some positions of R's binding sites to form another SPSP representation with smaller p-value. This merging step is repeated until the p-value cannot be further reduced. Definitely, this algorithm cannot guarantee the finding optimal motif in SPSP representation. However, when more seed sequences are considered, the longer is the running time, the better will be the solution.

3.1 Seed Searching

Voting Algorithm [9,17] is applied to discover length-l string motifs. Voting is used for solving the planted (l,d)-motif problem where a motif is represented by a length-l string S and the binding sites are d-variants of S (d-*variant* of S is a length-l string derivable from S with at most d symbol substitutions). This algorithm is based on the idea that if each length-l substring in the input sequences T gives a vote to each of its d-variants, a string S with b d-variants in T will get exactly b votes. Finding the number of d-variants in T of each length-l string S takes $O(nt(3l)^d)$ time [9,17].

Since the occurrence probability of each length-l string in T is different, we modify the Voting Algorithm such that each string σ gives $1/P(\sigma)$ vote to each of its d-variants where $P(\sigma)$ is σ's occurrence probability based on the background modeled by HMM. A string σ with low occurrence probability in the background will contribute a higher score to its d-variants. As the string S with relative more d-variants of low occurrence probabilities is more likely to be one of the motif patterns, we refine each length-l string one by one (Section 3.2) in decreasing order of the sum of (weighted) votes received.

3.2 Refining the SPSP Representation

Given representative string S (motif candidate), we can find all length-l d-variants of S (potential binding sites) in the t length-n DNA sequences T. By aligning these length-l d-variants, we can construct a restricted SPSP representation P for these d-variants by considering the consensus substrings as the representative strings R_i. However, as some of these d-variants might not be binding sites, the value of $\Sigma_k P(k)$ as well as the p-value may be very large. In order to reduce the value of $\Sigma_k P(k)$, we shall construct restricted SPSP representation for subsets of the d-variants. Since finding the optimal subset of d-variants (subset with the lowest p-value) is NP-complete (see Appendix), heuristic approach is being considered. We begin with the set of all d-variants. At each iteration, we remove the d-variant whose removal decreases the p-value most. Note that the restricted SPSP representation will change if the set of d-varinats is different. If we find a motif candidate M with smaller p-value than the best motif M* found so far, we update M* by M. We repeat this step until the p-value of new motif candidate M cannot be lowered.

After considering (or refining) one string, we shall consider (or refine) the next candidate string having the largest weighted votes. When the number of d-variants of the remaining candidates (as the candidates have been sorted in decreasing order of weighted votes) is too small to be refined to a better motif than M*, we stop the process and report M* as the answer.

4 Experimental Results

Based on the ideas in Section 3, we have implemented SPSP-Finder in C++. SPSP-Finder was used to find motifs in both simulated and real biological data. All experiments were performed on a 2.4GHz P4 CPU with 1 GB memory. The performance of SPSP-Finder was compared with various existing motif-finding algorithms.

Table 1. Experimental results on simulated data – the success rate of each software

l	α	Weeder	MEME	AlignACE	SPSP-Finder
5	0	100%	68%	10%	100%
5	1	6%	16%	4%	42%
7	1	100%	72%	14%	100%
7	2	7%	12%	2%	38%
9	2	100%	68%	18%	100%
9	3	2%	12%	8%	52%
11	3	98%	74%	14%	100%
11	4	3%	14%	0%	32%

4.1 Simulated Data

The simulated data were generated in the following manner. Twenty length-600 sequences were generated with each nucleotide having the same occurrence probability 0.25. Then a length-l motif M in SPSP representation with $l_{max} = 4$ was picked randomly according to the following two steps.

(1) A set of c numbers $l_1 \ldots l_c$ such that $c \leq l$ and $\sum_i l_i = l$, corresponding to the parameters of an SPSP representation, was generated randomly;

(2) For $i = 1, \ldots, c$, an integer r_i was randomly picked from 1 to 4 with equal probability. r_i length-l_i random strings were generated independently with each nucleotide having the same occurrence probability 0.25.

A binding site of M was randomly picked with equal probability and planted at a random position of each sequence in T. The Weeder [25], MEME [16], AlignACE [12] and SPSP-Finder were used to discover this hidden motif M. SPSP-Finder calculated the score of a predicted motif using an order-0 Markov model with the same occurrence probability for each nucleotide. The accuracy for each motif predicted by the above algorithms is defined as:

$$\text{accuracy} = \frac{|\text{predicted sites} \cap \text{planted sites}|}{|\text{predicted sites} \cup \text{planted sites}|}$$

A planted binding site is *correctly predicted* if that binding site overlaps with at least one predicted binding site. An algorithm is said to have predicted the hidden motif correctly if the accuracy ≥ 0.5. For each set of parameters, i.e. length l and threshold (Hamming distance) α, we ran 50 test cases. Table 1 shows the success rate of the algorithms in discovering the motif.

Buhler and Tompa [4] proved that when α is large with respect to l (e.g. the (5,1), (7,2), (9,3) and (11,4) problems), there are many random patterns having the same number of α-variants as the motif, so algorithms are unlikely to be able to discover the motif without extra information. Indeed, the Weeder, MEME and AlignACE do not perform well in these cases. MEME has a better performance than the other two algorithms because it allows different occurrence probabilities for nucleotides at each position. Since SPSP-Finder considers the dependence of the nucleotides, it has better performance than Weeder, MEME and AlignACE.

Table 2. Experimental results on real biological data in SCPD database

Factor Name	Pattern	Weeder	MEME	AlignACE	SPSP-Finder
13nt	ACGAGGCTTACCG	-	-	ACGAGGCTTACCG	(ACGA)(GGCT)(TACC)(G)
ACE2	GCTGGT	-	-	-	(GCTG)(GT)
ADR1	TCTCC	-	TCTCC	TCTCC	(TCTC)(C)
AP1	TTANTAA	-	-	-	(TTA)([C G])(TAA)
BAS2	TAATGA	-	-	-	(TAAA)([A G])(A)
CCBF	CNCGAAA	CACGAAA	-	-	(C)([A C G T])(CGAA)(A)
CPF1	TCACGTG	CACGTG	TCACGTG	-	(CACG)(TG)
CSRE	CGGAYRRAWGG	-	-	-	(CGGA)([TGA TAA CGG])(A)([A T])(GG)
CuRE	TTTGCTC	TTTGCTCA		-	(TTT)([GCT TCG])(C)
GATA	CTTATC	CTTATC	-	-	(CTTA)(TC)
HAP2/3/4	CCAATCA	-	-	-	(CCAA)([TC TG CC])(A)
LEU	CCGNNNNCGG	CCGGGACCGG	CCGGAACCGG	-	(CGG)([A G])(ACCGG)
MATα2	CRTGTWWWW	CATGTAATTA	-	CATGTAATT	([GA GT TA])(AATT)([AC TA TC TG])([A C])
MCM1	CCNNNWWRGG	CCCGTTTAGG	CCTAATTAGG	-	-
SFF	GTMAACAA	-	GTCAACAA	-	-
UASCAR	TTTCCATTAGG	-	-	-	(T)([GCCC TCAC TCCA])(TT)([AGCG AGGA])

Motifs of transcription factors that cannot be found by any algorithms were not shown in this table. 'M' stands for 'A' or 'C', 'N' stands for any nucleotide. 'R' stands for 'A' or 'G', 'W' stands for 'A' or 'T', 'Y' stands for 'C' or 'T'. Those motifs that all four algorithms can/cannot discover are not shown.

4.2 Real Biological Data

SCPD [37] contains information of different transcription factors for yeast. For each set of genes regulated by the same transcription factor, we chose the 600 base pairs in the upstream of these genes as the input sequences *T*. The Weeder, MEME, AlignACE and SPSP-Finder were used to discover the motifs. SPSP-Finder used an order-0 Markov model calculated based on the input sequence when calculating score of each predicted motif. Table 2 showed the experimental results of all transcription factors in SCPD except those motifs which cannot be discovered by any algorithms. As shown in Table 2, SPSP-Finder performs better than other algorithms in most cases. There are six motifs, ACE2, AP1, BAS2, CSRE, HAP2/3/4 and UASCAR and their binding sites could be discovered (accuracy ≥ 0.5) by SPSP-Finder but not by the other algorithms. Refer to the published binding sites in SCPD database, there are nucleotide dependencies in these binding sites.

For example, the HAP2/3/4 complex is a CCAAT-binding complex which mainly binds to the sequence "CCAATCA" in yeast. Although it also binds to the sequences "CCAATGA" and "CCAACCA", it cannot bind to sequences "CCAAACA" and "CCAAAGA" [30]. Since the binding sites are short and there are two non-conserved positions (positions 5 and 6), Weeder failed to discover the published motif because there were many length-7 random patterns whose 2-variants occurred more frequently than the binding sites of "CCAATCA". In this case, Weeder cannot distinguish the published motif "CCAATCA" from these random patterns. Similarly, MEME and AlignACE failed because there were many PSSMs having higher scores than the score of the published motif if the nucleotide dependency in positions 5 and 6 were not considered. By considering the nucleotide dependency in positions 5 and 6, SPSP-Finder discovered the motif in SPSP representation

$$(\mathrm{CCAA})\begin{pmatrix}\mathrm{TC}\\ \mathrm{TG}\\ \mathrm{CC}\end{pmatrix}(\mathrm{A})$$

which had a lower *p*-value than "CCAAYSA" and other random patterns.

CSRE is a transcription factor responsible for the transcriptional activation of gluconeogenic structural genes. There are five binding sites in the data set which can be represented by the motif "CGGAYRRAWGG". This motif contains 4 wildcard symbols and represents 16 different binding sequences instead of 5. Since this motif cannot model the binding sites specifically, many length-11 random patterns had frequently-occurring 4-variants and could be mistaken as the hidden motif. Therefore, Weeder could not discover the motif. Similarly, MEME and AlignACE failed even using the more precise PSSM representation. SPSP-Finder discovered the following motif in SPSP representation

$$(\mathrm{CGGA})\begin{pmatrix}\mathrm{TGA}\\ \mathrm{TAA}\\ \mathrm{CGG}\end{pmatrix}(\mathrm{A})\begin{pmatrix}\mathrm{A}\\ \mathrm{T}\end{pmatrix}(\mathrm{GG})$$

Although this motif in SPSP representation represented 6 instead of 5 binding patterns, it could describe the binding sites better than those motifs in string representation or PSSM.

Therefore, SPSP-Finder could discover the published motif successfully while Weeder, MEME and AlignACE failed.

For those cases that SPSP-Finder and other algorithms could discover the published binding sites, SPSP-Finder had an advantage that it can represent the binding sites better. For example, the CCBF transcription factor can bind to sequences "CNCGAAA" where 'N' represents any nucleotides. Although both Weeder and SPSP-Finder could discover the published motif, Weeder represented the motif as "CACGAAA" with at most 1 point substitution which will wrongly consider "TACGAAA", "CAAGAAA", etc as binding sites. On the other hand, SPSP-Finder represented the motif in the following format

$$(\mathrm{C})\begin{pmatrix}\mathrm{A}\\ \mathrm{C}\\ \mathrm{G}\\ \mathrm{T}\end{pmatrix}(\mathrm{CGAA})(\mathrm{A})$$

which can represent the motif better than Weeder. Similarily, SPSP-Finder had better representations for the CuRE, LEU and MATα2 motifs.

There were two cases that SPSP-Finder failed while some of the other algorithms success. SPSP-Finder could not discover the motifs of MCM1 (SPSP-Finder discovered the published motif at rank 25) while Weeder was successful because there were no strong bias at most positions of this motif and the information contained in the input sequences was little. Weeder could discover the motifs because it had extra information about different background models for discovering motifs in different species. Since SPSP-Finder only constructed a Markov model from the input sequences, it id not have any extra information on the background model and thus failed to discover the motif.

Similarly, SPSP-Finder could not discover the motifs of MCM1 and SFF while MEME were successful because there were no strong bias at most positions of this motif. In these cases, a matrix representation can model the motif better than a string representation and the restricted SPSP representation used in RMD problem (because PSSM or PWM is a more direct and efficient representations in these cases). Excluding these two data sets, SPSP-Finder had the best performance among the algorithms.

We have also tested the performance of SPSP-Finder on the fruitfly data from the TRANSFAC database [38]. Experimental results were shown in Table 3. SPSP-Finder also had the best performance among the four algorithms. Among the 16 data sets, a total of six motifs, BEAF-32B, Cad, D_MEF2, Eve, Su_Hw and TBP, and their binding sites could be discovered by SPSP-Finder but not by the other algorithms. Again we did not list out those motifs which could not be discovered by any of the algorithms.

Table 3. Experimental results on real biological data in TRANSFAC database

Factor Name	Pattern	Weeder	MEME	AlignACE	SPSP-Finder
Ac	CGCAGGTG	CGCAGGTG	CGCAGGTG	-	$(\mathrm{CGCA})\begin{pmatrix}\mathrm{GGTG}\\ \mathrm{GCTC}\end{pmatrix}$
Antp	TTWTYMT	-	ATTTTTA	-	-
AS-CT3	CAGGTG	CAGGTG	-	-	$(\mathrm{CAGG})(\mathrm{TG})$
BEAF-32B	CGATA	-	-	-	$(\mathrm{CGAT})(\mathrm{A})$
Cad	TTTAKG	-	-	-	$(\mathrm{TTTA})(\mathrm{GG})$
Ci	TGGGTGGTC	GGGTGGTCCA	GGGTGGACC	GGGTGGTCC	$\begin{pmatrix}\mathrm{TTT}\\ \mathrm{GATG}\end{pmatrix}(\mathrm{GGGT})(\mathrm{GG})$

D_MEF2	TTAAAAATAA	-	-	-	(TTTT)[AA CG](AAA)[T A]
D1	GGGTTTTTCCN	-	GGTTTTTCCCA	-	-
DREF	ASCTATC GATADNY	GCCACC TATCGA	GCCACCT ATCGATA	-	(AGC)[TA TT](TCG)[ATA AAT TTT](A)
Eve	TNWSSYCTGC	-	-	-	[TTA TTC TTG GTG][GCT GCC GCA](CTCC)
GCM	NNACCCGCATNNN	ACCCTCATGAGT	-	-	-
Kr	AMYGGGTTAW	-	-	ACGGGTTAAG C	[TAAA TCGA GGGT](AGGG)[TT AT]
Sc	CGCAGGTG	CGCAGGTG	CGCAGGTG	-	(CGCA)[GGTG GCTC]
Su_Hw	YRYTGCATAYYY	-	-	-	(T)[G A](TTGC)(ATAC)
TBP	STATAAAW	-	-	-	[GCT ACC CCC](ATAA)(A)
Zeste	WNTTGAGTGNN	-	ACTTGAGTGA G	TTTGAGTGAGT	[TT TC GT](GAGT)(GTTT)(T)

Motifs of transcription factors that cannot be found by any algorithms were not shown in this table. 'M' stands for 'A' or 'C', 'N' stands for any nucleotide. 'D' stands for 'A', 'G' or 'T', 'K' stands for 'G' or 'T', 'R' stands for 'A' or 'G', 'S' stands for 'C' or 'G', 'W' stands for 'A' or 'T', 'Y' stands for 'C' or 'T'. Those motifs that all four algorithms can/cannot discover are not shown.

5 Concluding Remarks

In this paper, we have proposed a new and better representation based on Scored Position Specific Pattern (SPSP) to describe a motif and its binding sites. With the proposed heuristic algorithm for the Restricted Motif Discovering (RMD) Problem, we can successfully find motifs and their binding sites even in some situations for which existing popular software fail. In the RMD problem, the possible scores received by the binding sites are limited to a small set of integers. In the real biological situation, each binding site should have a different score. With this assumption, we would expect an increase in the success rate of finding the correct motif. However, finding the optimal motif for the general Motif Discovering Problem without restrictions is very difficult and should be no easier than finding the optimal motif in matrix representation. The difficulty lies not only with the large solution space of the score $\{s_{i,j}\}$, but also with the exponential number of possible sets of patterns for a length-l motif. At this moment, no heuristic algorithm for the general MD problem with reasonable performance is known.

References

1 Bailey T., Elkan C.: Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning (1995) 21:51-80

2 Barash Y., Bejerano G., Friedman N.: A simple hyper-geometric approach for discovering putative transcription factor binding sites. WABI (2001) 278-293

3 Barash Y., Elidan G., Friedman N. Kaplan T.: Modeling Dependencies in Protein-DNA Binding Sites. RECOMB (2003) 28-37

4 Buhler J., Tompa M.: Finding motifs using random projections. RECOMB (2001) 69-76

5 Bulyk M.L., Johnson P.L.F., Church G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nuc. Acids Res. (2002) 30:1255-1261

6 Chin F., Leung H.: An Efficient Algorithm for String Motif Discovery. APBC (2006) 79-88

7 Chin F., Leung H.: An Efficient Algorithm for the Extended (*l*,*d*)-Motif Problem With Unknown Number of Binding Sites. BIBE (2005) 11-18

8 Chin F., Leung H.: Voting Algorithms for Discovering Long Motifs. APBC (2005) 261-271

9 Chin F., Leung H., Yiu S.M., Lam T.W., Rosenfeld R., Tsang W.W., Smith D., Jiang Y.: Finding Motifs for Insufficient Number of Sequences with Strong Binding to Transcription Factor. RECOMB (2004) 125-132

10 Hannenhalli S., Wang L.S.: Enhanced Position Weight Matrices Using Mixture Models. Bioinformatics (2005) 21(Supp 1):i204-i212

11 Hertz G. Z., Stormo G. D.: Identification of consensus patterns in unaligned dna and protein sequences: a large-deviation statistical basis for penalizing gaps. The 3rd International Conference on Bioinformatics and Genome Research (1995) 201-216

12 Hughes J.D., Estep P.W., Tavazoie S., Church G.M.: Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. Journal of Molecular Biology (2000) 296(5):1205-14

13 Keich U., Pevzner P.: Finding motifs in the twilight zone. RECOMB (2002) 195-204

14 Kielbasa S., Korbel J., Beule D., Schuchhardt J., Herzel H.: Combining frequency and positional information to predict transcription factor binding sites. Bioinformatics (2001) 17:1019-1026

15 Lawrence C., Altschul S., Boguski M., Liu J., Neuwald A., Wootton J.: Detecting subtule sequence signals: a Gibbs sampling strategy for multiple alignment. Science (1993) 262:208-214

16 Lawrence C., Reilly A.: An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins: Structure, Function and Genetics (1990) 7:41-51

17 Leung H., Chin F.: Algorithms for Challenging motif problems. JBCB (2005) 43-58

18 Leung H., Chin F.: Discovering Motifs with Transcription Factor Domain Knowledge. PSB (2007), 472-483

19 Leung H., Chin F.: Finding Exact Optimal Motif in Matrix Representation by Partitioning. Bioinformatics (2005) 22:ii86-ii92

20 Leung H. and Chin F.: Generalized Planted (*l*,*d*)-Motif Problem with Negative Set**.** WABI (2005) 264-275

21 Leung H., Chin F., Yiu S.M., Rosenfeld R., Tsang W.W.: Finding Motifs with Insufficient Number of Strong Binding Sites. Jour. Comp. Biol. (2005) 12(6):686-701

22 Li M., Ma B., Wang L.: Finding similar regions in many strings. Journal of Computer and System Sciences (2002) 65:73-96

23 Liang S.: cWINNOWER Algorithm for Finding Fuzzy DNA Motifs. Computer Society Bioinformatics Conference (2003) 260-265

24 Man T.K., Stormo G.D. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nuc. Acids Res. (2001) 29:2471-2478.

25 Pavesi, G., Mereghetti, P., Zambelli, F., Stefani, M., Mauri, G., Pesole, G.: MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. Nuc. Acids Res. (2006) 34:566-570.

26 Pesole G., Prunella N., Liuni S., Attimonelli M., Saccone C.: Wordup: an efficient algorithm for discovering statistically significant patterns in dna sequences. Nucl. Acids Res. (1992) 20(11):2871-2875

27 Pevzner P. and Sze S.H. Combinatorial approaches to finding subtle signals in dna sequences. The Eighth International Conference on Intelligent Systems for Molecular Biology (2000) 269-278

28 Rajasekaran S., Balla S., Huang C.H.: Exact algorithms for planted motif challenge problem. APBC (2005) 249-259

29 Sinha S.: Discriminative motifs. In Proc. of the Sixth Annual International Conference on Computational Biology (2002) 291-298

30 Sinha S., Maity S.N., Lu J., and Crombrugghe B.: Recombinant rat CBF-C, the third subunit of CBF/NFY, allows formation of a protein-DNA complex with CBF-A and CBF-B and with yeast HAP2 and HAP3. Proc. Natl. Acad. Sci. USA (1995) 92(5):1624-1628

31 Sinha S., Tompa M.: A statistical method for finding transcription factor binding sites. The 8th International Conference on Intelligent Systems for Molecular Biology (2000) 344-354

32 Takusagawa K.T., Gifford D.K.: Negative information for motif discovery. PSB, (2004) 360-371

33 Tompa, M.: An exact method for finding short motifs in sequences with application to the ribosome binding site problem. The 7th International Conference on Intelligent Systems for Molecular Biology (1999) 262-271

34 Xing Y., Fikes J.D., Guarente L. Mutations in yeast HAP2 HAP3 define a hybrid CCAAT box binding domain. EMBO Journal (1993) 12:4647-4655

35 Wolfe S., Greisman H., Ramm E. and Pabo C. Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. JMB (1999) 285(5):1917-1934

36 Zhao X., Huang H., Speed T.P.: Finding Short DNA Motifs Using Permuted Markov Models. RECOMB (2004) 68-75

37 Zhu J., Zhang M.: SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics (1999) 15:563-577 http://cgsigma.cshl.org/jian/

38 TRANSFAC database http://www.gene-regulation.com/pub/databases.html

Appendix

Restricted Motif Discovering Problem is NP-complete

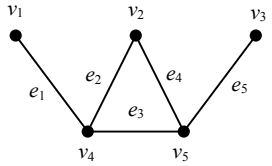
In this section, we will show that the Restricted Motif Discovering (RMD) Problem is NP-complete. In order to answer whether the RMD problem is NP-complete, we convert the RMD problem into a decision problem

> **RMD Decision Problem:** Given t length-n DNA sequences T and we assume the occurrence probability of each length-l pattern in the input sequences T is the same, whether there exists a motif P, $\prod_i |P_i| \le w$ (i.e. $\sum P(B_k) = w/4^l$), that has exactly b binding sites in T?

It is easy to see that the RMD decision problem is in NP because given a motif P, we can verify whether P has b binding sites in T and $\prod_i |P_i| \le w$ in polynomial time. In order to show that the RMD decision problem is NP-complete, we reduce the *Clique Decision Problem* (CDP) to it.

> **Clique Decision Problem (CDP):** Given a graph $G = (V,E)$ and an integer $k > 0$, the CDP is to determine whether G contains a clique of size k.

Denote $V = \{v_i, 1 \le i \le n\}$ and $E = \{e_j, 1 \le j \le m\}$. Let $\deg(v_i)$ be the degree of vertex v_i and $D = \max_i\{\deg(v_i)\}$. We construct $2n$ length-$(nD - m)$ DNA sequences as follows: For each vertex v_i, a length-$(nD - m)$ DNA sequence σ_i representing a binding site is constructed such that σ_i has all symbols 'T' except D symbols of 'A' or 'C'. The first m symbols of these n sequences (one for each vertex) resemble the incidence matrix such that the j^{th} symbols of σ_i and $\sigma_{i'}$ are 'A' and 'C' corresponding to the j^{th} edge connecting v_i and $v_{i'}$ respectively. Thus σ_i should have $\deg(\sigma_i)$ symbols of 'A' or 'C' in its first m symbols. If $\deg(\sigma_i) < D$, then $D - \deg(\sigma_i)$ symbols of 'A' will be packed after the first m symbols such that no two sequences have symbol 'A' at the same position and each σ_i has exactly D symbols of 'A' or 'C'. Precisely, we have



$G = \{V,E\}$

CDP: whether there is a clique of size 3 in G?

	1 2 3 4 5 6 7 8 9 10	
v_1	A T T T T A A T T T	T_1
v_2	T A T A T T T A T T	
v_3	T T T T A T T T A A	
v_4	C C A T T T T T T T	
v_5	T T C C C T T T T T	
	T T T T T T T T T T	T_2
	T T T T T T T T T T	
	T T T T T T T T T T	
	T T T T T T T T T T	
	T T T T T T T T T T	

RMP Problem: whether there is a motif, $\prod_i |P_i| \le 2^{3 \cdot 3 - 3(3-1)}\, 3^{3(3-1)/2} = 2^3 3^3$ having $5 + k$ binding sites in T?

Fig. 1. An example of the reduction from CDP to RMD Decision Problem

$$\sigma_i[j] = \begin{cases} \text{'A'} & \exists \mu,\ \{v_\mu, v_i\} = e_j, \mu > i \\ \text{'C'} & \exists \mu,\ \{v_\mu, v_i\} = e_j, \mu < i \\ \text{'A'} & |E| + \sum_{i'=1}^{i-1} \left(D - \deg(v_{i'})\right) < j \le |E| + \sum_{i'=1}^{i} \left(D - \deg(v_{i'})\right) \\ \text{'T'} & \text{otherwise} \end{cases}$$

Denote this set of n strings by T_1.

In addition to these n length-$(nD - m)$ DNA sequences, we have another n length-$(nD - m)$ DNA sequences with symbol 'T' only. Denote this set of strings by T_2, $T = T_1 \cup T_2$. We solve the RMD decision problem with $l = \alpha = nD - m$, $l_{\max} = 1$ and $w = 2^{kD - k(k-1)} 3^{k(k-1)/2}$. If there exists a motif P, $\prod_i |P_i| \le w$ having $b = n + k$ binding sites in T, the answer of CDP is "yes", otherwise, the answer is "no". Figure 1 shows an example of this reduction. Theorem 1 proves the correctness of this reduction.

Theorem 1: There is a motif P with $\prod_i |P_i| \le 2^{kD - k(k-1)}\, 3^{k(k-1)/2}$ and having $b = n + k$ binding sites in T if and only if there is a clique of size k in G.

Proof: w.l.o.g assume $\{v_i \mid 1 \le i \le k\}$ with $\{e_j \mid 1 \le j \le k(k-1)/2\}$ forms a clique of size k in G, the set of binding sites should contain all the strings in T_2 and k strings (corresponding to the vertices of the clique) in T_1, i.e. $n + k$ strings. The motif should have its first $k(k-1)/2$ positions having the symbols 'A', 'C" and 'T', i.e.

$$P_1 = P_2 = \ldots = P_{k(k-1)/2} = \begin{pmatrix} \text{A} \\ \text{C} \\ \text{T} \end{pmatrix}$$

and exactly $kD - k(k-1)$ positions having the pair of symbols 'A', 'T' or 'C', 'T'. Note that all the other positions should be conserved and have symbol 'T'. Thus, these $n + k$ strings can be represented in SPSP representation with $\prod_i |P_i| = 2^{kD - k(k-1)} 3^{k(k-1)/2}$.

Assume there is a motif P in the SPSP representation with $\prod_i |P_i| \le 2^{kD - k(k-1)} 3^{k(k-1)/2}$ and having exactly $n + k$ binding sites and y ($y \ge k$) out of these $n + k$ binding sites in set T_1. Since each binding site in T_1 has exactly D symbols of 'A' or 'C' and each of these yD symbols can be either represented by a partition with two symbols or three symbols as follows.

$$\begin{pmatrix} A \\ T \end{pmatrix} \text{or} \begin{pmatrix} C \\ T \end{pmatrix} \text{or} \begin{pmatrix} A \\ C \\ T \end{pmatrix}$$

The motif P has the smallest $\prod_i |P_i|$ when it has the largest possible number $y(y-1)/2$ of partitions with three symbols and the smallest value of y ($y = k$). We have $\prod_i |P_i| \geq 2^{yD - y(y-1)} 3^{y(y-1)/2} \geq 2^{kD - k(k-1)} 3^{k(k-1)/2}$. Therefore $\prod_i |P_i| = 2^{kD - k(k-1)} 3^{k(k-1)/2}$ and $y = k$. Since 2 and 3 are prime numbers, there are $k(k–1)/2$ pattern sets P_i with 3 symbols of 'A', 'C' and 'T' and the corresponding vertices of the k sequences in T_1 form a clique of size k in G. □