以混合 0-1 變數線性規劃方法
尋找基因轉錄因子結合位點

研究生：傅昶瑞　　　　　　　　　　指導教授：黎漢林

國立交通大學資訊管理研究所

摘要

基因轉錄因子結合位點 (Transcription factor binding site, TFBS) 的搜尋在基因組的功能性分析上扮演關鍵性的角色。在眾多搜尋方法之中，以共同序列(consensus sequence)為基礎的窮舉法相對最為準確，但其指數成長的運算量卻讓這類方法無法搜尋較長的序列。藉由預先給定的樣版來輔助搜尋能顯著地減少運算量，只是這類資訊較難以取得，也因此目前 TFBS 的搜尋大多仍仰賴準確率低但速度較快的啟發式方法。

為了能有效準確搜尋 TFBS，本研究發展一套混合 0-1 線性規劃法來求解三種不同類型的 TFBS 搜尋問題。這包括固定樣版 TFBS 搜尋，模糊樣版 TFBS 搜尋，以及無樣版 TFBS 搜尋。本方法的優點包括：(1)以共同序列為基礎的設計，(2)可得到全域最佳解，以及(3)可套用結構性的限制以加速運算並提高準確率。而在無樣版 TFBS 搜尋中，本方法更可以成功找到因結構鬆散而難以發現的 TFBS。本研究以多個範例來針對三種不同類型的 TFBS 搜尋問題進行一系列實驗，也都在可接受的時間下成功找到了實際存在的 TFBS。

關鍵詞：DNA-蛋白質交互作用，基因調控，轉錄因子結合位點，線性規劃，整數規劃

A mixed 0-1 linear programming approach for finding transcription factor binding sites

Student: James Changjui Fu　　　　Advisor：Hanlin Li

Institute of Information Management
National Chiao Tung University

ABSTRACT

The discrimination of transcription factor binding sites (TFBS) in multiple DNA sequences is an essential work for function analysis of gene expression. Enumeration methods that search all possible patterns have best precision among all current algorithms but require an exponential computational time and have difficulties to search for longer patterns. A predefined shared pattern can notably prunes the searching space but such information is often unavailable. Finding unframed TFBS today still relies on heuristic approaches which compromise to accuracy.

To effectively find TFBS, this study develops a mixed 0-1 linear programming approach to solve a series of problems for issues including fixed-pattern TFBS finding, ambiguous spacer TFBS finding and pattern-free TFBS finding. The proposed method has the following advantages over current methods: (1) A pattern-driven instead of sample-driven (or sequence-driven) design; (2) A global optimal solution is promised; (3) Structural features of motifs are embeddable to help facilitate search process. And with pattern-free approaches we can successfully determine TFBS within dispersed spacers. We apply several experiments on every kind of TFBS finding programs and in these examples the real TFBS are successfully determined in an acceptable computational time.

Keywords: DNA-protein interaction, gene regulation, transcription factor binding site, linear programming, integer programming.

誌謝

本研究的完成，首先要感謝我的指導教授黎漢林老師。沒有他當初的提點與指導，我不會有今日豐碩的研究成果。另外我要特別感謝我的口試委員游伯龍老師、曾國雄老師、陳茂生老師、唐傳義老師、林妙聰老師、以及盧錦隆老師，在我撰寫博士論文期間，他們所給予的寶貴建議與協助，使我的博士論文能夠有更為嚴謹完善的內容。

我更要感謝我最親愛的父母以及哥哥，他們對我不變的信心以及鼓勵，是我完成學業最大的原動力。還有我的好友晏安，總能讓我在意想不到的地方，得到更深一層的領悟與靈感。在博士研究的期間有許多快樂美好的回憶，感謝在這段時間陪伴我的諸位同學朋友，讓我的生活過得充實且多采多姿。深深地感謝大家，希望未來我們能在更大的舞台上相聚。

CONTENTS

TABLE

FIGURE

Chapter 1 Introduction

For past two decades, biologists have sequenced more and more complete genome sets of various species. To extract all the secrets of life from these huge data, procedures of how genes work in organism are continuously researched and discussed. Gene transcription, a primary gateway to gene function, is controlled by a complex regulatory mechanism in which many specific regulatory proteins bind to local regions of gene upstream, called *transcription factor binding sites* (TFBS), to control the gene expression. Therefore, the discrimination of TFBS from DNA sequences therefore becomes an essential work for genome function analysis.

1.1 DNA-binding Motifs and Their Binding Sites

Before the discussion about TFBS, we need to know the mechanism of gene regulatory. DNA transcription is the very first stage of gene expression. The complexes of *Rribonucleic acid (RNA) polymerases* and *general transcription factors* transcribe all kinds of genes at a basal level—like an idling engine—to remain the minimum operation. In fact, the transcription of active genes generally rises far above this basal level. To provide the needed extra boost in transcription, additional *gene-specific transcription factors* (TF) play the critical role to control the throttle. These transcription factors, also called *regulators* or *activators*, are like a set of keys capable of unlocking or locking the transcription. They bind to specific locations—like many particular keyholes—to stimulate or inhibit RNA polymerase to transcribe a gene. The activation of a gene relies on presence of all required enhancers and absence of all inhibitors (or at a low safe level).

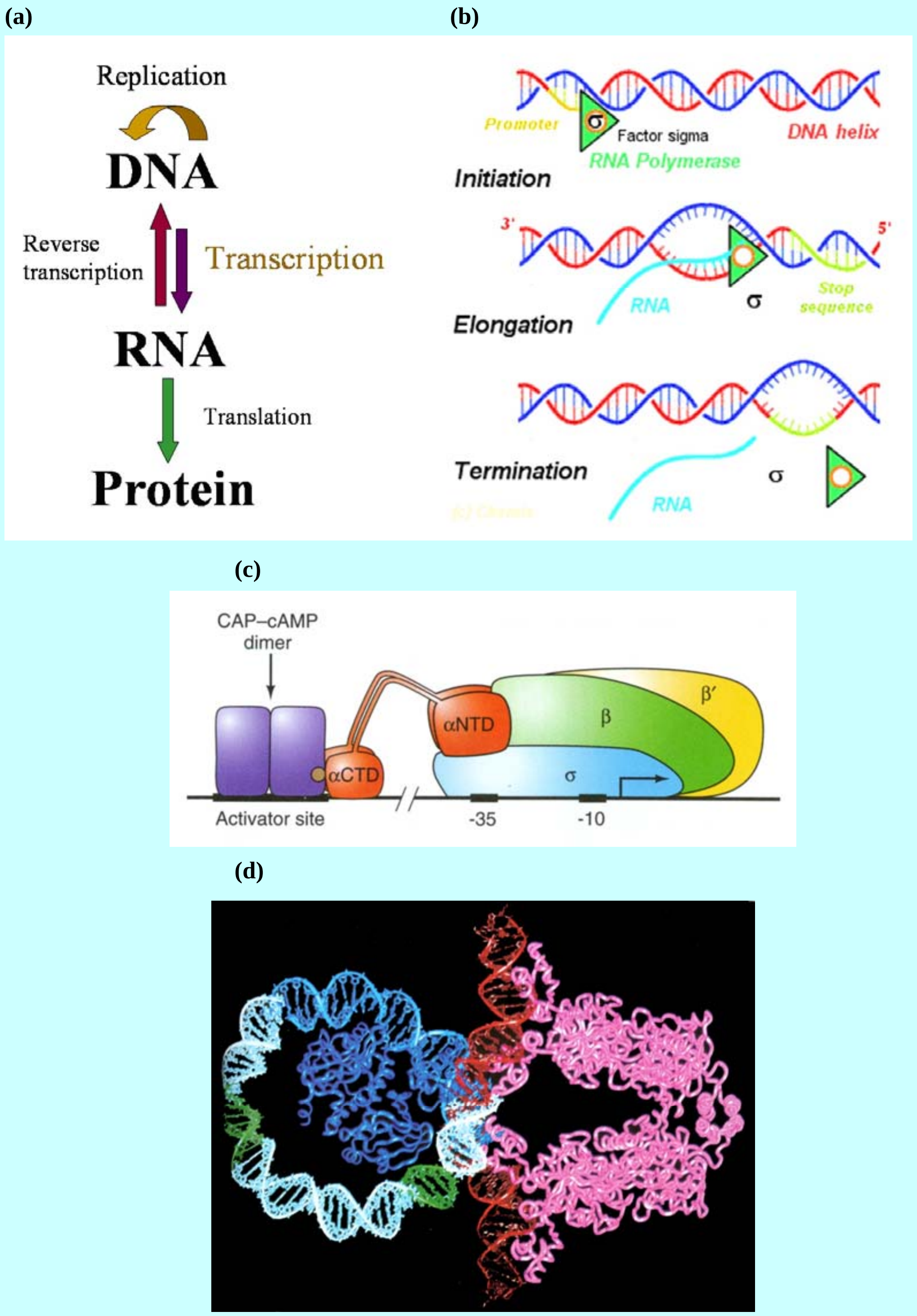


Figure 1 Gene expression. **(a)** Central dogma; **(b)** Transcription of a gene in prokaryotes; **(c)** The complex of DNA strand, RNA polymerase, general transcription factors and CAP-cAMP (CRP) dimer—a gene specific transcription factor; **(d)** Computer graphic of *lac* repressor (pink) and CRP dimer (blue) binding to DNA.

Activators have at least two functional domains: a DNA-binding domain and a transcription-activation domain. Many also have a dimerization domain that allows

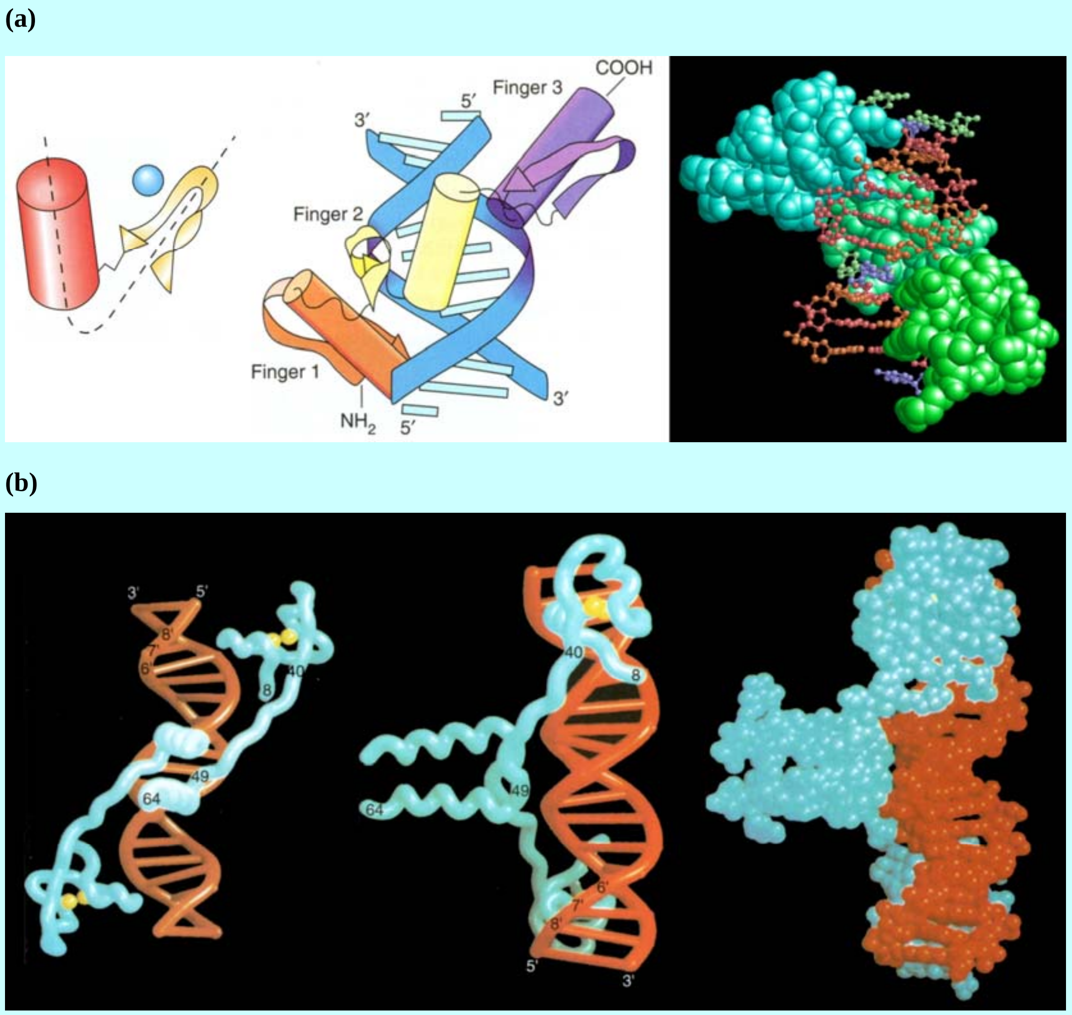


Figure 2 Zinc-containing modules: (a) Zinc fingers (Zif268), consisted by a series of zinc finger which contains a zinc ion; (b) The GAL4 protein, a dimerized motif which contains two zinc ions in each monomer.

the activators to bind to each others, forming homodimer (two identical monomers bound together), heterodimers (two different monomers bound together), or even higher multimers such as tetramers [Weaver, 2002]. Each DNA-binding domain, the most part we concern about, has a DNA-binding motif, which is the part of the domain that has a characteristic specialized for specific DNA binding. Most DNA-binding motifs fall into the following classes:

1. **Zinc-containing module**s. These modules use one or more zinc ions to create a proper shape to bind to DNA and include at least three kinds of modules. The most often seen is zinc fingers, which is a chain of two or more zinc finger

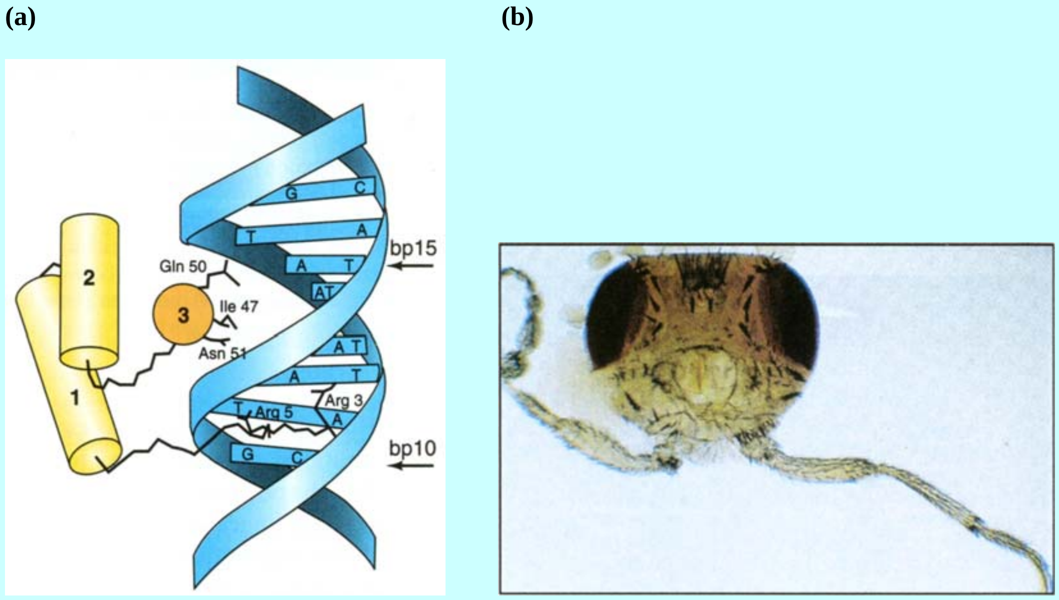


Figure 3 Homeodomain-DNA complex in fruit fly *Drosophila*--an example of mono-type interaction. **(a)** Schematic representation; **(b)** A deformation caused by mutations in genes of these regulators: *Antennapedia.* It grows legs where antennae would normally be.

monomers. Some zinc containing motifs also have dimerization domain containing two identical monomers, e.g. the GAL4 motif.

2. **Homeodomains (HDs).** These resemble in structure and function the helix-turn-helix DNA-binding domains such as the λ phage repressor. The mutation of their gene may cause severe deformation. Most homeodomain proteins have weak DNA-binding specificity and rely on other proteins to help them bind specifically and efficiently to their DNA targets.

3. **bZIP and bHLH motifs.** Most DNA-binding motifs are of this type. They have a highly basic DNA-binding motif linked to one or both of the protein dimerization motifs known as leucine zippers and helix-loop-helix (HLH) motifs. This kind of motifs have very strong DNA-binding specificity.

These three classes cover a large majority of DNA-binding motifs but certainly the list is not exhaustive. There are still other kinds of DNA-binding motifs not falling into any of these categories.

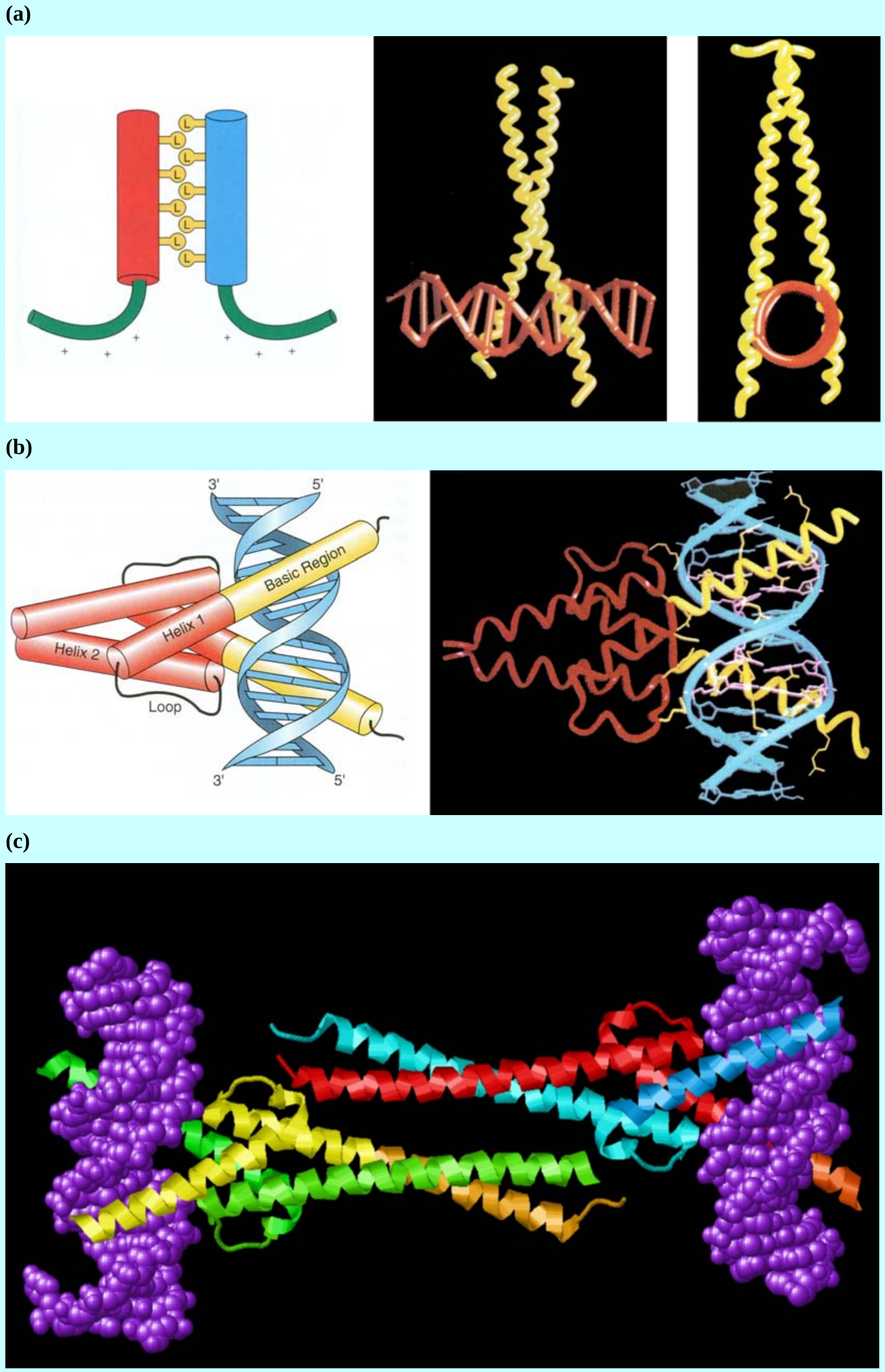


Figure 4 Dimerized DNA binding domain: **(a)** Leucine zipper (bZIP) complex. From left to right: dimerization of leucine zipper and two computer graph illustrating binding domain; **(b)** Two schematic diagrams of Helix-loop-helix (bHLH) complex; **(c)** Max-Myc heterotetramer.

Transcription factor binding site is a short region within a particular nucleotide sequence for a specific activator to bind. Because of various domains of DNA-protein interaction, TFBS linked to different kinds of DNA-binding motifs has particular characteristics for binding. Most TFBS can be categorized into three types:

(1) **Mono-type TFBS.** This kind of TFBS is for binding a monomer. DNA binding domains like homeodomains have their binding sites of this type. Most mono-type TFBS are relatively weak signals and difficult to determined. In fact, their binding motifs usually require other auxiliary protein-protein interaction domains or DNA-protein binding domains to help their binding.

(2) **Dyad-type TFBS.** Dimerized regulators bind to this kind of TFBS. Dyad-type TFBS is the most often seen type and generally not longer than 22 bases. It consists of two symmetrical half binding sites with a fixed number of in-between spacers. As a result this kind of TFBS has very strong binding specificity to regulators and relatively easy to determined.

(3) **Series-type TFBS.** The binding sites of chain-like regulatory protein like zinc-fingers are of this type. This kind of TFBS contains several adjacent short units of the same size. For example of zinc fingers Zif268, it has binding site consisting of three units each of which is three-base long.

These types of TFBS have different features that make the specificity for recognition. These features can be regarded as logical rules that might be helpful for TFBS determination.

1.2 TFBS finding problem

To find TFBS, one has a collection of sequences that are known to contain binding sites for a common factor, but neither the positions of the sites nor the specificity of the factor are known. Besides that, TFBS are usually with some degree of ambiguity. These make TFBS finding a difficult and challenging problem. Experimental methods like DNA microarray (DeRisi et al., 1997; Lockhart et al., 1996) and SAGE (Velculescu et al., 2000) are capable to precisely elucidate TFBS. However, they are too laborious and time consuming to analyze enormous genome data. More and more computer based methods like enumeration methods, probability models and heuristics have been developed to find these conserved signals. In this section we discuss current computer-based (say *in silico*) approaches and their limitations.

Site Representation

Most transcription factor binding sites have variability on their component bases. With this ambiguity regulatory system can take advantage of level control on the gene expression. This makes the representation of DNA binding sites more complicated. How to precisely describe this variability depends on what kind of methods is applied in searching TFBS. Generally TFBS searching methods can be classified into two categories: *pattern-driven* approaches and *sequence-driven* (also called *alignment-driven*) approaches. Pattern-driven approaches search for a *consensus sequence* which best fits all site occurrences. And the representation of this consensus sequence includes simple DNA sequence and *IUPAC* (acronym of: *International Union of Pure and Applied Chemistry*) code sequence. Sequence-driven approaches identify the site occurrences which maximize *position weight matrix* (PWM) and *information content* (IC).

The simplest TFBS representation is merely a DNA sequence consisted only by A, T, G and C. Although incapable of describing base variability, this expression is still useful in pattern-driven enumeration methods. This is because flexible representation like IUPAC code will lead to enormous searching space in enumeration methods.

IUPAC is a degenerate naming rule consisting of 16 alphabets which describe various combinations of nucleic acids codes, shown in Table 1. Any kind of ambiguities in nucleic acids has a corresponding code and so IUPAC code can be used to completely describe a TFBS consensus. An obvious defect of IUPAC code is that it fails to describe the base preference level at each position. Position weight matrix (PWM) is designed for more precisely describing base variability.

In PWM the significance of a particular TFBS consensus is given by a measure of statistical surprise from multiple aligned short sequences. It calculates *log*

Table 1 IUPAC code for nucleic acids

IUPAC code	Description
A	Adenine
T	Thymine
C	Cytosine
G	Guanine
U	Uracil
R	A, G (purine)
Y	T, C (pyrimidine)
K	G, T (keto)
M	A, C (amino)
S	G, C (strong)
W	A, T (weak)
B	T, G, C
D	A, T, G
H	A, T, C
V	A, G, C
N	A, T, G, C (any)

(a) Alignment matrix

	Base	Position				
Site Sequence	1	2	3	4	5	6
Site #1	**A**	**A**	**T**	**T**	**C**	**A**
Site #2	**A**	**G**	**G**	**T**	**A**	**C**
Site #3	**A**	**G**	**T**	**C**	**C**	**G**
Site #4	**A**	**A**	**T**	**T**	**C**	**A**
Site #5	**A**	**G**	**G**	**T**	**A**	**T**
Site #6	**A**	**G**	**G**	**T**	**C**	**C**
Site #7	**A**	**G**	**G**	**A**	**T**	**G**
Site #8	**A**	**G**	**G**	**C**	**C**	**T**
IUPAC sequence	**A**	**R**	**K**	**H**	**H**	**N**

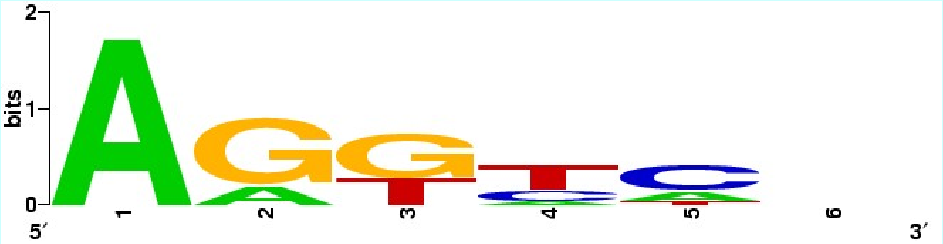
(b) Profile

$f_{b,i}$	1	2	3	4	5	6
A	**1**	**0.25**	**0**	**0.125**	**0.25**	**0.25**
T	**0**	**0**	**0.375**	**0.625**	**0.125**	**0.25**
G	**0**	**0.75**	**0.625**	**0**	**0**	**0.25**
C	**0**	**0**	**0**	**0.25**	**0.625**	**0.25**

(c) Position Weight Matrix

$LLR(b,i)$	1	2	3	4	5	6
A	**2**	**0**	$-\infty$	**-1**	**0**	**0**
T	$-\infty$	$-\infty$	**0.585**	**1.322**	**-1**	**0**
G	$-\infty$	**1.585**	**1.322**	$-\infty$	$-\infty$	**0**
C	$-\infty$	$-\infty$	$-\infty$	**0**	**1.322**	**0**

(d) Logo model



(e) Relative entropy

Position	1	2	3	4	5	6
Relative Entropy	**2**	**1.189**	**1.046**	**0.701**	**0.701**	**0**

Information content: I_{seq} **= 5.637**

Figure 5 PWM and IC representation: **(a)** aligned site sequences and their consensus as an IUPAC sequence; **(b)** the profile of these sites; **(c)** Position Weight Matrix; **(d)** Logo display of site sequences; **(e)** Information content of site sequences.

likelihood ratio (LLR) of four nucleic acids at each position as:

$$LLR(b,i) = \log_2 \frac{f_{b,i}}{p_b} \qquad \textbf{(1)}$$

where *i* is the position within the site, $b \in \{\mathrm{A, T, G, C}\}$ refers to each of the possible bases, $f_{b,i}$ is the observed frequency of each base at that position and p_b is the frequency of base *b* in the whole genome. The maximum LLR among each position are summed up as the significance of a given set of sites.

Information content (IC, Schneider et al., 1986), which is also known as the Kullback-Leibler distance, is the sum of all relative entropies of four types of bases in all positions defined as below:

$$I_{seq} = \sum_i \sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b} \qquad \textbf{(2)}$$

For pattern-driven methods, the criterion for best conserved consensus sequence is to find the one with maximum matches among all site occurrences. For sequence-driven methods, the criterion for identifying the best alignment of potential sites is to choose the one with highest information content I_{seq}.

Site Discovering

Pattern recognition approaches can also be categorized into two classes: pattern-driven approaches and sequence-driven approaches. As previously mentioned, pattern-driven approaches search for consensus sequence which best fits all site occurrences. For consensus-based TFBS finding (Stormo, 2000), pattern-driven algorithms that test all 4^m *m*-wide possible consensus sequences promise an optimal solution but are very time consuming and impractical for large *m*. (Pesole et al., 1992; Tompa, 1999) Many heuristics are developed to prune the huge searching space including testing only the substrings in the sequences (Li et al., 1999; Gelfand et al., 2000), specifying a shared pattern to restrict the locations of mismatches (Brazma et al., 1998; Califano, 2000; Sinha and Tompa, 2003; Régnier and Denise, 2004; Li and

Fu, 2005) and clustering (Buhler and Tompa, 2002; Pevzner and Sze, 2000; Liang et al., 2004). In addition to the exact enumeration methods, efficient data structure like suffix tree with fixed mismatches (Pavesi et al., 2001; 2004) can search for patterns of longer length. This kind of approaches is not exact enumeration algorithm and takes advantages of searching time polynomial to pattern length and exponential to the number of tolerant mutations.

Sequence-driven methods are designed based on probabilistic modeling. The challenge of sequence-driven approaches is to find the location of the sites and representative PWM using only the sequence data, without any assumptions on the statistical distributions of patterns in the sequences. The criterion for the best alignment is the one with maximum IC. Current methods include a greedy algorithm that builds up an entire alignment of sites by adding in new ones in each iteration (Stormo and Hartzell, 1989; Hertz et al., 1990) and expectation maximization (EM) that iteratively substitute the location of sites by expected locations (Lawrence and Reilly, 1990) and its variant, Gibbs sampling (Lawrence et al., 1993) as a type of Markov chain Monte Carlo (MCMC) algorithm. EM algorithm is also implemented in the MEME program (Bailey and Elkan, 1995) which allows for the simultaneous identification of multiple patterns. Other implements of sequence-driven approaches include CONSENSUS (Hertz and Stormo, 1999), AlignACE (Hughes et al., 2000), ANN-spec (Workman and Stormo, 2000), BioProspector (Liu et al., 2001), MotifSampler (Thijs et al., 2001), GLAM (Frith et al., 2004), The Improbizer (Ao et al., 2004), QuickScore (Régnier and Denise, 2004), SesiMCMC (Favorov et al., 2004) and TFBSfinder (Tsai et al., 2006).

In most current TFBS finding methods, all the letters in the consensus sequence are treated as independent variables. Because only some bases in binding region are reactive to the transcription factor, solving this problem by calculating scores of all bases may involve noise from bases inducting no interactions. Beside that, the assumption of independent and identically distributed bases in background is too strong. Even with a probability calculated from the sequence data, the contribution to the accuracy is still limited. Another type of heuristics include testing only the substrings in sequences and constructing data structures like a suffix tree or a graph to extract overrepresented signals. This kind of methods compromises to a possible situation of weeding out the exact consensus when all the motifs in sequences are somehow ambiguous.

Most current methods also have obstacles to involve specific TFBS features like inverted palindrome or direct repeats. By limitations from original concepts, statistical models like EM or HMM need a much more complex design to embed the structural features. In some tree-based enumeration methods it is even impossible to utilize these structural features.

1.3 Formulation of Pattern-driven TFBS Finding

In this study a pattern-driven approach utilizing mixed 0-1 linear program is proposed. A pattern-driven concept of discovering TFBS is to find the consensus which has maximum matches among all proposed sites from multiple sequences. This is a mixed 0-1 optimization problem and can be formulated as a mixed 0-1 nonlinear program. We start by formulating a fixed-pattern TFBS finding problem as a mixed 0-1 nonlinear program. In many cases a predefined shared-pattern is available from some preprocesses. This shared pattern provides information about positions of reactive bases in the binding sites and makes a TFBS finding problem relatively easier to solve.

Representations of fixed-pattern TFBS finding

To find TFBS of a specific regulation, a set of DNA sequences upstream genes known co-regulated by the same factor is firstly prepared for analysis. A prerequisite condition is that this DNA sequence set shall be prepared having at least one occurrence per sequence (OOPS). Namely, there exists at least one similar TFBS in each sequence. A pattern-driven TFBS finding problem is defined as:

> Given
>
> (i) A sequence set containing L sequences,
>
> (ii) A shared pattern "NNNNN******NNNNN" in which 'N' and '*' represent reactive and inactive bases respectively.
>
> To find the best conserved consensus sequence
>
> $x_1x_2x_3x_4x_5 ******x_6x_7x_8x_9x_{10}$,
>
> where $x_i \in \{ \text{A, T, G, C} \}$ and i is the index of reactive base.

To find best conserved consensus sequence among the given sequence set, the first step is to generate a set of candidate sites from sequence data. We use the example of CRP-binding sites among DNA sequences of *Escherichia coli* (Stormo et al., 1989), shown in Appendix, to illustrate the formulation. According to the predefined shared pattern, candidate sites are extracted from each starting position of each sequence, as shown in Figure 6, and indexed by (*l*, *s*) where *l* is the sequence index and *s* is the start position. Denote $d_{l,s,i} \in \{\text{A, T, G, C}\}$ as the i^{th} base present in candidate site from (*l*, *s*) position.

With the pattern-driven concept, denote the consensus sequence to find as a series of binary variables. Every reactive base (i.e., the notation 'N' in the shared

(a)

Sequence 1 (cole1): TAATGTTTGTGCTGGTT

Sequence 2 (eco arabop): GACAAAAACGCGTAACA

(b)

$(l, s) = (1,1)$ TAATG......CTGGT

$(l, s) = (1,2)$ AATGT......TGGTT

$(l, s) = (1,3)$ ATGTT......GGTTT

$(l, s) = (1,4)$ TGTTT......GTTTT

$(l, s) = (2,1)$ GACAA......GTAAC

$(l, s) = (2,2)$ ACAAA......TAACA

$(l, s) = (2,3)$ CAAAA......AACAA

$(l, s) = (2,4)$ AAAAA......ACAAA

Figure 6 Site extraction for fixed-pattern TFBS finding (The first two sequences in Appendix): **(a)** original sequence data; **(b)** schematic representation of the candidate sites.

pattern) in a consensus sequence is represented by two binary variables, u and v for four different nucleotides A, T, C and G, and indexed by its relative position, *i*. Obviously the example of CRP-binding sites needs 20 binary variables to represent the consensus sequence. The binary codes for four nucleotide types are defined in Table.2. Each feasible consensus sequence with a vector of (*u*, *v*) pairs is scored by summing up base matches compared with the best fitting candidate site in every sequences. To formulate the comparison, every base appearing in a candidate site is represented by four kinds of comparison functions for different base types, as defined in follows.

$$\begin{aligned} y_{\mathrm{A},i} &= (1-u_i)(1-v_i), \\ y_{\mathrm{T},i} &= u_i v_i, \\ y_{\mathrm{G},i} &= u_i(1-v_i), \\ y_{\mathrm{C},i} &= (1-u_i)v_i. \end{aligned} \qquad \textbf{(3)}$$

The illustrative table for the base comparison between the consensus sequence and candidate sites is listed in Table 3.

Illustrative formulation for maximizing matches

The objective of a fixed-pattern TFBS finding is to find the consensus sequence best conserved among all the input DNA sequences. For every single DNA sequence, the scoring criterion is to compare its best fitting candidate sites with the consensus sequence and to count the base matches. This can be formulated as below:

$$Score_l = \max\left\{ \sum_s z_{l,s}(\theta_{l,s,1} + \theta_{l,s,2} + \ldots + \theta_{l,s,10}) \;\middle|\; \sum_s z_{l,s} = 1 \right\}, \qquad \textbf{(4)}$$

where $z_{l,s}$ is the binary indicator of whether the candidate site at (l, s) is chosen to compare with the consensus. A candidate site is scored only when its corresponding z equals 1. All other non-basic candidate sites will have its corresponding z valued 0. For the assumption of one occurrence per sequence (OOPS), only the candidate site

Table 2 The binary coding for each four bases

Base in consensus sequence	u	v
A	0	0
T	1	1
G	1	0
C	0	1

Table 3 Illustrative table of base comparison

Comparison Table		Base in consensus sequence			
Base in candidate sites	Comparison function	A	T	G	C
A	(1-u)(1-v)	1	0	0	0
T	uv	0	1	0	0
G	u(1-v)	0	0	1	0
C	(1-u)v	0	0	0	1

that best fits the consensus is to be scored in a sequence. That means for all *l*,

$$\sum_{s} z_{l,s} = 1, \quad z_{l,s} \in \{0,1\}. \tag{5}$$

In sequence scoring fiunction (4), $\theta_{l,s,i}$ is the comparison function defined by the i^{th} base in the candidate site from (*l*, *s*). That is,

$$\theta_{l,s,i} = \begin{cases} y_{\mathrm{A},i} & \text{if } d_{l,s,i} = \mathrm{A} \\ y_{\mathrm{T},i} & \text{if } d_{l,s,i} = \mathrm{T} \\ y_{\mathrm{C},i} & \text{if } d_{l,s,i} = \mathrm{C} \\ y_{\mathrm{G},i} & \text{if } d_{l,s,i} = \mathrm{G}. \end{cases} \tag{6}$$

Every candidate site is evaluated by summing up $\theta_{l,s,i}$'s for a given (*u*, *v*) pair. For example of the first candidate site in Figure 6, "TAATG......CTGGT", the site score (i.e. number of base matches) is obtained from $\sum_{i=1}^{10}\theta_{l,s,i}$ = $(y_{\mathrm{T},1} + y_{\mathrm{A},2} + y_{\mathrm{A},3} + y_{\mathrm{T},4} + y_{\mathrm{G},5} + y_{\mathrm{C},6} + y_{\mathrm{T},7} + y_{\mathrm{G},8} + y_{\mathrm{G},9} + y_{\mathrm{T},10})$. When comparing with a consensus "TCATG******CATGA", this score function will give 6 as the site score for six matched letters.

The matching score of sequence 1 in Figure 6 is formulated as

$$\begin{aligned} Score_1 = \max\Big\{ & z_{1,1}(y_{\mathrm{T},1} + y_{\mathrm{A},2} + y_{\mathrm{A},3} + y_{\mathrm{T},4} + y_{\mathrm{G},5} + y_{\mathrm{C},6} + y_{\mathrm{T},7} + y_{\mathrm{G},8} + y_{\mathrm{G},9} + y_{\mathrm{T},10}) \\ & + z_{1,2}(y_{\mathrm{A},1} + y_{\mathrm{A},2} + y_{\mathrm{T},3} + y_{\mathrm{G},4} + y_{\mathrm{T},5} + y_{\mathrm{T},6} + y_{\mathrm{G},7} + y_{\mathrm{G},8} + y_{\mathrm{T},9} + y_{\mathrm{T},10}) \\ & + z_{1,3}(y_{\mathrm{A},1} + y_{\mathrm{T},2} + y_{\mathrm{G},3} + y_{\mathrm{T},4} + y_{\mathrm{T},5} + y_{\mathrm{G},6} + y_{\mathrm{G},7} + y_{\mathrm{T},8} + y_{\mathrm{T},9} + y_{\mathrm{T},10}) \\ & + z_{1,4}(y_{\mathrm{T},1} + y_{\mathrm{G},2} + y_{\mathrm{T},3} + y_{\mathrm{T},4} + y_{\mathrm{T},5} + y_{\mathrm{G},6} + y_{\mathrm{T},7} + y_{\mathrm{T},8} + y_{\mathrm{T},9} + y_{\mathrm{T},10}) \\ & \Big| \sum_{s} z_{1,s} = 1 \Big\}. \end{aligned} \tag{7}$$

For a fixed-pattern TFBS finding problem, the objective is to maximize the total matches among all the sequence, i.e., $\max\sum_{l} Score_l$. When finding CRP-binding

$$
\begin{aligned}
&\text{Max} \sum_{l=1}^{18} \text{Score}_l \\
&\text{s.t.} \\
&\text{Score}_1 = z_{1,1}[y_{\text{T},1} + y_{\text{A},2} + y_{\text{A},3} + y_{\text{T},4} + y_{\text{G},5} + \ldots + y_{\text{T},10}] \\
&\qquad + z_{1,2}[y_{\text{A},1} + y_{\text{A},2} + y_{\text{T},3} + y_{\text{G},4} + y_{\text{T},5} + \ldots + y_{\text{T},10}] \\
&\qquad + \ldots \\
&\qquad + z_{1,90}[y_{\text{T},1} + y_{\text{C},2} + y_{\text{C},3} + y_{\text{A},4} + y_{\text{C},5} + \ldots + y_{\text{G},10}], \\
&\text{Score}_2 = z_{2,1}[y_{\text{G},1} + y_{\text{A},2} + y_{\text{C},3} + y_{\text{A},4} + y_{\text{A},5} + \ldots + y_{\text{C},10}] \\
&\qquad + z_{2,2}[y_{\text{A},1} + y_{\text{C},2} + y_{\text{A},3} + y_{\text{A},4} + y_{\text{A},5} + \ldots + y_{\text{A},10}] \\
&\qquad + \ldots \\
&\qquad + z_{2,90}[y_{\text{C},1} + y_{\text{A},2} + y_{\text{T},3} + y_{\text{T},4} + y_{\text{T},5} + \ldots + y_{\text{G},10}], \\
&\text{Score}_3 = \ldots, \\
&\quad \vdots \\
&\text{Score}_{18} = \ldots, \\
&\sum_{s=1}^{90} z_{l,s} = 1 \quad \forall l \in \{1,\ldots,18\}, \\
&\left.\begin{aligned} y_{\text{A},i} &= (1-u_i)(1-v_i), \quad y_{\text{T},i} = u_i v_i, \\ y_{\text{C},i} &= (1-u_i)v_i, \quad y_{\text{G},i} = u_i(1-v_i), \end{aligned}\right\} \forall i \in \{1,\ldots,10\}, \\
&u_i, v_i \in \{0,1\}, \quad z_{l,s} \in [0,1], \\
&b \in \{\text{A},\text{T},\text{G},\text{C}\}, \quad i \in \{1,\ldots,10\}, \\
&l \in \{1,\ldots,18\}, \quad s \in \{1,\ldots,90\}.
\end{aligned}
$$

Figure 7 A mixed 0-1 nonlinear program for finding CRP-binding sites.

sites in *E.Coli* (see Appendix for complete data set), the mixed 0-1 nonlinear program is formulated as shown in Figure 7. There are 18 sequences each of which 105-bp long in this example. Because the length of the given pattern is 16 (i.e. “NNNNN******NNNNN”), we have 90 candidate sites in each sequence. The independent variables include 20 binary variables (i.e. *u* and *v*) for consensus sequence and 18*90 binary variables (i.e. *z*) for indicating proposed sites. The notation *y* in this program is used as comparison function for different cases and the number of these constraints for comparison is 40. This program has 18 conservation constraints (i.e. $\sum_s z_{l,s} = 1$) for assumption of one occurrence per sequence (OOPS) and 18 scoring constraints. And so the total number of constraints is 76.

General formulation of fixed-pattern TFBS finding

The objective of fixed-pattern TFBS finding is to maximize the total matches among all the sequence, i.e., $\max \sum_l Score_l$. With sequence score defined as $Score_l = \max\left\{\sum_s \left(z_{l,s} \sum_i \theta_{l,s,i}\right)\right\}$, we have the objective function more precisely described as:

$$\max \sum_l \sum_s \left(z_{l,s} \sum_i \theta_{l,s,i} \right). \tag{8}$$

Therefore a mixed 0-1 nonlinear program for fixed-pattern TFBS finding problem can be generally formulated as program (P1).

Mixed 0-1 Nonlinear Program for Fixed-pattern TFBS Finding

Maximize $$\sum_l \sum_s \left(z_{l,s} \sum_i \theta_{l,s,i} \right) \tag{P1}$$

Subject to

$$\left.\begin{aligned} y_{\mathrm{A},i} &= (1-u_i)(1-v_i),\\ y_{\mathrm{T},i} &= u_i v_i,\\ y_{\mathrm{G},i} &= u_i(1-v_i),\\ y_{\mathrm{C},i} &= (1-u_i)v_i, \end{aligned}\right\} \forall i \in \{1,\ldots,M\},$$

$$\theta_{l,s,i} = \begin{cases} y_{\mathrm{A},i} & \text{if } d_{l,s,i} = \mathrm{A},\\ y_{\mathrm{T},i} & \text{if } d_{l,s,i} = \mathrm{T},\\ y_{\mathrm{C},i} & \text{if } d_{l,s,i} = \mathrm{C},\\ y_{\mathrm{G},i} & \text{if } d_{l,s,i} = \mathrm{G}, \end{cases} \quad \forall l \in \{1,\ldots,L\} \quad \forall s \in \aleph \quad \forall i \in \{1,\ldots,M\},$$

$$\sum_s z_{l,s} = 1, \qquad \forall l \in \{1,\ldots,L\},$$

$$u_i, v_i \in \{0,1\}, \quad z_{l,s} \in [0,1],$$
$$b \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}, \quad i \in \{1,\ldots,M\},$$
$$l \in \{1,\ldots,L\}, \quad s \in \aleph.$$

Chapter 2 Propositions

In the previous chapter we formulate a nonlinear program (P1) for fixed-pattern TFBS finding on DNA sequences. Unfortunately, (P1) is very hard to solve with current optimization tools because of numerous binary variables. On the other hand, with natures of nonlinear program containing product terms, (P1) can only obtain a local optimum. These defects quality make (P1) impractical. In this chapter we discuss techniques utilized to make (P1) solvable and even to conduct linearization which can obtain the global optimal solution.

2.1 Relaxation of Binary Indicator *z*

Program (P1) contains many binary variables which make it difficult to solve. The largest part of binary variables is from the indicators *z*. Every candidate site has a binary variable *z* indicating whether it best fits the consensus sequence. For example of finding CRP-binding sites, the formulation as program (P1) will have 1620 *z*'s. This large number of binary variables makes (P1) intractable. A linear relaxation on *z* is applicable to make (P1) solvable. For a TFBS finding problem, this relaxation provides a very tight bound to (P1). In fact, by the following proposition, it is proven having the same optimal value as (P1) has.

PROPOSITION 1 *A selection problem as*

Maximize (or minimize) $\sum_i c_i x_i$

Subject to $\sum_i x_i = 1$,

where c_i be constants and $x_i \in \{0,1\}$.

has a linear relaxation by loosing x_i *to be continuous from* 0 *to* 1 *which shares* *the same optimal value* $\max_i c_i$ *(or* $\min_i c_i$*).* ■

PROOF The proof is trivial. □

For Program (P1), the objective function (8) is separable by sequences, i.e. $\max\sum_l\sum_s\left(z_{l,s}\sum_i\theta_{l,s,i}\right)=\sum_l\left(\max\sum_s\left(z_{l,s}\sum_i\theta_{l,s,i}\right)\right)$. Because only one chosen (u, v) pair is involved in each iteration, $\sum_i\theta_{l,s,i}$ can be regarded to as a constant. With Proposition 1 we can have the result that Program (P1) share the same optimal value, $\max\sum_l\sum_s\left(z_{l,s}\sum_i\theta_{l,s,i}\right)=\sum_l\left(\max\sum_s\left(z_{l,s}\sum_i\theta_{l,s,i}\right)\right)=\sum_l\left(\max\sum_i\theta_{l,s,i}\right)$, with a relaxation where z's are loosen to as continuous variables between 0 and 1. The enormous binary variables are therefore eliminated successfully and Program (P1) becomes manageable.

2.2 Disaggregated Nonlinear Formulation

To obtain the global optimal solution, program (P1) needs reformulated to a mixed 0-1 linear program. Before utilizing the linearization approach proposed in the following section, Program (P1) is firstly transformed to another formulation for effective elimination on all product terms. The formulation underlying the linearization process discussed in this chapter can be viewed as a disaggregated version of Program (P1).

Denote $S_{b,i}$ as the index set of candidate sites having their i^{th} base as nucleotide type b, as defined as follows:

$$S_{b,i}=\{(l,s)\mid d_{l,s,i}=b,b\in\{\mathrm{A,T,G,C}\}\}. \qquad \textbf{(9)}$$

From the definition of comparison function $\theta_{l,s,i}$, it can be restated that

$$\theta_{l,s,i} = y_{b,i} \text{ for all } (l,s) \in S_{b,i} \quad . \tag{10}$$

Then, with (9) and (10) the objective function (8)

$$\max \sum_{l,s} \left(z_{l,s} \sum_{i} \theta_{l,s,i} \right)$$

has an equivalent disaggregated formulation as

$$\max \sum_{b,i} \left(y_{b,i} \sum_{(l,s)\in S_{b,i}} z_{l,s} \right) . \tag{11}$$

Therefore program (P1) is reformulated as program (P1a) shown below:

Disaggregated version of (P1)

Maximize $\sum_{b,i} \left(y_{b,i} \sum_{(l,s)\in S_{b,i}} z_{l,s} \right)$ **(P1a)**

Subject to

$$\left.\begin{aligned} y_{\mathrm{A},i} &= (1-u_i)(1-v_i), \\ y_{\mathrm{T},i} &= u_i v_i, \\ y_{\mathrm{G},i} &= u_i(1-v_i), \\ y_{\mathrm{C},i} &= (1-u_i)v_i, \end{aligned}\right\} \forall i \in \{1,\ldots,10\},$$

$$S_{b,i} = \{(l,s) \mid d_{l,s,i} = b\} \quad \forall b \in \{\mathrm{A,T,G,C}\} \;\; \forall i \in \{1,\ldots,10\},$$

$$\sum_{s} z_{l,s} = 1, \qquad \forall l \in \{1, \ldots, L\},$$

$$\begin{aligned} &u_i, v_i \in \{0, 1\}, \quad z_{l,s} \in [0, 1], \\ &b \in \{\mathrm{A,T,G,C}\}, \quad i \in \{1,\ldots,10\}, \\ &l \in \{1,\ldots,L\}, \quad s \in \aleph. \end{aligned}$$

An important progression from (P1) to (P1a) is elimination of an ambiguous term $\theta_{l,s,i}$. This is very important for further linearization because it makes the product term $z_{l,s}\sum_i \theta_{l,s,i}$ more explicit to eliminate.

2.3 Replacement of Mixed 0-1 Product Terms

Program (P1a) cannot find the global optimum because the product terms contained in the formulation. There are two kinds of product terms, $u_i v_i$ and $y_{b,i}\sum z_{l,s}$, conducting nonlinearity of (P1a). To make the program globally solvable, here we discuss how to eliminate product terms by a series of constraints.

The first kind of product term is $u_i v_i$ which exists in $y_{b,i}$. This product term consists only by binary variable and can be replaced by a continuous variable based on the following proposition.

Proposition 2 *A general binary product* $\alpha\prod_{j=1}^{n} u_j$ *where* $u_j \in \{0,1\}$ *and α is nonzero constant can be replaced by a continuous variable w accompanied with the following bounding constraints:*

(i) $w \le \alpha u_j \quad \forall j,$

(ii) $w \ge 0,$ ■

(iii) $w \ge \alpha\left(\sum_j u_j - n + 1\right).$

Proof Consider $f(u) = \alpha\prod_{j=1}^{n} u_j$. Because product of binary variables is also binary, there are only two possible values for *f*: 0 and *α*. Because *α* is nonzero, the case of *f* = 0 implies $\prod_{j=1}^{n} u_j = 0$ and there must be at least one $u_j = 0$. The bounding constraints (i) and (ii) can make *w* = 0 = *f* when any of $u_j = 0$ and meanwhile constraints (iii) and (iv) are inactive. The other case of $f = \alpha$ implies $\prod_{j=1}^{n} u_j = 1$ and $u_j = 1$ for all *j*. Consider $\alpha\left(\sum_j u_j - n + 1\right)$. If $f = \alpha$ (i.e., all $u_j = 1$) then $\alpha\left(\sum_j u_j - n + 1\right) = \alpha$. If $f = 0$ then

$\alpha\left(\sum_j u_j - n + 1\right) \le 0$. That means with constraints (i) and (ii), constraint (iii) can make $w = \alpha$ when $f = \alpha$ but will become inactive when $f = 0$. Therefore, with constraints (i), (ii) and (iii) the nonnegative variable *w* can completely substitute *f*. □

Therefore, the first kind of product term $u_i v_i$ is a simplified case with $\alpha = 1$ and can be replaced by a continuous variable w_i accompanied with the following constraints

$$\begin{aligned} & w_i \le u_i, \\ & w_i \le v_i, \\ & w_i \ge 0, \\ & w_i \ge u_i + v_i - 1. \end{aligned} \qquad \textbf{(12)}$$

The second kind of product terms to eliminate is $y_{b,i} \sum_{(l,s)\in S_{b,i}} z_{l,s}$. In the relaxation version of Program (P1a), $\sum_{(l,s)\in S_{b,i}} z_{l,s}$ is a continuous variable within [0, *L*] (i.e., *L* is the number of sequences). This kind of mixed 0-1 product terms can be eliminated with Corollary 1.

Corollary 1 *A mixed 0-1 product term* $\alpha x \prod_{j=1}^{n} u_j$ *where* $u_j \in \{0,1\}$, $x \in (0, \varsigma]$ *and α is nonzero constants can be replaced by a continuous variable w accompanied with the following bounding constraints:*

$$\begin{aligned} & \text{(i) } w \le \alpha \varsigma u_j \quad \forall j, \\ & \text{(ii) } w \le \alpha x, \\ & \text{(iii) } w \ge 0, \\ & \text{(iv) } w \ge \alpha\left(x + \varsigma\left(\sum_j u_j - n\right)\right) \end{aligned} \qquad \blacksquare$$

Proof Denote $p(u) = \prod_{j=1}^{n} u_j$ and obviously *p*(*u*) is also binary. From the proof of Proposition 2 we know that when any $u_j = 0$, *p*(*u*) becomes 0 and so does $\alpha x \prod_{j=1}^{n} u_j$. In this condition constraints (i) and (iii) make *w*=0 without violating (ii) and (iv). In the other case of $u_j = 1$ for all *j*, *p*(*u*) becomes 1 and therefore $\alpha x \prod_{j=1}^{n} u_j$ equals *αx*. For this case constraints (ii) and (iv) make *w* tightly bounded to *αx* without violating (i) and (iii). □

In Program (P1a), the second kind of product terms to eliminate is $y_{b,i} \sum z_{l,s}$. The upper bound of $\sum_{(l,s)\in S_{b,i}} z_{l,s}$ is *L*, the number of sequences, because every sequence has only one candidate site to propose, i.e. one occurrence per sequence (OOPS). With Corollary 1, $y_{b,i} \sum z_{l,s}$ can be replaced by a continuous variable $q_{b,i}$ accompanied with following constraints:

$$
\begin{aligned}
& q_{b,i} \le \sum\nolimits_{(l,s)\in S_{b,i}} z_{l,s} \ , \\
& q_{b,i} \le y_{b,i} L \ , \\
& q_{b,i} \ge \sum\nolimits_{(l,s)\in S_{b,i}} z_{l,s} + (y_{b,i} - 1)L \ , \\
& q_{b,i} \ge 0.
\end{aligned}
\qquad \textbf{(13)}
$$

where *L* is the number of sequences.

Therefore, all the product terms in Program (P1a) can be successfully replaced by other single continuous variables and a globally optimal solution is then available for fixed-pattern TFBS finding. In fact, based on the techniques discussed in this chapter, more flexible and complicated TFBS finding problems can also be formulated as mixed 0-1 linear programs. In the following chapters we have a more detailed discussion on these formulations.

Chapter 3 Model 1: Fixed-pattern TFBS Finding

With linearization techniques discussed in Chapter 2, program (P1) can be transformed into a mixed 0-1 linear program which is solvable and promising on global optimum. In this chapter we firstly illustrate the mixed 0-1 linear program for fixed-pattern TFBS finding. Then, more details on TFBS finding is discussed and formulated to appropriated logical constraints which help accuracy. Finally, software designed using the proposed mixed 0-1 linear program is introduced and we discuss on the experimental results about searching for TFBS by this software.

3.1 Mixed 0-1 linear program for Fixed-pattern TFBS Finding

After applying relaxation and linearization discussed in Chapter 2 on (P1), we have (P2), a mixed 0-1 linear program for finding fixed-pattern TFBS. From the nature of binary variable and mixed 0-1 linear program, (P2) has advantages over many current methods:

(i) A globally optimal solution is promised. Because the nonlinear formulation is successfully replaced by a linear relaxation proven exactly to match the original formulation at optimal points, this program can provide globally optimal solution.

(ii) Logical constraints are applicable for better searching quality. With binary variables utilized, structured information profiting accuracy can be formulated as logical constraints like structural constraints and exception rules. Some of

Mixed 0-1 Linear Program for Fixed-pattern TFBS Finding

Maximize $\sum_{b,i} q_{b,i}$ **(P2)**

Subject to

$$\left.\begin{aligned}
&y_{\mathrm{A},i} = 1 - u_i - v_i + w_i,\\
&y_{\mathrm{T},i} = w_i,\\
&y_{\mathrm{G},i} = u_i - w_i,\\
&y_{\mathrm{C},i} = v_i - w_i,\\
&w_i \le u_i\,, \quad w_i \le v_i\,,\\
&w_i \ge 0\,, \quad w_i \ge u_i + v_i - 1,
\end{aligned}\right\} \forall i \in \{1,\ldots,10\},$$

$$\left.\begin{aligned}
&q_{b,i} \le \sum_{(l,s)\in S_{b,i}} z_{l,s}\,,\\
&q_{b,i} \le y_{b,i} L,\\
&q_{b,i} \ge \sum_{(l,s)\in S_{b,i}} z_{l,s} + (y_{b,i} - 1)L,\\
&q_{b,i} \ge 0,\\
&S_{b,i} = \{(l,s) \mid d_{l,s,i} = b\}
\end{aligned}\right\} \forall b \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\} \quad \forall i \in \{1,\ldots,10\},$$

$$\sum_{s} z_{l,s} = 1, \qquad \forall l \in \{1,\ldots,L\},$$

$$\begin{aligned}
&u_i, v_i \in \{0,1\}, \quad z_{l,s} \in [0,1],\\
&b \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}, \quad i \in \{1,\ldots,10\},\\
&l \in \{1,\ldots,L\}, \quad s \in \aleph.
\end{aligned}$$

these constraints, especially structural constraints, can also notably reduce the searching space and computation.

(iii) The program can be extended for more complicated formulation with considerations of practical use. For situations of poor information of target TFBS, e.g. spacer number unknown, this program can still find the TFBS with some modification on the formulation.

(iv) Suboptimal solutions are available by excluding specific solutions. For case of searching for weak signals in DNA sequences, this program can find more than one solution to help explore the correct binding targets with further empirical examinations.

(v) It is very straightforward to find the complete set of the second, third, etc. best consensus sequences.

For utilizing information which helps accuracy, we discuss the formulation of several types of logical constraints in the following sections.

3.2 Structural Constraints

Most TFBS are not only conserved signals but having some specific features reflecting structures of the corresponding regulatory proteins. The proposed mixed 0-1 linear program is convenient for embedding logical constraints to elucidate specific TFBS precisely and efficiently. Structural features of various types of TFBS can be formulated as logical constraints to help facilitate the search process. There are three general types of TFBS: mono-type TFBS like binding sites for homeodomains, dyad-type TFBS like bHLH and bZIP binding sites, and serial-type TFBS like zinc-finger binding sites. To find TFBS with specific structures, program (P2) is further modified with several logical constraints incorporated.

The most often seen TFBS are dyad-type. This is because most gene regulators are dimers or tetramers. This kind of TFBS usually has a length less than or equal to 22 and has two symmetric half parts forming an inverted palindrome or direct repeats. For an inverted palindrome the homologous nucleotide bases are supposed complement, i.e. adenine (A) should be paired with thymine (T) and guanine (G) should be paired with cytosine (C). The logical constraint set describing inverted palindrome, for example of CRP-binding sites, can be formulated as:

$$\begin{aligned} u_i + u_{11-i} &= 1, \\ v_i + v_{11-i} &= 1 . \end{aligned} \qquad \textbf{(14)}$$

Another type of TFBS for binding dimerized protein has direct repeats where the same sequence repeats tandem. The logical constraint set of direct repeat can be formulated as:

$$u_i = u_{5+i} \text{ ,} \quad v_i = v_{5+i} \text{ .} \qquad \textbf{(15)}$$

Obviously, this kind of logical constraints establishes tight relationships between two half sites and prunes a very large portion of searching space. Therefore, applying such a constraint can notably improve both accuracy and computational performance.

3.3 Suboptimal Consensus

The proposed program can find the globally optimal solution. But practically TFBS finding need more than one solution for further verification. This is because there may be more than one kind of regulatory binding sites and the target TFBS may relatively weaker than other signals. To find the suboptimal solutions, we need to embed exception constraints to banish previously obtained solutions and iteratively run the program with these exception constraints.

To exclude one or more solutions previously determined meaningless or not of interest (e.g. *ATGT******ACAT*), a constraint to be involved is as follows:

$$y_{\mathrm{A},2} + y_{\mathrm{T},3} + y_{\mathrm{G},4} + y_{\mathrm{T},5} + y_{\mathrm{A},6} + y_{\mathrm{C},7} + y_{\mathrm{A},8} + y_{\mathrm{T},9} \le 8 - \delta, \qquad \textbf{(16)}$$

where $0 \le \delta \le 8$ is exclusiveness degree.

The exclusiveness degree decides the banishing range. All solutions having 8-δ

matches with “*ATGT******ACAT*” will be filtered. $\delta = 0$ means no exclusion. In this example the right hand side of (16) is set as 8-δ because the number of reactive letters in the excluded solution is 8. Note that this number need not equal to the number of reactive bases.

3.4 General Exception Constraints

A reality among regulatory TFBS is that the background nucleotides are not independently and identically distributed. There are always other noises than the target TFBS in the data. When finding a weak signal, we will need exception constraints to help dig out the target. For instance, the often seen poly-A and poly-T tails should be excluded when searching for a direct repeat.

Solutions to be excluded may come from several sources: meaningless repeats, binding sites for a co-regulator, and regions to form stem-loops in mRNA when searching for binding sites of negative regulators, etc. Two kinds of exception constraints are formulated for different cases of noise source:

Repeats with uncertain length

Repeats of arbitrary length like poly-A tail or poly-T tail are meaningless and should be filtered out. For instance of poly-A tail, the constraint should be formulated as follows to banish all the possible solutions containing too many ‘A’:

$$y_{\mathrm{A},1} + y_{\mathrm{A},2} + y_{\mathrm{A},3} + \ldots + y_{\mathrm{A},10} \le 10 - \delta \cdot \qquad \textbf{(17)}$$

Empirical exception rules

Constraints for excluding a specific set of solutions can also be formulated conveniently. For instance, consensus sequences consisting only of weak bases (A and T) or only of strong bases (C and G) are usually not a regulatory site of concern. If this kind of solutions is not expected, exclusive constraints can be attached as:

$$\sum_{i=1}^{10}\left(u_i + v_i - 2w_i\right) \geq 1 \qquad \text{for all-weak consensus exclusion,} \qquad \textbf{(18)}$$

$$\sum_{i=1}^{10}\left(u_i + v_i - 2w_i\right) \leq 9 \qquad \text{for all-strong consensus exclusion.} \qquad \textbf{(19)}$$

By utilizing binary variables, any if-then rules can also be formulated as logical constraints. These constraints vary by cases and they notably help discriminate weakly conserved TFBS.

3.5 Experimental Results

CRP-binding sites

For the example of finding CRP-binding sites on DNA of *E. Coli*, after solving program (P2) we can obtain the globally optimum solution "TGTGA******TCACA" with objective value 147. The related nonzero $z_{l,s}$ values indicate the starting positions of the binding sites in the 18 sequences, as listed below:

$$\begin{aligned} & z_{1,64} = z_{2,58} = z_{3,79} = z_{4,66} = z_{5,53} = z_{6,63} = z_{7,27} = z_{8,42} = z_{9,12} = z_{10,17} \\ & = z_{11,64} = z_{12,44} = z_{13,51} = z_{14,74} = z_{15,20} = z_{16,56} = z_{17,87} = z_{18,81} = 1. \end{aligned} \qquad \textbf{(20)}$$

Based on the solution, we can also apply an exception constraint to find suboptimal solutions. Program (P2) can find the exact global optimum solution. But in some cases this globally optimal solution may be only an overrepresented but meaningless repeat more conserved than target TFBS. For further discovery of target TFBS, we can apply exclusive constraints to find the suboptimal solutions. For example of CRP-binding sites, the second best solution of (P2) can be obtained by adding a new constraint as

$$y_{\mathrm{T},1} + y_{\mathrm{G},2} + y_{\mathrm{T},3} + y_{\mathrm{G},4} + y_{\mathrm{A},5} + y_{\mathrm{T},6} + y_{\mathrm{C},7} + y_{\mathrm{A},8} + y_{\mathrm{C},9} + y_{\mathrm{A},10} \le 9\,. \quad \textbf{(21)}$$

The new constraint is used to force the program to find a new solution different from the solution of (P2). The found second best consensus sequence is "AAATT******AATTT" with score 129. This is a solution consisted by only weak bases (i.e. A and T), so we can regard it as a meaningless solution. Similarly we can find another solution by adding following constraint.

$$y_{\mathrm{A},1} + y_{\mathrm{A},2} + y_{\mathrm{A},3} + y_{\mathrm{T},4} + y_{\mathrm{T},5} + y_{\mathrm{A},6} + y_{\mathrm{A},7} + y_{\mathrm{T},8} + y_{\mathrm{T},9} + y_{\mathrm{T},10} \le 9\,. \quad \textbf{(22)}$$

The third best solution obtained is "TTTGA******TCAAA" with score 129.

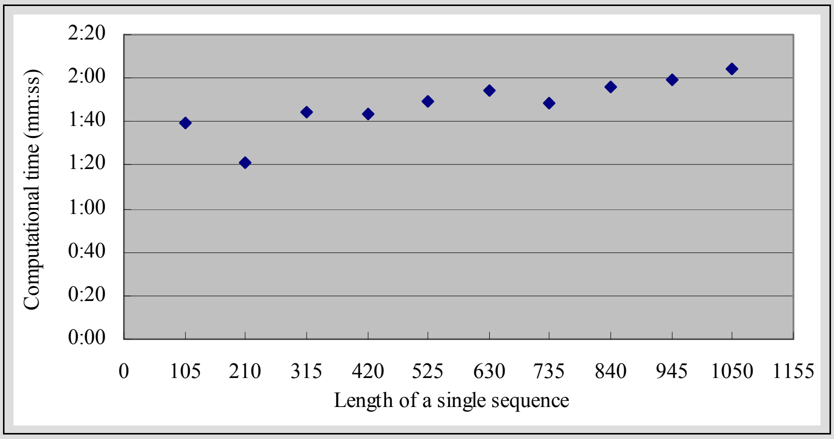
Computational experiments

To analyze the effect of sequence length and number of sequences on the computational time, several experiments are tested using the example of CRP-binding sites. The solving engine for optimization is LINGO (Schrage, 1999), a widely used optimization software, on a personal computer with a Pentium 4 2.0G CPU.

Figure 8 illustrates the experimental results for analyzing the time complexity. Figure 8(a) is the computational time given various sequence lengths, where the number of sequences is fixed at 18. The results show that the computational time changes slightly even if the sequence length is increased from 105 to 1050. Figure 8(b)

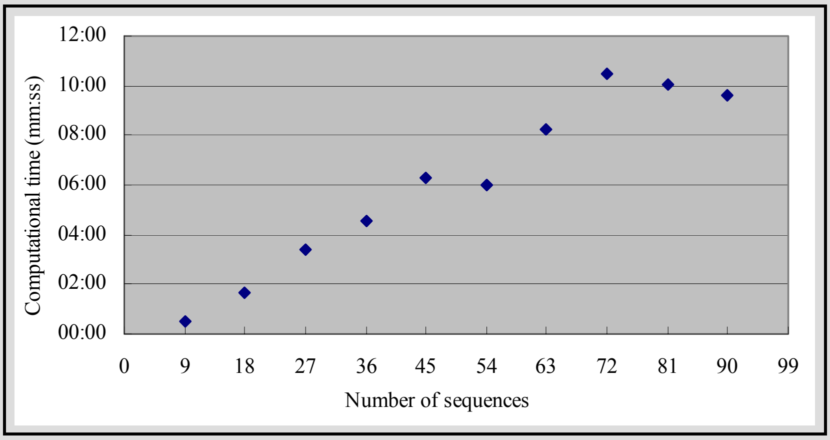
(a) Computational time versus sequence length

Sequence Length	Solving Time (mm:ss)
105	1:39
210	1:21
315	1:44
420	1:43
525	1:48
630	1:54
735	1:48
840	1:56
945	1:59
1050	2:04



(b) Computational time versus number of sequences

Number of Sequences	Solving Time (mm:ss)
9	0:30
18	1:39
27	3:21
36	4:32
45	6:15
54	6:01
63	8:16
72	10:29
81	10:01
90	9:37



(c) Computational time versus number of independent positions

Number of Indep Pos	Solving Time (h:mm:ss)
2	0:00:01
3	0:00:03
4	0:00:21
5	0:01:23
6	0:03:38
7	0:05:18
8	0:08:25
9	0:15:52
10	0:53:27
11	2:33:20

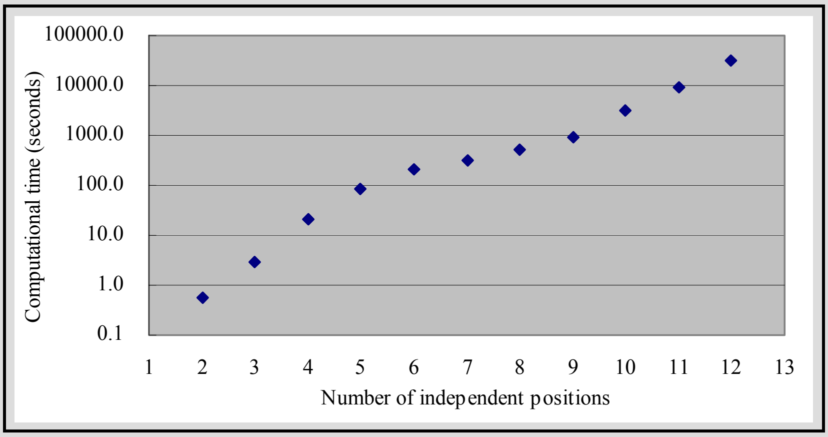


Figure 8 Computational experiments for fixed-pattern TFBS finding. The relationship between computational time and various factors involved in a motif finding problem. This figure illustrates the computational time of solving Program 2 with (a) various sequences sizes; (b) various number of sequences and (c) various independent positions.

is the computational time with various numbers of sequences. It shows that the solving time is roughly proportional to the number of sequences. The proposed model is quite promising for treating the TFBS finding problems with long sequences and a large number of sequences. Figure 8(c) shows that the computational time rises exponentially as the number of independent positions increases.

3.6 Software Package: “Global Site Seer”

A software package named “Global Site Seer” is developed based on program (P2) for solving fixed-pattern TFBS finding problems. This software is available from http://www.iim.nctu.edu.tw/~cjfu/gss.htm

Chapter 4 Model 2: Ambiguous-spacer TFBS Finding

A more complicated TFBS finding problem is to find the consensus sequence in an uncertain pattern format where the number of ignored letters between the two half sites is unknown. In this chapter we introduce a modification of program (P2) to solve this kind of TFBS finding problems.

4.1 Problem of Ambiguous-spacer TFBS Finding

A TFBS finding problem with Ambiguous spacers is defined as follows:

Given

(i) A sequence set containing L sequences co-regulated by a dimerized activator,

(ii) An inverted palindrome shared pattern which has 5 adjacent reactive bases in each half sites but in-between spacers unknown, i.e. "NNNNN*…*NNNNN".

To find the best conserved consensus sequence

$$x_1x_2x_3x_4x_5(k)x_6x_7x_8x_9x_{10},$$

where $x_i \in \{\text{A, T, G, C}\}$, i is the index of reactive base, k is the spacer number to find and comp(·) means a complement base.

Generally the TFBS for binding dimerized regulators have their length not more than 22 bases. This is because of the space restriction on binding α-helices of both protein units to the major groove of DNA double strand helix structure. That means, a reasonable range of spacer numbers is limited from 0 to 12 in the problem defined above. Therefore, the concept of mixed 0-1 linear programming approach for this kind of TFBS finding problem is to enumerate all possible spacer numbers k and reformulate program (P2) to cover these enumerations.

(a)

Sequence #1: AAGACTGTTTTTTTGATC

Sequence #2: …

(b)

$$D_0 = \{(l,s,k) \mid k = 0\},$$

$(l, s, k) = (1,1,0)$ $\boxed{\texttt{AAGACTGTTT}}\texttt{TTTTGATC}$

$(l, s, k) = (1,2,0)$ $\texttt{A}\boxed{\texttt{AGACTGTTTT}}\texttt{TTTGATC}$

… …

$$D_1 = \{(l,s,k) \mid k = 1\},$$

$(l, s, k) = (1,1,1)$ $\boxed{\texttt{AAGAC}}\texttt{T}\boxed{\texttt{GTTTT}}\texttt{TTTGATC}$

$(l, s, k) = (1,2,1)$ $\texttt{A}\boxed{\texttt{AGACT}}\texttt{G}\boxed{\texttt{TTTTT}}\texttt{TTGATC}$

… …

$$D_2 = \{(l,s,k) \mid k = 2\},$$

$(l, s, k) = (1,1,2)$ $\boxed{\texttt{AAGAC}}\texttt{TG}\boxed{\texttt{TTTTT}}\texttt{TTGATC}$

$(l, s, k) = (1,2,2)$ $\texttt{A}\boxed{\texttt{AGACT}}\texttt{GT}\boxed{\texttt{TTTTT}}\texttt{TGATC}$

… …

Figure 9 Site extraction for ambiguous-spacer TFBS finding: (a) original sequence data; (b) schematic representation of the candidate sites.

The data preparation step is similar to the preparation procedure of fixed-pattern TFBS finding. To enumerate all possible *k*, a candidate set $D = \{D_0, D_1, ..., D_k, ..., D_{10}\}$, where D_k is constructed as a fixed-pattern candidate set, is prepared for different *k* from 0 to 10. The candidate sites are thus indexed by (*l*,*s*,*k*). A simple illustration of constructing *D* is shown in Figure 9. And therefore we can redefine the index set $S_{b,i}$ as

$$S_{b,i} = \{(l,s,k) \mid d_{l,s,i}^{k} = b\}, \qquad \textbf{(23)}$$

where $d_{l,s,i}^{k}$ is the i^{th} base of a candidate site contained by D_k.

4.2 Mixed 0-1 linear program for Ambiguous-spacer TFBS Finding

To solve a TFBS finding problem with an ambiguous spacer number, we need to apply some modification on program (P2) to enumerate different *k*. Because the target consensus sequence is fixed on its spacer number, we need to find *k* with the maximum matching score. With the assumption of OOPS where each sequence has only the best fitting candidate site proposed, the conservation constraints of site indicator $z_{l,s,k}$ is reformulated as follows:

$$\sum_{s,k} z_{l,s,k} = 1 \qquad \forall l \in \{1, ..., L\}, \tag{24}$$

$$\sum_{s} z_{1,s,k} = \sum_{s} z_{2,s,k} = ... = \sum_{s} z_{L,s,k} \quad \forall k \in \{0, ..., 10\}. \tag{25}$$

Constraint set (24) is a modification of (5) which is based on the assumption of OOPS: only one $z_{l,s,k}$ is supposed to be nonnegative in a sequence. Constraint (25) is used to make sure that all the nonzero $z_{l,s,k}$ have their corresponding candidate sites from the same set D_k. By applying these two constraints we can then obtain a solution with a fixed spacer number.

Because this kind of TFBS finding problem is only for analyzing dimerized activator, structural constraint of inverted palindrome must be incorporated as

$$u_i + u_{11-i} = 1,$$
$$v_i + v_{11-i} = 1.$$

Mixed 0-1 Linear Program for TFBS finding with ambiguous spacers

$$\text{Maximize} \quad \sum_{b,i} q_{b,i} \qquad \textbf{(P3)}$$

$$\text{Subject to} \quad \left.\begin{aligned}
& y_{\mathrm{A},i} = 1 - u_i - v_i + w_i, \\
& y_{\mathrm{T},i} = w_i, \\
& y_{\mathrm{G},i} = u_i - w_i, \\
& y_{\mathrm{C},i} = v_i - w_i, \\
& w_i \le u_i, \quad w_i \le v_i, \\
& w_i \ge 0, \quad w_i \ge u_i + v_i - 1, \\
& u_i + u_{11-i} = 1, \\
& v_i + v_{11-i} = 1,
\end{aligned}\right\} \forall i \in \{1,\ldots,10\},$$

$$\left.\begin{aligned}
& q_{b,i} \le \sum_{(l,s,k)\in S_{b,i}} z_{l,s,k}, \\
& q_{b,i} \le y_{b,i} L, \\
& q_{b,i} \ge \sum_{(l,s,k)\in S_{b,i}} z_{l,s,k} + (y_{b,i} - 1)L, \\
& q_{b,i} \ge 0, \\
& S_{b,i} = \{(l,s,k) \mid d^{k}_{l,s,i} = b\}
\end{aligned}\right\} \forall b \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\} \;\; \forall i \in \{1,\ldots,10\},$$

$$\begin{aligned}
& \sum_{s,k} z_{l,s,k} = 1 \qquad \forall l \in \{1, \ldots, L\}, \\
& \sum_{s} z_{1,s,k} = \sum_{s} z_{2,s,k} = \ldots = \sum_{s} z_{L,s,k} \quad \forall k \in \{0,\ldots,10\}, \\
& u_i, v_i \in \{0, 1\}, \quad z_{l,s} \in [0, 1], \\
& b \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}, \quad i \in \{1,\ldots,10\}, \\
& l \in \{1,\ldots,L\}, \quad s \in \aleph, \quad k \in \{0,\ldots,10\}.
\end{aligned}$$

Most of the exception rules for this problem are the same as those discussed in the fixed-pattern motif finding case except the constraints for excluding solutions. The modified constraint is as follows, for example of excluding “*ATGT******ACAT*”:

$$\begin{aligned}
& y_{\mathrm{A},2} + y_{\mathrm{T},3} + y_{\mathrm{G},4} + y_{\mathrm{T},5} + y_{\mathrm{A},6} + y_{\mathrm{C},7} + y_{\mathrm{A},8} + y_{\mathrm{T},9} \\
& + \left(\sum\nolimits_l \sum\nolimits_s z_{l,s,k^*} - L\right) M \le 8 - \delta,
\end{aligned} \qquad \textbf{(26)}$$

where $k^* = 6$ is the spacer number of the excluded solution and $M \geq 8$.

The addendum term $\left(\sum_l \sum_s z_{l,s,k^*} - L\right)M$ lets constraints (26) exclude only the solution "ATGT(k^*)ACAT" with a specified spacer number $k^* = 6$. If all the nonzero $z_{l,s,k}$ are not from D_{k^*}, the constraints will become inactive.

After applying these constraints we have program (P3), the mixed 0-1 linear program for ambiguous-spacer TFBS finding.

4.3 Experimental Results

CRP-binding sites

Using program (P3) to search for CRP binding sites we obtain the globally optimal solution as "TGTGA******TCACA" with score 147, which is exactly the solution found in program (P2). And the second best solution is "GTGAA****TTCAC" with score 134. The relationship between the computational time and the number of possible *k*'s (i.e. |*k*|) is linear, as shown in the experiment result listed in Figure 10. The number of ignored letter k is between 0 and $\bar{k}$, the upper bound of *k*, and thus we have |*k*| = $\bar{k}$ + 1 in this experiment.

FNR-binding sites

Program (P3) is also applied to solve an example of searching for binding sites of *fumarate and nitrate reduction* (FNR) regulatory protein in *E. coli*. Both CRP and FNR belong to the CRP/FNR helix-turn-helix transcription factor superfamily (Tan et al., 2001). The sequence data, which is taken from GenBank, contains 12 DNA sequences with lengths varied from 96 to 781. Owing to the dimer structure of the binding protein, the consensus sequence in this example also has a constraint of

(a)

$\bar{k}$	$\lvert k \rvert$	Consensus sequence	Score	Computational Time
0	1	**TGTTT(0)AAACA**	126	4:51
2	3	**TGAAA(2)TTTCA**	129	12:32
4	5	**GTGAA(4)TTCAC**	134	19:46
6	7	**TGTGA(6)TCACA**	147	24:28
8	9	**TGTGA(6)TCACA**	147	25:49
10	11	**TGTGA(6)TCACA**	147	32:35

(b)

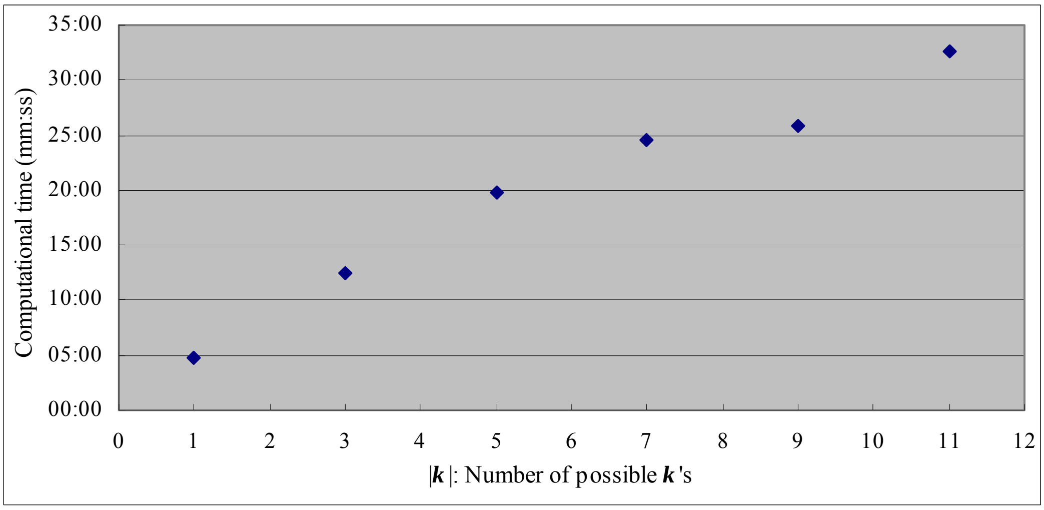


Figure 10 Computational experiments of ambiguous-spacer TFBS finding using program (P3) with various numbers of possible k's (the example in Appendix): **(a)** Solutions of various upper bound of spacer numbers and their corresponding computational time; **(b)** Illustrative plot of relationship between $|k|$ and computational time. The number enclosed in the consensus sequence is the spacer number k^*.

inverse symmetry. The RegulonDB database (Huerta et al., 1998) lists the found regulatory binding sites for eight of these twelve sequences while the exact positions of other four sequences are not listed yet. Solving this example by program (P3) we obtained the global optimal consensus sequence as "TTGAT****ATCAA" with score 107, which is the same consensus sequence as indicated by Tan et al. (2001). Table 4 illustrates the result including the consensus sequence and the predicted binding sites for all of the 12 sequences. Some sites downstream of the transcription start (i.e. with positive indices) are also listed because there are a few known cases in which regulatory sites appear within transcription units (Tan et al., 2001). The proposed method has found some sites not listed in RegulonDB but having scores higher than those listed in RegulonDB (e.g. the third solution in the Operon *ansB* row of Table 4). The best predicted sites in the four undetermined sequences are also listed in Table 4.

Table 4 FNR binding sites found by program (P3)

Operon	Seq. length	Site seq. found by program (P3)	Predicted Position	Score	Site seq. listed in RegulonDB*	Center Position
Consensus:		**TTGAT----ATCAA**				
narK	338	ATGAT----ATCAA	-86	9	actatgGGTAATGATAA**A**TATCAATGATagataa	-79.5
		TTGAT----ATCAA	-48	10	atcttaTCGTTTGATTT**A**CATCAAATTGcctta	-41.5
ansB	345	TTGTT----GTCAA	-48	8	acgttgTAAATTGTTTA**A**CGTCAAATTTcccata	-41.5
		TTGTA----TCCAA	-81	6	gcctctAACTTTGTAGA**T**CTCCAAAATAtattca	-74.5
		TTTAT----TTTAA	-123	7		
narG	525	TTGAT----ATCAA	-55	10	ctcttgATCGTTATCAA**T**TCCCACGCTGtttcag	-41.5
dmsA	325	TTGAT----AACAA	-48	9	ctttgaTACCGAACAAT**A**ATTACTCCTCacttac	-33
frd	781	TTCAG----ATCCA	-37	7	AAAAATCGATC**T**CGTCAAATTTcagacttatcca	-47
		TTAAT----TTCAG	-98	7		
nirB	262	TTGAT----ATCAA	-48	10	aaaggtGAATTTGATTT**A**CATCAATAAGcggggt	-41.5
sodA	284	TTGAT----ATTTT	-42	7	agtacgGCATTGATAAT**C**ATTTTCAATAtcattt	-34
fnr**	96	TTGAC----ATCAA	-7	9	atgttaAAATTGACAAA**T**ATCAATTACGgcttga	1
					ccttaaCAACTTAAGGG**T**TTTCAAATAGatagac	-103.5
(cyoA)	599	CTTCT----ATCAA	-113	7	N/A	N/A
		TTGTT----TTCAC	-198	7		
(icdA)	290	ATGAC----AACAA	16	7	N/A	N/A
		TTGCT----AGCAT	73	7		
(sdhC)	708	TTGAT----AATAA	-330	8	N/A	N/A
(ulaA)	346	TCAAT----ATCAA	-278	8	N/A	N/A
		TTGGT----ATTAA	-257	8		

* For visualizing the comparison, the letters in uppercase represent the binding site listed in RegulonDB; the letter in bold face is the center of the site sequence; and the encompassed letters represent the exact binding site obtained by program (P3).

** The second site listed in RegulonDB is not contained in the sequence data, which is only 96 bases long, from GenBank.

Chapter 5 Model 3: Pattern-free TFBS Finding

In the previous chapters we have discussed mixed 0-1 linear programming approaches for finding TFBS with a given pattern. This pattern may be definitely given or defined with ambiguity on center spacers. In this chapter we discuss a mixed 0-1 linear programming formulation of TFBS finding without any predefined share patterns.

A predefined shared pattern can notably help discriminate the TFBS when applying consensus based TFBS finding approaches. Unfortunately, such information is unavailable in most cases, especially when analyzing an unknown functional regulation. Finding unframed TFBS today still relies on heuristic approaches which compromise to accuracy. This is because that exact enumeration approaches which test all 4^M M-wide patterns are very time consuming and only capable of searching very short patterns. In this chapter a mixed 0-1 linear program for finding unframed binding sites is introduced. This approach can exactly find the best conserved signals in acceptable computational time without any predefined shared pattern.

5.1 Problem of Pattern-free TFBS Finding

Given

(i) A set of DNA sequences each of which contains at least one motif of a specific regulator,

(ii) M, the length of consensus sequence,

(iii) K, the number of reactive bases,

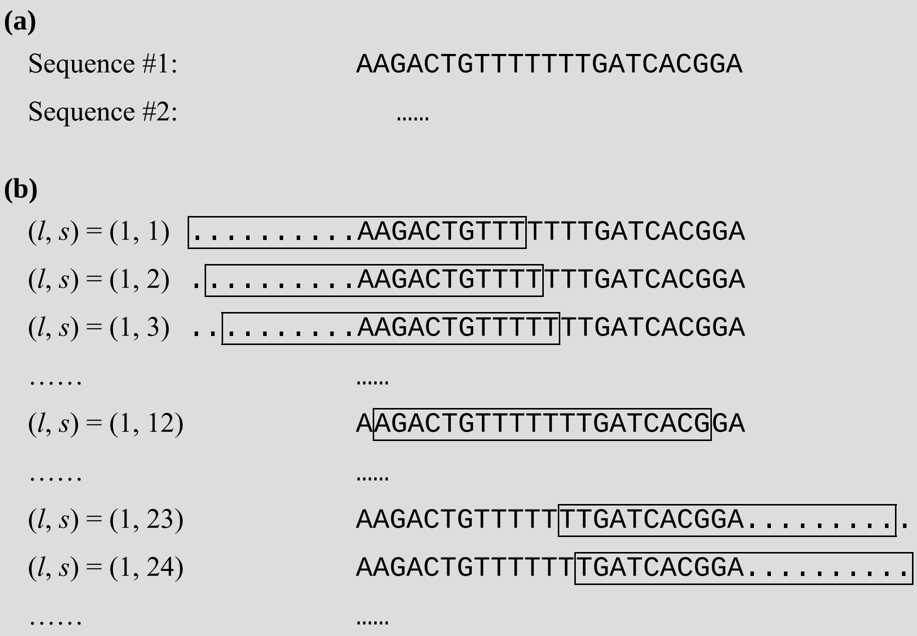
To find the best conserved consensus sequence

$x_1 x_2 x_3 \dots x_{M-1} x_M$,

where $x_i \in \{ \mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}, * \}$ in which '*' means an inactive base.

To find TFBS without any given shared pattern, a better idea than enumerating all 4^M *M*-wide consensus sequences is to enumerate all reasonable combination of reactive bases on the consensus sequence. Based on this concept, only two parameters, the length of regulatory region, *M*, and the number of reactive bases, *K*, are required for TFBS finding. A pattern-free TFBS finding problem can then be defined as follows:

In this definition only K bases in the consensus sequence are reactive, although their positions in the consensus are unknown. Therefore x_i's in this definition have five alternatives including 4 nucleotide types and an inactive type (i.e. '*'). The candidate set to prepare is simpler than those to construct in previous chapters. As shown in Figure 11, all candidate sites are extracted with a given length *M*. A difference on site extraction is that candidate sites are indexed by positions of their center bases. That means the first few and the last few sites will contain some virtual meaningless bases (i.e. bases represented by '.'). This prevents from a case ignoring short signals which locate at the beginning or ending regions.



```
(a)
  Sequence #1:              AAGACTGTTTTTTTGATCACGGA
  Sequence #2:                 ......

(b)
  (l, s) = (1, 1)   ..........AAGACTGTTTTTTTGATCACGGA
  (l, s) = (1, 2)   ..........AAGACTGTTTTTTTGATCACGGA
  (l, s) = (1, 3)   ..........AAGACTGTTTTTTTGATCACGGA
  ......                    ......
  (l, s) = (1, 12)          AAGACTGTTTTTTTGATCACGGA
  ......                    ......
  (l, s) = (1, 23)          AAGACTGTTTTTTTGATCACGGA..........
  (l, s) = (1, 24)          AAGACTGTTTTTTTGATCACGGA..........
  ......                    ......
```

Figure 11 Site extraction for pattern-free TFBS finding ($M = 20$): **(a)** original sequence data; **(b)** schematic representation of the candidate sites.

5.2 Formulation and Linearization

For presentation of the consensus sequence, each base x_i is represented by three binary variables, u_i, v_i and e_i, with a relation shown in Table 5. u_i and v_i decide the nucleotide type and e_i decides sensitivity as follows:

$$e_i = \begin{cases} 1, \text{ if } i^{\text{th}} \text{ base is reactive} \\ 0, \text{otherwise.} \end{cases} \qquad \textbf{(27)}$$

Since the number of reactive bases in a consensus is K, we have

$$\sum_i e_i = K \, . \qquad \textbf{(28)}$$

To determine base matching between a consensus and a candidate site, a series of comparing functions $y_{b,i}$ are defined as

$$\begin{aligned} y_{\mathrm{A},i} &= e_i(1-u_i)(1-v_i), \\ y_{\mathrm{T},i} &= e_i u_i v_i, \\ y_{\mathrm{G},i} &= e_i(1-u_i)v_i, \\ y_{\mathrm{C},i} &= e_i u_i(1-v_i). \end{aligned} \tag{29}$$

For base comparison we use the same notation $\theta_{l,s,i}$ as defined in (6). That is, $\theta_{l,s,i} = y_{b,i}$ if $d_{l,s,i} = b$. A candidate site is evaluated by summing up $\theta_{l,s,i}$'s. Take a candidate site "CGGTCAG" for example, the site score (i.e. number of base matches) is obtained from $\sum_{i=1}^{7}\theta_{l,s,i}$ = $(y_{\mathrm{C},1} + y_{\mathrm{G},2} + y_{\mathrm{G},3} + y_{\mathrm{T},4} + y_{\mathrm{C},5} + y_{\mathrm{A},6} + y_{\mathrm{G},7})$. When comparing with a consensus "CTG*CAG" (*M* = 7 and *K* = 6), this score function will give 5 as the site score for five matched letters.

As defined in Chapter 2, binary variable $z_{l,s}$ is used in (4) to flag the candidate site at (*l*,*s*). The best matching site in a sequence will have its corresponding $z_{l,s}$ be 1 whereas other $z_{l,s}$s be 0. According to the assumption of one occurrence per sequence (OOPS), the same conservation constraint (5) applied in (P1) is also involved as

Table 5 Binary base codes for pattern-free consensus.

x_i	u_i	v_i	e_i
A	0	0	1
T	1	1	1
G	1	0	1
C	0	1	1
*	0	0	0

Mixed 0-1 Nonlinear Program for Pattern-free Motif Finding

Maximize $$\sum_{i=1}^{M}\sum_{b}\left(y_{b,i} \sum_{(l,s)\in S_{b,i}} z_{l,s} \right) \qquad \textbf{(P4)}$$

Subject to
$$\begin{aligned}
& y_{\mathrm{A},i} = e_i(1-u_i)(1-v_i), \\
& y_{\mathrm{T},i} = e_i u_i v_i, \\
& y_{\mathrm{G},i} = e_i(1-u_i)v_i, \\
& y_{\mathrm{C},i} = e_i u_i (1-v_i), \\
& \sum_{i=1}^{M} e_i = K, \\
& \sum_{s} z_{l,s} = 1 \quad \forall l \in \{1,\ldots,L\}, \\
& e_i, u_i, v_i \in \{0,1\}, \quad z_{l,s} \in [0,1], \\
& b \in \{\mathrm{A,T,G,C}\}, \quad i \in \{1,\ldots,M\}, \\
& l \in \{1,\ldots,L\}, \quad s \in \aleph.
\end{aligned}$$

$$\sum_{s} z_{l,s} = 1, \quad z_{l,s} \in \{0,1\} \text{ for all } l. \qquad \textbf{(3)}$$

For a given consensus sequence (i.e. a sequence of (u_i, v_i, e_i)), the best candidate sites of sequences are extracted and sum scored to obtain the total matches by following formula:

$$Score = \sum_{l}\sum_{s}\left(z_{l,s} \sum_{i} \theta_{l,s,i} \right). \qquad \textbf{(30)}$$

A higher score means the consensus found is better conserved among all the sequences. Similarly, the objective function has an equivalent disaggregated form as

$$\sum_{i=1}^{M}\sum_{b}\left(y_{b,i} \sum_{(l,s)\in S_{b,i}} z_{l,s} \right).$$

A pattern-free TFBS finding problem can then be formulated as program (P4).

The objective is to find the consensus having the highest score among the search space of patterns under a given *M* and *K*.

Program (P4) has similar product terms like those in program (P1) which bring about nonlinearity. The product terms we encountered here include: $e_i u_i$, $e_i v_i$, $e_i u_i v_i$ and $y_{b,i} \sum z_{l,s}$. With the relaxation and linearization approaches discussed in Chapter 2, these product terms can also be completely linearized.

For product terms $e_i u_i$ and $e_i v_i$ the linearization can be simplified without loss of generality. A good idea of reducing searching range of binary variables is to add two constraints which make $u_i = 0$ and $v_i = 0$ when $e_i =0$:

$$u_i \le e_i \,, \qquad v_i \le e_i \,. \tag{31}$$

From this dependency we can induce two logical relations:

$$e_i u_i = u_i \,, \qquad e_i v_i = v_i \,. \tag{32}$$

According to Proposition 2 in Chapter 2, the binary product term $e_i u_i v_i$ can be replaced by a new continuous variable w_i accompanied with the following constraints:

$$\begin{aligned} & w_i \le u_i \,, \quad w_i \le v_i \,, \quad w_i \le e_i \,, \\ & w_i \ge 0 \,, \quad w_i \ge u_i + v_i + e_i - 2. \end{aligned} \tag{33}$$

Therefore the comparing functions $y_{b,i}$ have alternative definitions as follows:

$$\begin{aligned}
&y_{\mathrm{A},i} = e_i - u_i - v_i + w_i\,, \quad y_{\mathrm{T},i} = w_i\,,\\
&y_{\mathrm{G},i} = u_i - w_i\,, \quad y_{\mathrm{C},i} = v_i - w_i\,,\\
&u_i \le e_i\,, \quad v_i \le e_i\,,\\
&w_i \le u_i\,, \quad w_i \le v_i\,, \quad w_i \le e_i\,,\\
&w_i \ge 0\,, \quad w_i \ge u_i + v_i + e_i - 2.
\end{aligned} \tag{34}$$

Linearization of the last product term $y_{b,i}\sum z_{l,s}$ is quite similar to those terms in program (P2). That is, a continuous variable $q_{b,i}$ accompanied with following constraints can replace $y_{b,i}\sum z_{l,s}$.

$$\begin{aligned}
&q_{b,i} \le \sum_{(l,s)\in S_{b,i}} z_{l,s}\,,\\
&q_{b,i} \le y_{b,i}L,\\
&q_{b,i} \ge \sum_{(l,s)\in S_{b,i}} z_{l,s} + (y_{b,i} - 1)L,\\
&q_{b,i} \ge 0.
\end{aligned} \tag{10}$$

Therefore all the product terms in program (P4) are successfully linearized. After applying the relaxation of indicator $z_{l,s}$ and linearization of all product terms we can have program (P5), a mixed 0-1 linear program for pattern free TFBS finding.

5.3 Structural Constraints

Most TFBS have some specific features reflecting structures of the corresponding regulatory proteins. The proposed mixed 0-1 linear program is convenient to embed logical constraints for elucidating specific TFBS precisely and efficiently. Structural features of various types of TFBS can be formulated as logical constraints to help facilitate the search process. There are three general types of activators: mono-type TFBS like binding sites for homeodomains, dyad-type TFBS like bHLH and bZIP binding sites, and serial-type TFBS like zinc-finger binding sites.

Mixed 0-1 Linear Program for Pattern-free Motif Finding

Maximize $\sum_{i=1}^{M}\sum_{b} q_{b,i}$ **(P5)**

Subject to

$$
\begin{aligned}
& y_{\mathrm{A},i} = e_i - u_i - v_i + w_i\,, \quad y_{\mathrm{T},i} = w_i\,, \\
& y_{\mathrm{G},i} = u_i - w_i\,, \quad y_{\mathrm{C},i} = v_i - w_i\,, \\
& u_i \le e_i\,, \quad v_i \le e_i\,, \\
& w_i \le u_i\,, \quad w_i \le v_i\,, \quad w_i \le e_i\,, \\
& w_i \ge 0\,, \quad w_i \ge u_i + v_i + e_i - 2, \\
& \sum_{i=1}^{M} e_i = K, \\
& q_{b,i} \le \sum_{(l,s)\in S_{b,i}} z_{l,s}\,, \\
& q_{b,i} \le y_{b,i} L, \\
& q_{b,i} \ge \sum_{(l,s)\in S_{b,i}} z_{l,s} + (y_{b,i} - 1)L, \\
& q_{b,i} \ge 0, \\
& \sum_{s} z_{l,s} = 1 \qquad \forall l \in \{1,\ldots,L\}, \\
& e_i, u_i, v_i \in \{0,1\}, \quad z_{l,s} \in [0,1], \\
& b \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}, \quad i \in \{1,\ldots,M\}, \\
& l \in \{1,\ldots,L\}, \quad s \in \aleph.
\end{aligned}
$$

To find TFBS with specific structure, (P5) is further modified with several logical constraints applied.

Mono-type TFBS

In (P5), a consensus is scored by calculating matches with all proposed candidate sites. Nevertheless, some TFBS may occur in form of inverted complement. For example of a consensus “CACTCA”, the TFBS resembling to the inv/comp “TGAGCG” should also be considered when scoring the consensus. Scoring an inv/comp consensus is the same as to compare original consensus with a set of inv/comp candidate sites. For testing inverted complement simultaneously, another

```
Sequence      TAATGTTTGACAGTGCAACTGTGG
Candidate Set 0 (Original):
...           ...
s = 7         TAATGTTTGACA
s = 8          AATGTTTGACAG
s = 9           ATGTTTGACAGT
...              ...
Candidate Set 1 (Inv/comp):
...           ...
s = 7         TGTCAAACATTA
s = 8          CTGTCAAACATT
s = 9           ACTGTCAAACAT
...              ...
```

Figure 12 Extraction of two candidate sets for original and inv/comp motifs. The homologous candidate sites in different sets are inverse complement.

candidate set consisting of inv/comp candidate sites is involved. A new index t is introduced to distinguish the two sets of candidate sites: $t = 0$ for original set and $t = 1$ for inv/comp set. The index of candidate sites becomes (l, s, t) instead of (l, s) in program (P2). Variable $z_{l,s}$ is replaced by $z_{l,s,t}$ and $S'_{b,i}$, the replacement for $S_{b,i}$, is defined as:

$$S'_{b,i} = \{(l,s,0) \mid d_{l,s,i} = b\} \cup \{(l,s,1) \mid d_{l,s,M-i+1} = \text{comp}(b)\}, \qquad \textbf{(35)}$$

where comp(·) means a complement base.

And $q_{b,i}$, the replacement of product term $y_{b,i} \sum z_{l,s,t}$, is accompanied with constraints shown below:

$$\begin{aligned} & q_{b,i} \le \sum_{(l,s,t)\in S'_{b,i}} z_{l,s,t}, \\ & q_{b,i} \le L y_{b,i}, \\ & q_{b,i} \ge \sum_{(l,s,t)\in S'_{b,i}} z_{l,s,t} + L(y_{b,i} - 1), \\ & q_{b,i} \ge 0. \end{aligned} \qquad \textbf{(36)}$$

Mono-type Motif Finding

Maximize $\sum_{i=1}^{M}\sum_{b} q_{b,i}$ **(P6)**

Subject to

$$
\begin{aligned}
& y_{\mathrm{A},i} = e_i - u_i - v_i + w_i\ , \quad y_{\mathrm{T},i} = w_i\ , \\
& y_{\mathrm{G},i} = u_i - w_i\ , \quad y_{\mathrm{C},i} = v_i - w_i\ , \\
& u_i \le e_i\ , \quad v_i \le e_i\ , \\
& w_i \le u_i\ , \quad w_i \le v_i\ , \quad w_i \le e_i\ , \\
& w_i \ge 0\ , \quad w_i \ge u_i + v_i + e_i - 2, \\
& \sum_{i=1}^{M} e_i = K, \\
& q_{b,i} \le \sum_{(l,s,t)\in S_{b,i}} z_{l,s,t}\ , \\
& q_{b,i} \le y_{b,i} L, \\
& q_{b,i} \ge \sum_{(l,s,t)\in S_{b,i}} z_{l,s,t} + (y_{b,i} - 1)L, \\
& q_{b,i} \ge 0, \\
& \sum_{s,t} z_{l,s,t} = 1 \qquad \forall l \in \{1,\ldots,L\}, \\
& S'_{b,i} = \{(l,s,0) \mid d_{l,s,i} = b\} \cup \{(l,s,1) \mid d_{l,s,M-i+1} = \mathrm{comp}(b)\} \\
& \text{where comp}(\cdot) \text{ is complement operation,} \\
& e_i, u_i, v_i \in \{0, 1\}, \quad z_{l,s} \in [0, 1], \\
& b \in \{\mathrm{A, T, G, C}\}, \quad i \in \{1,\ldots,M\}, \\
& l \in \{1,\ldots,L\}, \quad s \in \aleph, \quad t \in \{0, 1\}.
\end{aligned}
$$

No matter from which candidate set, there is only one site proposed to match the consensus for every sequence. Thus constraints for $z_{l,s,t}$ are placed as follows:

$$\sum_{s,t} z_{l,s,t} = 1 \text{ for all } l\ , \quad 0 \le z_{l,s,t} \le 1\cdot \qquad \textbf{(37)}$$

And therefore we have program (P6) as the mixed 0-1 linear program for searching mono-type TFBS.

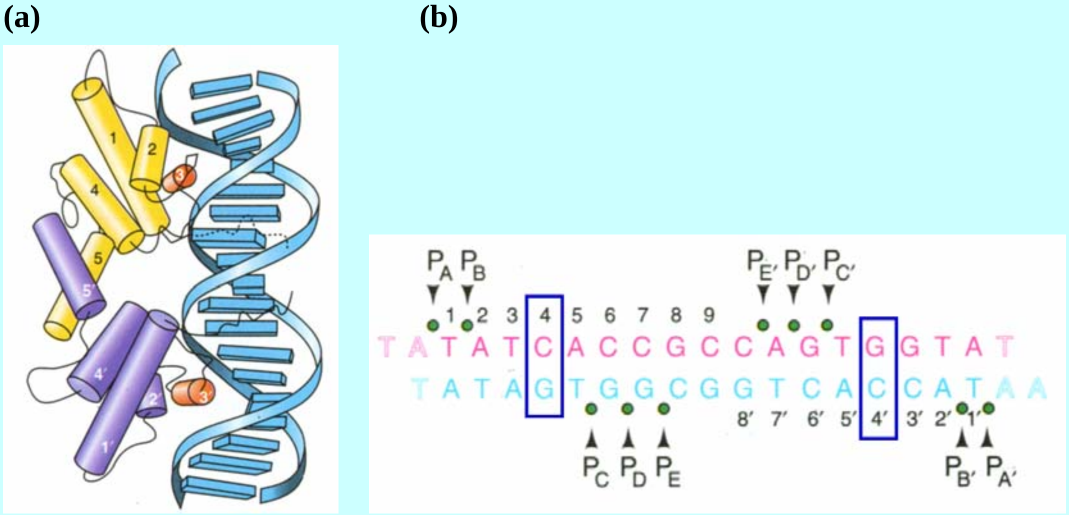


Figure 13 λ repressor, an example of dimerized binding protein. **(a)** Geometry of the λ repressor-operator complex. **(b)** The operator fragment. This 20-mer contains two λ *O*L1 half-sites, each of which binds a monomer of repressor. PA-PE are phosphate groups (backbone) important for recognition. Base pair 4 (Guanine) is also regarded as a reactive base with which Ser 45 of λ repressor makes a hydrogen bond.

Dyad-type TFBS

Most gene regulators are dimers or tetramers. The binding sites of this kind of regulators usually have length less/equal to 22 and have two symmetric half parts forming an inverted palindrome or direct repeats. To find this kind of TFBS, the consensus sequence will be like

$$x_1 x_2 \ldots x_M x_{M+1} \ldots x_{2M-1} x_{2M}, \tag{38}$$

where $x_i \in \{\mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}, *\}$ in which '*' means an inactive base.

The homologous bases in both half sites have the same sensitivity. Referring to constraint (31), both an inactive base and its homologous base have $u_i = v_i = 0$ since $e_i = 0$. That means inactive bases should be free from invert complement relations. The logical constraints for inverted palindrome and direct repeats are then formulated as

```
Sequence      TAATGTTTGACAGTGCAACTGTGG
Candidate Set 0 (even spacers, M=6):
……            ……
s = 7         TAATGTTTGACA
s = 8          AATGTTTGACAG
s = 9           ATGTTTGACAGT
……              ……
Candidate Set 1 (odd spacers, M=6):
……            ……
s = 7         TAATGT.TGACAG
s = 8          AATGTT.GACAGT
s = 9           ATGTTT.ACAGTG
……                ……
```

Figure 14 Extraction of two candidate sets for even-spacer and odd-spacer dyad motifs.

$$\begin{cases} u_i + u_{2M-i+1} = e_i, \\ v_i + v_{2M-i+1} = e_i, \\ e_i = e_{2M-i+1}. \end{cases} \quad \text{for inverted palindrome;} \qquad \textbf{(39)}$$

$$\begin{cases} u_i = u_{M+i}, \\ v_i = v_{M+i}, \\ e_i = e_{M+i}. \end{cases} \quad \text{for direct repeat.} \qquad \textbf{(40)}$$

Many dyad-type TFBS have spacers (i.e. inactive bases) between the two half sites. Because the two half sites are assumed connected in (38), there can only be even number of center spacers when searching an inverted palindrome. program (P5) needs modified to involve both even-spacer and odd-spacer cases. For testing odd-spacer solutions simultaneously, another candidate set is constructed by extracting substrings with a center position skipped, as illustrated in Figure 14. Both the two candidate sets are put into the same program and distinguished by a new index *p*, where *p* = 0 for even-spacer set and p = 1 for odd-spacer set. The index of candidate sites becomes (*l*, *s*, *p*). Variable $z_{l,s}$ is replaced by $z_{l,s,p}$ and $S''_{b,i}$, the replacement for $S_{b,i}$, is defined as:

Dyad-type Motif Finding

Maximize $\sum_{i=1}^{M}\sum_{b} q_{b,i}$ **(P7)**

Subject to
$$y_{\mathrm{A},i} = e_i - u_i - v_i + w_i\ ,\quad y_{\mathrm{T},i} = w_i\ ,$$
$$y_{\mathrm{G},i} = u_i - w_i\ ,\quad y_{\mathrm{C},i} = v_i - w_i\ ,$$
$$u_i \le e_i\ ,\quad v_i \le e_i\ ,$$
$$w_i \le u_i\ ,\quad w_i \le v_i\ ,\quad w_i \le e_i\ ,$$
$$w_i \ge 0\,,\quad w_i \ge u_i + v_i + e_i - 2,$$
$$\sum_{i=1}^{M} e_i = K,$$
$$\begin{cases} u_i + u_{2M-i+1} = e_i\,, & \\ v_i + v_{2M-i+1} = e_i\,, & \text{for inverted palindrome} \\ e_i = e_{2M-i+1}\,, & \end{cases}$$
$$q_{b,i} \le \sum_{(l,s,p)\in S''_{b,i}} z_{l,s,p}\,,$$
$$q_{b,i} \le Lq_{b,i}\,,$$
$$q_{b,i} \ge \sum_{(l,s,p)\in S''_{b,i}} z_{l,s,p} + L(q_{b,i} - 1),$$
$$q_{b,i} \ge 0,$$
$$\sum_{s} z_{l,s} = 1 \qquad \forall l \in \{1,\ldots,L\},$$
$$\sum_{s} z_{1,s,p} = \sum_{s} z_{2,s,p} = \ldots = \sum_{s} z_{L,s,p} \qquad \forall p \in \{0,1\},$$
$$S''_{b,i} = \{(l,s,p) \mid d_{l,s,i} = b,\, i \le M\} \cup \{(l,s,p) \mid d_{l,s,i+p} = b,\, M < i \le 2M\},$$
$$e_i, u_i, v_i \in \{0,1\},\quad z_{l,s} \in [0,1],$$
$$b \in \{\mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}\},\quad i \in \{1,\ldots,M\},$$
$$l \in \{1,\ldots,L\},\quad s \in \aleph,\quad p \in \{0,1\}.$$

$$S''_{b,i} = \{(l,s,p) \mid d_{l,s,i} = b,\, i \le M\} \cup \{(l,s,p) \mid d_{l,s,i+p} = b,\, M < i \le 2M\}. \quad \mathbf{(41)}$$

The linearization constraints for $q_{b,i} = y_{b,i}\sum z_{l,s,p}$ becomes:

$$\begin{aligned} &q_{b,i} \geq \sum_{(l,s,p)\in S''_{b,i}} z_{l,s,p} + L(y_{b,i}-1), \\ &q_{b,i} \geq 0, \\ &q_{b,i} \leq \sum_{(l,s,p)\in S''_{b,i}} z_{l,s,p}, \\ &q_{b,i} \leq Ly_{b,i}. \end{aligned} \tag{42}$$

To ensure that all the proposed sites are from the same set, additional constraints for $z_{l,s,p}$ are placed as follows:

$$\begin{aligned} &\sum_{s,p} z_{l,s,p} = 1 \text{ for all } l, \quad 0 \leq z_{l,s,p} \leq 1; \\ &\sum_{s} z_{1,s,p} = \sum_{s} z_{2,s,p} = \ldots = \sum_{s} z_{L,s,p} \text{ for all } p. \end{aligned} \tag{43}$$

Therefore we have program (P7) as the mixed 0-1 linear program for searching dyad-type TFBS.

Serial-type TFBS

Another often seen type of regulators is zinc-fingers which is a zinc-containing protein chain. Zinc-fingers binding site is a serial-type TFBS which is a chain of trinucleotide groups. Each of these trinucleotide groups contains two reactive bases. The logical constraint is formulated as follows:

$$e_i + e_{i+1} + e_{i+2} = 2 \quad \text{for } i \in \{1, 4, 7, \ldots\ldots\}. \tag{44}$$

Constraint (44) should be set depending on the length of consensus which is supposed a multiple of 3. *K* should be set equal to 2*M*/3 when searching zinc-fingers binding sites.

Another feature of zinc-fingers binding sites is that the first base of each triplet

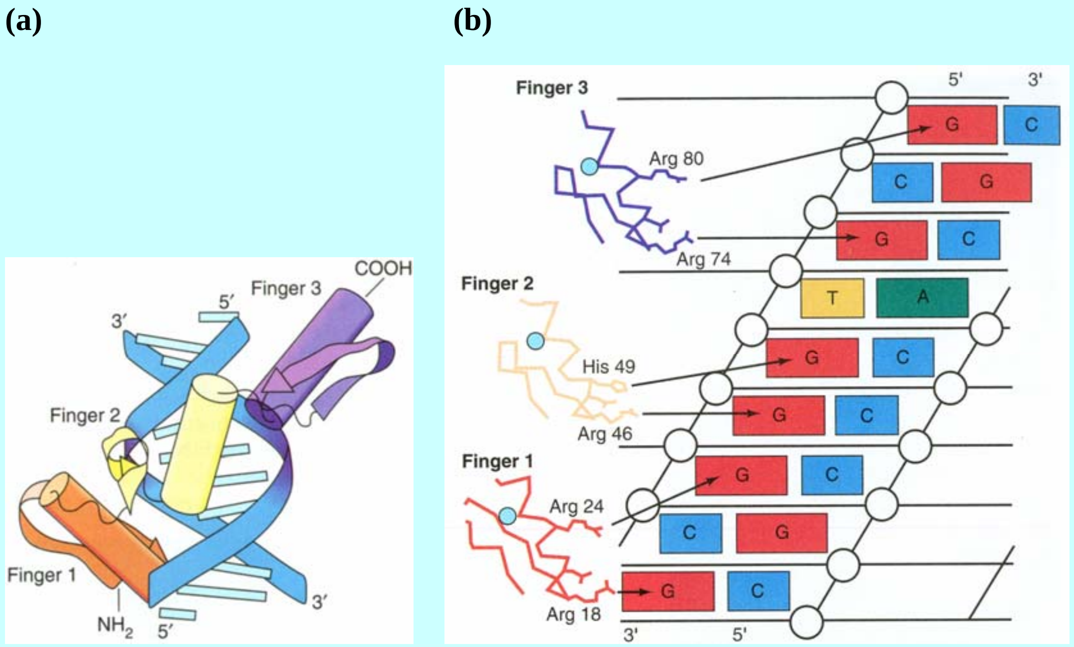


Figure 15 Zif268 zinc finger regulator, an example of series-type regulator. **(a)** Arrangement of the three zinc fingers of Zif268 in a curved shape to fit into the major groove of DNA. **(b)** Summary of interactions between Zif268 zinc finger amino acids and DNA bases. Each of the three fingers has two amino acids (all but one of the six are arginines) that make specific contact with guanines in the DNA major groove.(Pavletich and Pabo, 1991)

must be ‘G’. This is because the first two of the three DNA contacting amino acids are the same in every case: Arg and Asp. This arginine in each finger makes direct contact with a guanine in each triplet. A base assignment constraint is placed as:

$$e_i = 1,\ u_i = 1 \text{ and } v_i = 0 \qquad \text{for } i \in \{1, 4, 7, ...\}. \tag{45}$$

Like a mono-type TFBS, there may be occurrences with the opposite direction. Therefore the program for serial-type TFBS finding is formulated as program (P6) accompanied with constraints (44) and (45).

5.4 Suboptimal Solutions and Exception Rules

The proposed program can obtain only one global optimal solution. Practically biologists need more than one consensus sequence solution to make further discrimination and verification. In addition, there may be a case that the target signal is weaker than other noises. To successfully find out the TFBS of concern, we need exception constraints for two separate purposes: finding suboptimal solutions and exclude noises. The exception constraints include constraints for excluded solutions and general exception rules, as described in Chapter 3.

For the proposed program discussed in this chapter, the exception constraints are totally the same as those discussed in Chapter 3. These include constraints for finding suboptimal solutions (see §3.3) and general exception rules (see §3.4).

5.5 Complexity Analysis of Searching Space

When utilizing three binary variables (i.e. *u*, *v*, and *e*) to represent every base in the consensus sequence, we may encounter a searching space of 2^{3M} consensus sequences. Fortunately, not all of these binary variables are independent and a large part of searching space is pruned. First, only *K* of (u_i, v_i) pairs are independent because only the bases at reactive positions are compared. Second, because only *K* of e_is in a half site are allowed to be 1, the number of combinations of e_i values is C_K^M. Therefore, the total number of possible combinations of binary variables is $2^{2K} \cdot C_K^M$.

From the discussion in Chapter 3 we know the computational time is proportional to the number of sequences and almost no effect to the size of each sequence. The worst case of computational time may be roughly of order $\text{O}(l \cdot 2^{2K} \cdot C_K^M)$. Generally, the computational time is acceptable for most cases where *M* less than or equal to 12 and *K* less equal to 7.

(a)

K	*M*	Consensus Sequence	Score	Computational Time (h:mm:ss)
3	7	**TG..A....T..CA**	92	0:17:44
	8	**T.TG........CA.A**	94	0:21:31
	9	**.T.TG........CA.A.**	94	0:26:59
	10	**T...T.A......T.A...A**	94	0:34:59
	11	**.T...T.A......T.A...A.**	94	1:21:11
4	7	**GTGA......TCAC**	120	0:27:11
	8	**.GTGA......TCAC.**	120	0:28:44
	9	**..GTGA......TCAC..**	120	0:57:07
	10	**...GTGA......TCAC...**	120	0:58:08
	11	**....GTGA......TCAC....**	120	1:17:19
5	7	**GTGA.C..G.TCAC**	136	0:30:09
	8	**TGTGA......TCACA**	147	0:52:02
	9	**.TGTGA......TCACA.**	147	1:13:41
	10	**..TGTGA......TCACA..**	147	2:42:36
	11	**...TGTGA......TCACA...**	147	5:23:40
6	7	**GTGAA.TA.TTCAC**	151	0:27:10
	8	**TGTGA.C..G.TCACA**	164	0:51:50
	9	**.TGTGA.C..G.TCACA.**	164	4:37:04
	10	**T.TGTGA......TCACA.A**	165	10:58:17
	11	**A..TGTGA......TCACA..T**	166	33:23:59

(b)

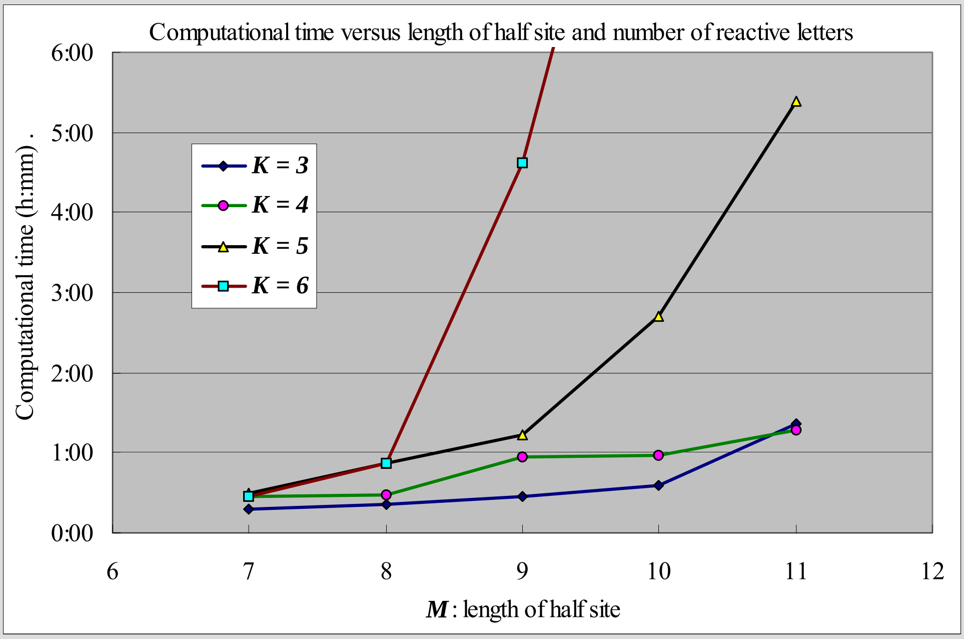


Figure 16 Experiments of pattern-free TFBS finding on searching for CRP-binding sites in 18 *E.coli* DNA sequences, which taken from Stormo et al. (1989) **(a)** the list of consensus sequences found under various reactive base numbers *K* and various half site lengths *M*. **(b)** polygon graph illustrating the computational time under various settings.

Figure 16 illustrates a prediction of CRP-binding sites of *E. Coli* by program (P4) with $7 \leq M \leq 11$ and $3 \leq K \leq 6$ on a Pentium 4-3.0GHz PC. This example has 18

DNA sequences each of which has 105 bp. The consensus is considered inverted palindrome and no exclusive constraints are applied. Although different solutions of the consensus sequence in experiments of reactive letter number greater or equal to 4, the positions of all the TFBS are the same as the results listed in Stormo et al. (1989).

5.6 Experimental Results

The proposed method is implemented and publicly available as the second version of software Global Site Seer (GSS2). This section reports results of GSS2 on several eukaryotic examples. All the examples are TFBS of dimerized regulators and benefit from inverted palindrome. These examples include two C. Elegans sequence sets regulated by daf-19 and lin-32 collected from WormBase (Stein et al., 2001), and a benchmark sequence set from assessment system of Tompa et al. (2005) originally collected from TRANSFAC (Matys et al., 2003). All the sequence sets in FASTA format are available at online supplementary.

daf-19 regulons

The C. elegans gene daf-19 encodes an RFX-type transcription factor that is expressed specifically in all ciliated sensory neurons. Target sites for mammalian RFX-type transcription factors (X boxes) typically are 13-14 bp imperfect inverted repeats (Swoboda et al., 2000). The example set contains five sequences listed in Swoboda et al. (2000) and three listed in WormBase. Our searching result is listed in Table 6. Compared GSS2 prediction with Swoboda et al. (2000), the same consensus is found but two variants on individual TFBS occur: TFBS in *che-2* exactly matches but at different location; and TFBS in osm-1 listed in Swoboda et al. (2000) doesn’t appear. In fact, the *osm-1* TFBS seems more like a silencer in our experiment because the TFBS listed in Swoboda et al. (2000) is at +4 position.

lin-32 regulons

lin-32 encodes a basic helix-loop-helix transcription factor that is required for development of several types of neurons, including the touch receptor neurons and the male sensory ray neurons (Krause et al., 1997; Portman and Emmons, 2000). This sequence set contains 9 upstream sequences of various genes regulated by lin-32 with length varying from 326 to 1050. The GSS2 prediction is listed in Table 7. The accuracy is not verified here because no related experimental report from laboratory is available for comparison. But in our opinion, this result is good because the reactive bases in consensus sequence are close together and have a TG group. Therefore, these TFBS are strong and meaningful signals and possibly the lin-32 binding sites.

hm17r (Tompa et al., 2005)

We also test a sequence set from an assessment system designed by Tompa et al. (2005). The whole dataset in this assessment system includes 3 classes each of which contains 56 sequence sets. This dataset is totally from real genome and the TFBS are very hard to find because they have various features for binding zinc-finger, HTH, HLH, and leucine-zipper, etc. Besides, the most intractable part is, the insertion and deletion errors. Because the dataset is designed for assessing TFBS finding tools designed in TCM (zero or more occurrence per sequence) mode, most of them are not appropriate for testing the proposed mixed 0-1 linear program, which is designed for OOPS (One Occurrence Per Sequence) mode from original concept. The sequence set we used here is hm17r, a human DNA sequence set from real class of Tompa's dataset. Every sequence in hm17r is 500 bp long with a TFBS for a dimerized regulator. The prediction is listed in Table 8. Comparing with answer of Tompa et al. (2005), there are two differences in this prediction: the TFBS position in Seq_3 is -328 differing from Tompa's answer, -173; in Seq_5 there is no occurrence by answer; and TFBS position in Seq_9 is -173 differing from Tompa's answer, -138. All different answers

Table 6 A Prediction of *daf-19* binding sites.

Gene regulated	GSS2		Score	*Swoboda et al.*(2000)	
	...GTT.CCATGG.AAC...	Posi.	**85**	Motifs	Posi.
che-2	ctg**GTT**g**TCATGG**t**GAC**tgc	-57	10	**GTT**g**TCATGG**t**GAC**	-130
daf-19	ttg**GTT**t**CCATGG**a**AAC**tac	-109	12	**GTT**t**CCATGG**a**AAC**	-109
osm-1	att**GTA**t**CCATAC**c**AAC**atc	-1211	9	**GCT**a**CCATGG**c**AAC**	-86
osm-6	cat**GTT**a**CCATAG**t**AAC**cac	-100	11	**GTT**a**CCATAG**t**AAC**	-100
xbx-1	ccc**GTT**t**CCATGG**t**AAC**cgt	-79	12	**GTT**t**CCATGG**t**AAC**	-79
dyf-3	gga**GTT**t**CTATGG**g**AAC**gga	-88	11	**N/A**	**N/A**
pkd-2	tcc**GTT**t**CTATGC**a**AAA**aac	-231	9	**N/A**	**N/A**
xbx-4	cta**GTT**g**CCATGA**c**AAC**cgc	-35	11	**N/A**	**N/A**

Table 7 A Prediction of *lin-32* binding sites.

Gene regulated	position	Consensus Sequence	Score
		TGAAA (9) TTTCA	**78**
hlh-2	-457	t**GGAAA**tattaaaga**ATTCT**t	7
cfi-1	-738	t**TAAAA**ttaaattat**TTTCA**a	9
cwp-4	-332	t**TTAAA**tatattttt**TTTCA**g	9
egl-46	-239	g**TGAAA**attgactag**ATTCA**c	9
lin-22	-348	t**TGAAT**ttctggga**TTTCT**t	8
mab-3	-184	t**TGAAA**atttgactt**TTCCA**c	9
mab-5	-56	g**TGAAA**tatgtgtcg**TTTCA**c	10
tbb-4	-300	c**AGAAA**aagtcaaca**TTACA**g	8
twk-21	-374	c**TGAAA**attcaagta**TTTAA**a	9

Table 8 A Prediction of DNA motifs in *hm17r* sequence set.

Seq. name	GSS2		Score	*Tompa et al.*(2005)	
	.....GGGAA.TTCCC.....	Posi.	**97**	Motifs	Posi.
Seq_0	actcc**GGGAA**t**TTCCC**tggcc	-83	10	tcc**GGGAA**t**TTCCC**tg	-81
Seq_1	gctcc**GGGAA**t**TTCCC**tggcc	-83	10	tcc**GGGAA**t**TTCCC**tg	-81
Seq_2	gctcc**GGGAA**t**TTCCC**tggcc	-85	10	tcc**GGGAA**t**TTCCC**tg	-83
Seq_3	ctccg**GGGAA**g**TTGGC**agtat	-328	8	gcttggaaattccggagc	-173
Seq_4	aaagt**GGGAA**a**TTCCT**ctgaa	-144	9	gt**GGGAA**a**TTCC**	-141
Seq_5	gtatc**GGGAA**t**TGCTC**cctcc	-274	8	*<No Instances>*	**N/A**
Seq_6	ggcag**GGGAA**t**CTCCC**tctcc	-274	9	g**GGGAA**t**CTCC**	-270
Seq_7	aatgt**GGGAT**t**TTCCC**atgag	-79	9	aaatgt**GGGAT**t**TTCCC**	-80
Seq_8	aatcg**TGGAA**t**TTCCT**ctgac	-86	8	**GGAA**t**TTCCT**	-80
Seq_9	catcg**TGGAT**a**TTCCC**gggaa	-173	8	attggggatttcctc	-138
Seq_10	gccct**GGGGG**c**TTCCC**cgggc	-136	8	t**GGGGG**c**TTCCC**c	-132

provided by Tompa have lower matches than the TFBS found by GSS2. These tests illustrate that the determined consensus successfully helps determinate most TFBS and can be regarded as a good result.

5.7 Software Package: "Global Site Seer v2"

A software package "Global Site Seer 2.0" is designed for pattern-free TFBS finding and is available by http://www.iim.nctu.edu.tw/~cjfu/gss2.htm.

Chapter 6 Discussion

6.1 Features of Proposed Methods

This study proposes a mixed 0-1 linear programming approach to search TFBS under various conditions. The final result of this study is a mixed 0-1 linear program for solving pattern-free TFBS finding problems. Advantageous features of this approach include:

(i) **A pattern-driven design which can search longer patterns than current enumeration approaches.** Because only the reactive bases are enumerated in consensus sequence, the computational time is notably reduced.

(ii) **A global optimal consensus is promised.** As a nature of mixed 0-1 linear program, the consensus sequence with maximum matches is surely obtained.

(iii) **No prerequisite shared pattern is needed.** The proposed method can search TFBS of an undiscovered regulation with limited information like length of regulatory region and number of reactive bases.

(iv) **Capable of identifying TFBS with spacers dispersed in regulatory region.** Most current TFBS finding methods have difficulty to search patterns containing inactive bases. Contrarily, the proposed method benefits from these inactive bases because searching space is pruned.

(v) **Structural features can be involved.** In the proposed method various structure features of TFBS can be formulated to help prune searching space and improve precision, e.g. inverted palindrome or direct repeat.

This approach also has several weaknesses as follows:

(i) **Exponential growing computational time to the number of reactive bases.** Although a notable feature over current pattern-driven enumeration methods is that the critical factor of searching time is number of reactive bases instead of pattern length, the limitations on length of regulatory sites still exist.

(ii) **Only one solution obtained.** By nature of optimization program, the proposed method cannot simultaneously search multiple patterns. Finding suboptimal solutions in this approaches still required individual program in which previously obtained optimal consensus sequences are excluded.

(iii) **Difficult to search consensus with base variability.** The proposed method utilizes consensus sequences consisted only by four distinct nucleic acid types. The consensus sequence is a distinct ideal model of TFBS and only exact base matches within sites contribute the matching score. But in fact, there may be some reactive bases replaceable by other nucleotides which have similar sensitivity to regulators.

As a nature of pattern-driven and mixed 0-1 linear programming design, the proposed method can find the optimal consensus in an acceptable computational time. The most advantaged property to current heuristic methods is the capability of embedding logical constraints. These logical constraints telling many kinds of specific features and exclusive rules notably increase the precision and efficiency.

6.2 Issues in Approach Design

Based on assumptions of occurrences in each sequence, there are several different searching modes for the computer-based determination of transcription binding sites. These modes are generally defined in studies of sequence-driven approaches which apply probability models to iteratively search the most significant conserved signal. CONSENSUS (Hertz and Stormo, 1995), a statistical based system for identifying consensus patterns of DNA sequence and protein sequences, provides three modes of searching: One Occurrence per Sequence (OOPS), One or More Occurrences per Sequence (OMOPS) and Zero or More Occurrences per Sequence (ZMOPS). Another TFBS searching tool, MEME (Bailey et al., 1995), also can search motif under three different modes: One Occurrence per Sequence (OOPS), Zero or One Occurrence per Sequence (ZOOPS) and Two-Component Mixture (TCM)—each sequence may contain any number of non-overlapping occurrences of each motif.

Which sequence mode is appropriate depends on the purpose of motif finding work. When a sequence set is given from any combination of upstream sequences of various genes and the purpose is to discover any possible regulations, searching tools capable of handling TCM mode are obviously much appropriate. For analyzing function of a particular regulator, the sequence set shall be prepared more conscientiously from sequences upstream genes regulated by the target transcription factor. And in this case OOPS and OMOPS are more suitable for finding the DNA motifs precisely.

The proposed approach is only designed for searching sequences in OOPS or OMOPS mode and is very powerful when analyzing a specific function regulatory. It is not appropriate to search sequences in ZOOPS, ZMOPS and TCM modes for any possible undiscovered regulatory.

6.3 Benchmark Results

Assessment system of Tompa et al. (2005)

In 2005 Tompa et al. designed an assessment system for TFBS finding tools. In this system they prepared three testing groups each of which contains 56 eukaryotic DNA sequence sets in various sizes. A statistical evaluation system is designed to measure the accuracy of all kinds of TFBS finding tools. The benchmark data set is designed in AMOPS or TCM modes, containing TFBS with insertion and deletion errors. These sequence sets contain all kinds of noises and most TFBS are relatively weak. Generally, a TFBS searching tool which has accuracy over 0.2 is regarded as a good design.

Testing results of proposed approach

The testing result of proposed method on benchmark system of Tompa et al. (2005) is poor. Possible defects of the proposed method which lead to this poor performance may include:

(i) Designed in OOPS mode

(ii) Unable to handle insertion and deletion errors

(iii) No further refining strategies for multiple suboptimal alternatives

(iv) Weak on treating base variability

And surely, all these weaknesses are active issues of further researches.

Chapter 7 Concluding Remarks

This study develops a series of mixed 0-1 linear programming approaches to search for transcription factor binding sites. The most advantageous property of the proposed method to existing DNA motif finding tools and is capable to find TFBS without any given shared patterns. Nevertheless, the accuracy still can be improved by more complex design. Some issues remain for further study:

(i) The first issue is about the treatment of multiple consensuses. For searching a weak target signal, one needs more than one solution for further verification. These solutions can be obtained by applying exception constraints for banishing previously known solutions in the proposed method. Two strategies to obtain the final solution are designed as follows: One is to apply different scoring functions like PSSM or log-likelihood function to verify these consensus sequences. That is, for every consensus we can make a log-likelihood evaluation of all the TFBS in sequences and then compare the score to obtain the final solution. Another strategy is to count the number of occurrences. As a post treatment step, more other TFBS may be determined based on a consensus accompanied with a score threshold. The consensus having most TFBS occurrences is more possibly the target.

(ii) Another issue is about the quality of sequence set. The collection step of a sequence set is critical to accuracy. The proposed method is originally designed for sequence sets which have one occurrence per sequence (OOPS), and may not be appropriate if not all the sequences contain target TFBS. Other modes like ZOOPS (Zero or One Occurrence per Sequence) or TCM (zero or more occurrences per sequence) are not considered in this method. Although TCM mode is regarded as the most convenient to biologists, it compromises to accuracy. The consensus sequence is supposed to be an ideal binding pattern so it allows no ambiguity. Any sequence with no occurrences is not recommended because it will dramatically affect the quality of the consensus sequence.

(iii) The third issue is to formulate various possible features. More complicated features can be articulated as logical constraints. One feature is the specific base group positioned case by case. An example is trinucleotide group: many dyad type TFBS have at least 3 reactive bases close together in half site. Another example is TG kink, an often seen structure in regulatory region (Schultz et al., 1991). It may not contact regulatory protein but is very important because it allows DNA strand bend to fit the regulatory protein. The exact positions of these specific groups in the consensus vary by cases. Formulating these features may be difficult but very helpful in finding related TFBS effectively.

Appendix

The *Escherichia coli* DNA sequences containing CRP-binding sites. This data set contains 18 gene upstream sequences, each of which is 105-bp long.

cole1	TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATAGCGCGTGGTGTGAAAGACTGTTTTTTTGATCGTTTTCACAAAAATGGAAGTCCACAGTCTTGACAG
ecoarabop	GACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTGCTATGCCATAGCATTTTTATCCATAAG
ecobglr1	ACAAATCCCAATAACTTAATACTTGACATTTGTTATATATAACTTTATAAATTCCTAAAATTACACAAAGTTAATAACTGTGAGCATGGTCATATTTTTATCAAT
ecocrp	CACAAAGCGAAAGCTATGCTAAAACAGTCAGGATGCTACAGTAATACATTGATGTACTGCATGTATGCAAAGGACGTCACATTACCGTGCAGTACAGTTGATAGC
ecocya	ACGGTGCTACACTTGTATGTAGCGCATCTTTCTTTACGGTCAATCAGCAAGGTGTTAAATTGATCACGTTTTAGACCATTTTTTCGTCGTGAAACTAAAAAAACC
ecodeop	AGTGAATTATTTGAACCAGATCGCATTACAGTGATGCAAACTTGTAAGTAGATTTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGGAGTAGATGTTAGAATA
ecogale	GCGCATAAAAAACGGCTAAATTCTTGTGTAAACGATTCCACTAATTTATTCCATGGCACACTTTTCGCATCTTTGTTATGCTATGGTTATTTCATACCATAAGCC
ecoilvbpr	GCTCCGGCGGGGTTTTTTGTTATCTGCAATTCAGTACAAAACGTGATCAACCCCTCAATTTTCCCTTTGCTGAAAAATTTTCCATTGTCTCCCCTGTAAAGCTGT
ecolac	AACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGATCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCAC
ecomale	ACATTACCGCCAATTCTGTAACAGAGATCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGTTGCCGTATAAAGAAACTAGAGTCCGTTTA
ecomalk	GGAGGAGGCGGGAGGATGAGAACACGGCTTCTGTGAACTAAACCGAGGTCATGTAAGGAATTTCGTGATGTTGCTTGCAAAAATCGTGGCGATTTTATGTGCGCA
ecomalt	GATCAGCGTCGTTTTAGGTGAGTTGTTAATAAAGATTTGGAATTGTGACACAGTGCAAATTCAGACACATAAAAAAACGTCATCGCTTGCATTAGAAAGGTTTCT
ecoompa	GCTGACAAAAAAGATTAAACATACCTTATACAAGACTTTTTTTTCATATGCCTGACGGAGTTCACACTTGTAAGTTTTCAACTACGTTGTAGACTTTACATCGCC
ecotnaa	TTTTTTAAACATTAAAATTCTTACGTAATTTATAATCTTTAAAAAAAGCATTTAATATTGCTCCCCGAACGATTGTGATTCGATTCACATTTAAACAATTTCAGA
ecouxul	CCCATGAGAGTGAAATTGTTGTGATGTGGTTAACCCAATTAGAATTCGGGATTGACATGTCTTACCAAAAGGTAGAACTTATACGCCATCTCATCCGATGCAAGC
pbr-p4	CTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTC
trn9cat	CTGTGACGGAAGATCACTTCGCAGAATAAATAAATCCTGGTGTCCCTGTTGATACCGGGAAGCCCTGGGCCAACTTTTGGCGAAAATGAGACGTTGATCGGCACG
(tdc)	GATTTTTATACTTTAACTTGTTGATATTTAAAGGTATTTAATTGTAATAACGATACTCTGGAAAGTATTGAAAGTTAATTTGTGAGTGGTCGCACATATCCTGTT

Reference

Ao, W., Gaudet, J., Kent, W.J., Muttumu, S. and Mango, S.E. (2004) Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* **305** (5691), 1743–1746.

Bailey, T. and Elkan, C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **21** (1-2), 51-80.

Bailey, T. and Elkan, C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. 21–29 (AAAI Press, Menlo Park, CA, 1995).

Blanchette, M., Schwikowski, B. and Tompa, M. (2000) An exact algorithm to identify motifs in orthologous sequences from multiple species. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 37-45, San Diego, CA.

Brāzma, A., Jonassen, I., Eidhammer, I. and Gilbert, D. (1998) Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology* **5** (2), 279-305.

Buhler, J. and Tompa, M. (2002) Finding Motifs Using Random Projections. *Journal of Computational Biology* **9** (2), 225-242.

Califano, A. (2000): SPLASH: structural pattern localization analysis by sequential histograms. *Bioinformatics* **16** (4), 341-357

Eskin, E. and Pevzner, P. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics (Supplement 1)* **18** (1), S354–S363.

Favorov, A.V., Gelfand, M.S., Gerasimova, A.V., Mironov, A.A. and Makeev, V.J. (2004) Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length and its validation on the ArcA binding sites. *Proceedings of BGRS 2004* (BGRS, Novosibirsk, 2004).

Fratkin, E., Naughton, B. T., Brutlag, D. L. and Batzoglou, S. (2006) MotifCut: Regulatory motifs finding with maximum density subgraphs. *ISMB (Supplement of Bioinformatics)* 2006, 156-157

Frith, M.C., Hansen, U., Spouge, J.L. and Weng,Z. (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research* **32** (1), 189–200.

Galas, D., Eggert, M. and Waterman, M (1985) Rigorous pattern-recognition methods for DNA sequences: analysis of promoter sequences from Escherichia coli. *Journal of Molecular Biology* **186** (1), 117-128.

Gelfand, M, Koonin, E. and Mironov, A. (2000) Prediction of transcription regulatory sites in archaea by a comparative genomic approach. *Nucleic Acids Research* **28** (3), 695-705

Hertz, G.Z., Hartzell, G.W. and Stormo, G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Computer Applications in the Biosciences* **6** (2), 81-92.

Hertz, G.Z. and Stormo, G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15** (7-8), 563–577.

Huerta, A. M., Salgado, H., Thieffry, D. and Collado-Vides, J. (1998) RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Research* **26** (1), 55-59.

Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000) Computational identification of cis-regulatory elements associated with functionally coherent groups of genes in Saccharomyces cerevisiae. *Journal of Molecular Biology* **296** (5), 1205–1214.

Krause, M., Park, M., Zhang, J.M., Yuan, J., Harfe, B., Xu, S.Q., Greenwald, I., Cole, M., Paterson, B. and Fire, A., (1997) A C. elegans E/Daughterless bHLH protein marks neuronal but not striated muscle development. *Development* **124** (11), pp. 2179–2189.

Lawrence, C.E., Altschul, S. Boguski, M. Liu, J. Neuwald, A. and J, Wootton. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262** (5131), 208-214.

Lawrence, C.E. and Reilly, A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *PROTEINS: Structure, Function, and Genetics* **7** (1), 41-51.

Liang, S., Samanta, M. and Biegel, B. (2004) cWINNOWER algorithm for finding fuzzy DNA motifs. *Journal of Bioinformatics and Computational Biology* **2** (1), 47–60.

Liu, X.S., Brutlag, D.L. and Liu, J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing*, 127-138.

Li, H.L. and Fu, C.J. (2005) A linear programming approach for identifying a consensus sequence on DNA sequences. *Bioinformatics* **21** (9), 1838-1845.

Li, M., Ma, B. and Wang, L. (1999) Finding similar regions in many strings. *Proceedings of the 31st ACM Annual Symposium on Theory of Computing*, 473-482.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* **31** (1), 374-378.

Needleman, S. and Wunsch, C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48** (3), 443-453.

Pavesi, G., Mauri, G. and Pesole, G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *ISMB 2001 (Supplement of Bioinformatics)*, 207-214.

Pavesi, G., Mereghetti, P., Mauri, G. and Pesole, G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research* **32** (Web Server Issue), W199–W203.

Peng, C.H., Hsu, J.T., Chung, Y.S., Lin, Y.J., Chow, W.Y., Hsu, D.F. and Tang, C.Y. (2006), Identification of Degenerate Motifs Using Position Restricted Selection and Hybrid Ranking Combination, Nucleic Acids Research, 34, pp. 6379-6391. Pesole, G.., Prunella, N., Liuni, S., Attimonelli, M. and Saccone, C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucleic Acids Research* **20** (11), 2871-2875.

Pevzner, P. and Sze, H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, 269-278.

Portman, D.S. and Emmons, S.W. (2000) The basic helix-loop-helix transcription factors LIN-32 and HLH-2 function together in multiple steps of a C. elegans neuronal sublineage. *Development* **127**(24), 5415–5426.

Régnier, M. and Denise, A. (2004) Rare events and conditional events on random strings. *Discrete Mathematics and Theoretical Computer Science* **6** (2), 191–214.

Schrage, L. (1999) Optimization Modeling With Lingo, LINDO Systems Inc., Chicago.

Sinha, S. and Tompa, M. (2003) Performance comparison of algorithms for finding transcription factor binding sites. *3rd IEEE Symposium on Bioinformatics and Bioengineering* (ed. Bourbakis, N.G.). 214–220 (IEEE Computer Society, New York, 2003).

Sinha, S. and Tompa, M. (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research* **31** (13), 3586–3588.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J. and Spieth, J., 2001. WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Research* **29** (1), 82–86.

Stormo, G.D. and Hartzell, G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. Proceedings of the National Academy of Sciences of the USA. **86** (4), 1183-1187.

Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics* **16** (1), 16-23.

Swoboda, P., Adler, H.T. and Thomas, J.H., 2000. The RFX-type transcription factor DAF-19 regulates sensory neuron cilium formation in C. elegans. *Molecular Cell* **5** (3), 411–421.

Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J. and Stormo, G.D. (2001) A comparative genomics approach to prediction of new members of regulons. *Genome Research* **11** (4), 566-584.

Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouzé P, Moreau Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17** (12), 1113–1122.

Tompa, M. (1999) An exact method for finding short motifs in sequences with application to the Ribosome Binding Site problem. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 262-271.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu,Y., Kent,W.J., Makeev,V.J., Mironov,A.A., Noble,W.S., Pavesi,G., Pesole,G., Regnier,M., Simonis,N., Sinha,S., Thijs,G., van Helden,J., Vandenbogaert,M., Weng,Z., Workman,C., Ye,C. and Zhu,Z. (2005) Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites. *Nature Biotechnology* **23** (1), 137-144.

Tsai, H.K., Huang, G.T.W., Chou, M.Y., Lu, H.H.S. and Li, W.H. (2006) Method for identifying transcription factor binding sites in yeast. *Bioinformatics* **22** (14), 1675-1681.

Waterman, M., Galas, D., and Arratia, R. (1984) Pattern recognition in several sequences: consensus and alignment. *Bulletin of Mathematical Biology* **46** (4), 512-527.

Workman, C.T. and Stormo, G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pacific Symposium on Biocomputing* (ed. Altman, R., Dunker, A.K., Hunter, L. and Klein, T.E.). 467–478 (Stanford University, Stanford, CA, 2000).