行政院國家科學委員會專題研究計畫　期中進度報告

以全域最佳化方法求解 DNA 序列之共同區間定址問題(1/3)

計畫類別：個別型計畫
計畫編號：NSC93-2213-E-009-070-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：國立交通大學資訊管理研究所

計畫主持人：黎漢林

計畫參與人員：傅昶瑞

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 94 年 5 月 24 日

行政院國家科學委員會專題研究計畫期中精簡報告

以全域最佳化方法求解 DNA 序列之共同區間定址問題(1/3)

A Global Optimization Method for Identifying Common Sites on DNA Sequences (1/3)

計畫編號：NSC 93-2213-E-009-070
執行期限：93 年 8 月 1 日至 94 年 7 月 31 日
主持人：黎漢林　　國立交通大學資訊管理研究所

一、中英文摘要

DNA 序列之共同區間定址問題從過去 20 年至今一直被廣泛討論。目前許多求解 CSI(Consensus Sequence Identification)問題的方法多引用最大相似技術[Stormo 1989, Ecker et al 2002]。這類方法不僅無法確認得到的解答是否為全域最佳解，而且無法處理包含有上百筆序列之大尺度 CSI 問題。本研究計畫提出一新的方法以求解大尺度 CSI 問題以得到全域最佳解。我們首先將 CSI 問題轉成一非線性 0-1 規劃問題；此一問題可經由線性化的處理後轉化成一線性 0-1 規劃問題。給定 *L* 筆 DNA 序列資料，欲找出在此 *L* 條序列中長度為 *k* 的共同區間，則我們可將其相關的 CSI 問題轉化為一只含有 *4k* 個 0-1 變數之線性 0-1 規劃問題。由於 *L*（序列的筆數）不會影響 0-1 變數的個數，我們的模式可以處理大筆數的 CSI 問題。因此我們可以發展一分散式計算系統求解包含數百筆 DNA 序列的問題。本研究計畫擬分三年進行。第一年著重於發展 CSI 問題的最佳化求解模式。第二年為拓展此模式尋找多個次佳解，以提供生物學家更多 CSI 的參考答案。第三年則為在 PC 上發展一分散式計算軟體以處理共同區間定址問題。

關鍵詞：關鍵字：最佳化、分子生物學、蛋白質接合、共同區間

Abstract

Consensus sequence identification (CSI) in multiple DNA sequences has been discussed widely in last two decades. Many current methods of solving CSI problems are based on the maximum likelihood techniques [Stormo 1989, Ecker et al 2002]. These methods, however, have no guarantee to find a globally optimal solution of a CSI problem. In addition, they are difficult to handle large-size CSI problems with hundreds of DNA sequences. This project proposes a naïve method to solve a large CSI problem to find a global optimum. We first formulate the CSI problem as nonlinear 0-1 optimization model. Such a model is then converted into a linear 0-1 problem by linearization techniques to reach a best fitted solution.

Given *L* DNA sequences with a consensus sequence known having *k* locations, we can formulate the related CSI problem as a linear 0-1 model which only contains *4k* 0-1 variables. Since *L* (i.e., the number of sequences) has no effect on the number of 0-1 variables, our model can treat large size CSI problem. A distributed computation system is then developed to solve a CSI problem with hundreds of DNA sequences. The whole project will be executed in three years. The first year is emphasized on building an optimization model for solving CSI problems. The second year is to extend the previous model to discover the suboptimal solutions for biologists as more usable consultation. Software of distributed network computation system for solving the CSI problems will be developed in the third year.

Keywords: optimization, molecular biology, protein binding, consensus sequence

二、前言

The methods for determining a consensus pattern can be split into two parts. The first part is the model for describing the shared pattern; the second part is the algorithm for identifying the optimal common site according to its shared pattern. This study belongs to the second part. A consensus sequence identification (CSI) problem is, given a set of sequences known to contain binding sites for a common factor but not knowing where the site are, discover the location of the sites in each sequence (Stormo, 2000).

三、研究目的

This study proposes a linear programming method for solving a CSI problem to reach the globally optimal consensus sequence. Two examples of searching for CRP-binding sites and for FNR-binding sites in the Escherichia coli genome are used to illustrate the proposed method. The CSI problem is firstly formulated as a nonlinear mixed 0-1 program for alignment of DNA sequences, each of the four bases are coded with two binary variables and a matching score is designed. This nonlinear mixed 0-1 program is then converted into a linear mixed 0-1 program by linearization techniques. This study decomposes a CSI problem into several subprograms to be solved by a set of distributed computers linked via internet. Owing to some special features of the binary relationships, this linear 0-1 program includes 2m binary variables where m is the number of active letters in the common site. Some very attractive properties of this method are firstly that the required number of binary variables is independent of the number of sequences and the size of each sequence. That means, the proposed method is computationally efficient in solving a CSI problem with a large data size. Secondly, the proposed method is guaranteed to find the global optimum instead of a local optimum. Thirdly, many kinds of specific features accompanied with a CSI problem can be formulated straight forwardly as logical constraints and embedded into the linear program.

四、文獻探討

The CSI problem is critical in research on gene expression such as the protein-binding site in a DNA strand. For the last decade several good methods have been developed for solving such problems (Brazma et al., 1998). Of those methods, the maximum likelihood approach (Stormo et al., 1989; Hertz et al., 1990) is the best known. The traditional maximum likelihood approach, which measures information content to determine alignments, works fairly well and is reliable on discovering the common sites. However, they are still not able to determine the complete set of regulatory interactions for complicated promoters typical of metazoans (Stormo, 2000).

Recently, Ecker et al. (2002) utilized optimization techniques to reformulate the maximum likelihood approach for solving CSI problems. They adopted a probabilistic model and formulated a well-designed nonlinear model with reference to the expectation maximization algorithm of Lawrence and Reilly (1990). Their method, however, occasionally only finds a feasible solution or a local optimum: which means the best solution may not be found. Additionally, no further structural feature in a CSI problem can be embedded conveniently in their model.

An example of searching CRP-binding sites, as discussed in Stormo *et al*. (Stormo *et al*., 1989) and Ecker *et al*. (Ecker *et al*., 2002), is described as follows. Given eighteen letter sequences each 105 positions long, where each position contains a letter from the set {A, T, C, G}, find a common site of length16 with the pattern

$$L_1L_2L_3L_4L_5\,\square\square\square\square\square\square\,L_6L_7L_8L_9L_{10}$$

where L_i, $\square \in$ {A, T, C, G}and $\square$'s mean the positions of ignored letters.

Restated, the problem is to specify

(i) the L_i 's of the common site pattern

(ii) the location of the site in each given sequence, which can fit most closely the common site.

The following are difficulties associated with the method of Ecker *et al*. (2002) and other maximum likelihood methods (as reviewed in Brazma *et al*., 1998) for solving a CSI problem:

(i) Only a local optimal or feasible solution is obtained

Since Ecker *et al*. (2002) formulated a CSI problem as a non-convex nonlinear program, their method may only find local optima, as has been acknowledged (Ecker *et al*., 2002). Other maximum likelihood methods, which intend to maximize the probability of binding to the promoters in the sequences, may only find a feasible solution instead of finding a local optimal solution. It is not guaranteed that current maximum likelihood methods can reach the global optimum for general CSI problems.

(ii) Heavy computational burden

The nonlinear program in Ecker *et al*. (2002) contains too many nonlinear terms. The heavy computational burden in their method prohibits it from treating a CSI problem with a large number of sequences.

(iii)Difficulty of adding logical constraints

When identifying protein binding sites, there usually exists some specific features to be considered as logical constraints. For example, the letters of position L_i and L_{11-i} are expected to be complement (i.e. G with C and A with T). Formulating such a constraint in maximum likelihood approaches is a complex task. It is even impossible to formulate more complicated logical constraints (e.g. those with some ambiguity) when applying these approaches.

(iv)Fixed number of ignored letters

Maximum likelihood methods are mainly used to solve CSI problems with fixed number of ignored letters (e.g. six in the above example). However, in real world this number is unknown and need to be found by some preliminary processes.

(v) Difficulty of finding the second and the third best solutions

Since current methods may only find a local optimum. It is hard to find other solutions next to the best solution.

五、研究方法

In order to overcome the above difficulties of solving a CSI problem, this study proposes a novel method to treat the same problem that molecular biologists actually are interested in solving. We formulate a CSI problem as the identification of a consensus sequence that minimizes the number of differences between the proposed sites. Our basic concept is to reformulate a CSI problem as a mixed 0-1 linear program which only contains a limited number of 0-1 variables and most variables are continuous. Such a mixed 0-1 linear program can be solved effectively by commonly used branching-and-bound algorithms or a branch-cut algorithm (Balas *et al*. 1996). The advantages of the proposed method are listed below:

(i) It is guaranteed to find the globally optimal solution. Since the objective function and constraints are all linear, the program should converge to the global optimum.

(ii) It can effectively solve a CSI problem by a set of on-line computers as illustrated by our numerical experiments.

(iii) It is convenient to add logical constraints. Since the binary variables are very suitable to express logical relationship, various complicated constraints can be embedded directly into the proposed method.

(iv) It can be extended to treat CSI problems with unknown number of ignored letters.

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

This study firstly formulates a CSI problem as a nonlinear mixed 0-1 program. Then it converts this nonlinear mixed 0-1 program into a linear mixed 0-1 program using linearization techniques. To reduce the computational burden, many 0-1 variables in this linear mixed 0-1 program can actually be solved as continuous variables by an all or nothing assignment technique which improves greatly the computational efficiency of this program.

Nonlinear mixed 0-1 program

Here we use the example data in Stormo (1989), as listed in Appendix, to describe the proposed method. Firstly, represent the data in Appendix as an 18*105 data matrix D:

$$D = \begin{bmatrix} b_{1,1} & b_{1,2} & \Lambda & b_{1,105} \\ b_{2,1} & b_{2,2} & \Lambda & b_{2,105} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ b_{18,1} & b_{18,2} & \Lambda & b_{18,105} \end{bmatrix} \tag{1}$$

where $b_{l,p}$ is the letter in the position p of the sequence l.

Recall the example discussed in previous section, the common site we want to find has 16 positions (ten L_i's and six ignored letters), a sequence has 90 corresponding sites, so an 18*900 data matrix *D'* is generated from *D*.

$$D' = \begin{bmatrix} d_{1,1}^{1} & \Lambda & d_{1,1}^{10} & d_{1,2}^{1} & \Lambda & d_{1,2}^{10} & \Lambda & d_{1,90}^{1} & \Lambda & d_{1,90}^{10} \\ d_{2,1}^{1} & \Lambda & d_{2,1}^{10} & d_{2,2}^{1} & \Lambda & d_{2,2}^{10} & \Lambda & d_{2,90}^{1} & \Lambda & d_{2,90}^{10} \\ & \mathrm{M} & & & \mathrm{M} & & \mathrm{O} & & \mathrm{M} & \\ d_{18,1}^{1} & \Lambda & d_{18,1}^{10} & d_{18,2}^{1} & \Lambda & d_{18,2}^{10} & \Lambda & d_{18,90}^{1} & \Lambda & d_{18,90}^{10} \end{bmatrix} \qquad (2)$$

where

$$d_{l,s}^{i} = \begin{cases} b_{l,i+s-1} & (for\ \ i = 1,2,...,5) \\ b_{l,i+s+5} & (for\ \ i = 6,7,...,10) \end{cases},$$

and s = 1…90 is the starting position of each candidate site.

For $L_i \in$ {A, T, C, G}, two binary variables u_i and v_i can be used to express L_i, an element of the consensus sequence, as shown in Tab. 1.

Tab. 1 indicates that if L_i is A, T, C, or G respectively, then $a_i = 1$, $t_i = 1$, $c_i = 1$ or $g_i = 1$, which implies following conditions.

$$\begin{aligned} a_i &= (1-u_i)(1-v_i) \\ t_i &= u_i v_i \\ c_i &= (1-u_i)v_i \\ g_i &= u_i(1-v_i) \end{aligned} \qquad (3)$$

Now let $Score_l$ be the degree of fitting to the found common site, specified as

$$Score_l = \sum_{s=1}^{90} z_{l,s}(\theta_{l,s}^{1} + \theta_{l,s}^{2} + ... + \theta_{l,s}^{10}) \qquad (4)$$

where $\theta_{l,s}^{i}$ is the element of candidate sites extracted from *D'*. The constraints associated with (4) are below:

(i) $$\sum_{s=1}^{90} z_{l,s} = 1, \quad z_{l,s} \in \{0,1\} \text{ for all } l \text{ and } s. \qquad (5)$$

(ii) $$\theta_{l,s}^{i} = \begin{cases} a_i & if\ d_{l,s}^{i} = \mathrm{A} \\ t_i & if\ d_{l,s}^{i} = \mathrm{T} \\ c_i & if\ d_{l,s}^{i} = \mathrm{C} \\ g_i & if\ d_{l,s}^{i} = \mathrm{G} \end{cases} \qquad (6)$$

Clearly, $0 \le Score_l \le 10$. And the objective is to maximize the total sum of $Score_l$.

(a)

AAGACTGTTTTTTTGATC
GATTATTTGCACGGCGTC

(b)

$l = 1, s = 1$ $\boxed{\texttt{AAGAC}}\texttt{TGTTTT}\boxed{\texttt{TTTGA}}\texttt{TC}$
$l = 1, s = 2$ $\texttt{A}\boxed{\texttt{AGACT}}\texttt{GTTTTT}\boxed{\texttt{TTGAT}}\texttt{C}$
$l = 1, s = 3$ $\texttt{AA}\boxed{\texttt{GACTG}}\texttt{TTTTTT}\boxed{\texttt{TGATC}}$
$l = 2, s = 1$ $\boxed{\texttt{GATTA}}\texttt{TTTGCA}\boxed{\texttt{CGGCG}}\texttt{TC}$
$l = 2, s = 2$ $\texttt{G}\boxed{\texttt{ATTAT}}\texttt{TTGCAC}\boxed{\texttt{GGCGT}}\texttt{C}$
$l = 2, s = 3$ $\texttt{GA}\boxed{\texttt{TTATT}}\texttt{TGCACG}\boxed{\texttt{GCGTC}}$

(c)

$$\begin{bmatrix} AAGACTTTGA & AGACTTTGAT & GACTGTGATC \\ GATTACGGCG & ATTATGGCGT & TTATTGCGTC \end{bmatrix}$$

Fig. 1. A small example of finding consensus sequence: (a) two sequences to be compared; (b) Schematic representation of the candidate sites; (c) The associated *D'* matrix

Consider the sample data in Fig. 1 for instance:

$$\begin{aligned} Score_1 = {} & z_{1,1}(a_1 + a_2 + g_3 + a_4 + c_5 + t_6 + t_7 + t_8 + g_9 + a_{10}) \\ & + z_{1,2}(a_1 + g_2 + a_3 + c_4 + t_5 + t_6 + t_7 + g_8 + a_9 + t_{10}) \\ & + z_{1,3}(g_1 + a_2 + c_3 + t_4 + g_5 + t_6 + g_7 + a_8 + t_9 + c_{10}) \end{aligned} \quad (7)$$

$$\begin{aligned} Score_2 = {} & z_{2,1}(g_1 + a_2 + t_3 + t_4 + a_5 + c_6 + g_7 + g_8 + c_9 + g_{10}) \\ & + z_{2,2}(a_1 + t_2 + t_3 + a_4 + t_5 + g_6 + g_7 + c_8 + g_9 + t_{10}) \\ & + z_{2,3}(t_1 + t_2 + a_3 + t_4 + t_5 + g_6 + c_7 + g_8 + t_9 + c_{10}) \end{aligned} \quad (8)$$

All $z_{l,s}$ in (4) are binary variables. Equation (5) implies that for a sequence *l*, only one site is chosen and no other sites contribute to $Score_l$. Suppose the *k*'th site is selected, then $z_{l,k} = 1$ and $z_{l,s} = 0$ for all $s \in \{1, 2, ..., 90\}$, $s \neq k$. Since a huge amount of $z_{l,s}$ (i.e, $|l| * |s|$) are involved, to treat $z_{l,s}$ as binary variables would cause a heavy computational burden. Therefore $z_{l,s}$ should be resolved as continuous variables rather than binary variables. An important proposition is introduced below:

Proposition 1 (All or nothing assignment) Let $z_{l,s} \geq 0$ be continuous variables instead of

Tab. 1. Base code in the determined common site

Base	u_i	v_i	a_i	t_i	c_i	g_i
A	0	0	1	0	0	0
T	1	1	0	1	0	0
C	0	1	0	0	1	0
G	1	0	0	0	0	1

binary variables. If there is a k, $k \in \{1,2,...,90\}$, such that $\sum_{i=1}^{10}\theta_{l,k}^{i} = \max\{\sum_{i=1}^{10}\theta_{l,s}^{i} \ for\ s = 1,2,...,90\}$, then assigning $z_{l,k} = 1$ and $z_{l,s} = 0$ for all $s \neq k$, $s \in \{1,2,...,90\}$, can maximize the value of $Score_l$.

Proof Since $\sum_s z_{l,s} = 1$ and $z_{l,s} \geq 0$, it is true that

$$\max\{ \sum_s (z_{l,s} \sum_i \theta_{l,s}^{i}) \} \leq \max\{ \sum_i \theta_{l,s}^{i} \text{ for } s = 1,2,...,90 \} = \sum_i \theta_{l,k}^{i}$$

Remark 1 The objective function of a CSI problem *f*(x) can be rewritten as

$$f(x) = \sum_{i=1}^{10}\{a_i \sum_{(l,s)\in SA_i} z_{l,s} + t_i \sum_{(l,s)\in ST_i} z_{l,s} + c_i \sum_{(l,s)\in SC_i} z_{l,s} + g_i \sum_{(l,s)\in SG_i} z_{l,s}\} \qquad (9)$$

where $SA_i = \{(l,s) \mid d_{l,s}^{i} = A\}$, $ST_i = \{(l,s) \mid d_{l,s}^{i} = T\}$, $SC_i = \{(l,s) \mid d_{l,s}^{i} = C\}$, and $SG_i = \{(l,s) \mid d_{l,s}^{i} = G\}$ for *i*=1,2,…10.

This result implies that SA_i (or ST_i, SC_i, SG_i) is a set composed of (*l, s*) in which the product term $z_{l,s}a_i$ (or $z_{l,s}t_i$, $z_{l,s}c_i$, $z_{l,s}g_i$ respectively) appears on the right hand side of (4) because that $\theta_{l,s}^{i} = a_i$.

For instance, the sum of $Score_1$ and $Score_2$ in (7) and (8) becomes

$$\begin{aligned} Score_1 + Score_2 \ &= \ a_1(z_{1,1} + z_{1,2} + z_{2,2}) + ... + a_{10}z_{1,1} \\ &\quad + ... + g_1(z_{1,3} + z_{2,1}) + ... + g_{10}z_{2,1} \end{aligned} \qquad (10)$$

Some logical constraints can be conveniently expressed by binary variables. For instance, the constraint that a CRP dimer binds a symmetrical site requires that

$$\text{if } L_i = \begin{cases} \text{A} & \text{then } L_{11-i} = \text{T}, \\ \text{C} & \text{then } L_{11-i} = \text{G}. \end{cases}$$

Such a logical structure can be formulated conveniently as the following constraints.

$$\left.\begin{aligned} u_i + u_{11-i} &= 1 \\ v_i + v_{11-i} &= 1 \end{aligned}\right\} \text{for } i = 1,2,3,4,5 \qquad (11)$$

where $u_i, v_i, u_{11-i}, v_{11-i} \in \{0,1\}$.

With reference to Tab. 1, clearly if $L_i = A$ (i.e, $u_i = 0$ and $v_i = 0$) then $L_{11-i} = T$ (i.e, $u_{11-i} = 1$ and $v_{11-i} = 1$) and vice versa; (ii) if $L_i = C$ (i.e, $u_i = 0$ and $v_i = 1$) then $L_{11-i} = G$ (i.e, $u_{11-i} = 1$ and $v_{11-i} = 0$) and vice versa. A CSI problem can then be formulated as a nonlinear mixed 0-1 program below based on these constraints:

Program 1 (Nonlinear 0-1 CSI program)

$$\text{Maximize} \quad \sum_{l=1}^{18} Score_l = \sum_{i=1}^{10}\{ a_i \sum_{(l,s)\in SA_i} z_{l,s} + t_i \sum_{(l,s)\in ST_i} z_{l,s} + c_i \sum_{(l,s)\in SC_i} z_{l,s} + g_i \sum_{(l,s)\in SG_i} z_{l,s} \} \qquad (12)$$

subject to $\sum_{s=1}^{90} z_{l,s} = 1, \quad z_{l,s} \geq 0$ for all l, s

$$\left.\begin{array}{l} a_i = (1-u_i)(1-v_i) \\ t_i = u_i v_i \\ c_i = (1-u_i)v_i \\ g_i = u_i(1-v_i) \end{array}\right\} \begin{array}{l}\text{Conservative constraints} \\ \quad \text{for } i = 1, 2, ..., 10\end{array}$$

$$\left.\begin{array}{l} u_i + u_{11-i} = 1 \\ v_i + v_{11-i} = 1 \end{array}\right\} \begin{array}{l}\text{Logical constraints} \\ \quad \text{for } i = 1, 2, ..., 5\end{array}$$

$$u_i, v_i \in \{0, 1\} \quad \text{for } i = 1, 2, ..., 5$$

$$0 \leq u_i, v_i \leq 1 \quad \text{for } i = 6, 7, ..., 10$$

$$0 \leq a_i, t_i, c_i, g_i \leq 1 \quad \text{for } i = 1, 2, ..., 10$$

This program intends to solve { a_i, t_i, c_i, g_i } for i =1,2, …10 thus to maximize the total degree of fitting to the common site for the given 18 sequences, subjected to a possible logical constraint. A very important feature of Program 1 is that we can treat $z_{l,s}$ as continuous variables rather than binary variables, which can improve the computational efficiency dramatically. We can ensure all found $z_{l,s}$ still have binary values as discussed in the next section.

Linearization of Program 1

Program 1 is a mixed nonlinear 0-1 program where $q_i \sum z_{l,s}$ for $q_i \in \{a_i, t_i, g_i, c_i\}$ and $u_i v_i$ are product terms. These product terms can be linearized directly by the following propositions:

Proposition 2 The product term $\lambda_i = q_i \sum z_{l,s}$ where λ_i is to be maximized and $q_i \in \{0, 1\}$ can be linearized as follows:

$$\begin{array}{l} \lambda_i \geq \sum z_{l,s} + M(q_i - 1) \\ \lambda_i \geq 0 \\ \lambda_i \leq \sum z_{l,s} \\ \lambda_i \leq M\, q_i \end{array} \tag{13}$$

where M is a big constant larger than or equal to the number of sequences.

Proof If q_i= 1 then $\lambda_i = \sum z_{l,s}$; and otherwise λ_i = 0.

Proposition 3 The product term $w_i = u_i v_i$ where $u_i, v_i \in \{0,1\}$ can be linearized as follows:

$$\begin{array}{l} w_i \leq u_i \\ w_i \leq v_i \\ w_i \geq 0 \\ w_i \geq u_i + v_i - 1. \end{array} \tag{14}$$

Denote $Z(a_i) = a_i \sum_{(l,s)\in SA_i} z_{l,s}$, $Z(t_i) = t_i \sum_{(l,s)\in ST_i} z_{l,s}$, $Z(c_i) = c_i \sum_{(l,s)\in SC_i} z_{l,s}$, and $Z(g_i) = g_i \sum_{(l,s)\in SG_i} z_{l,s}$. Program 1 is then linearized into Program 2 below based on Proposition 2 and Proposition 3.

Program 2 (Linear mixed 0-1 CSI program)

$$\text{Maximize} \quad \sum_{l=1}^{18} Score_l = \sum_{i=1}^{10} (Z(a_i) + Z(t_i) + Z(c_i) + Z(g_i)) \qquad (15)$$

$$\text{subject to} \quad \sum_{s=1}^{90} z_{l,s} = 1, \qquad z_{l,s} \geq 0 \quad \text{for all } l, s$$

$$\left.\begin{array}{l} a_i = 1 - u_i - v_i + w_i \\ t_i = w_i \\ c_i = v_i - w_i \\ g_i = u_i - w_i \\ w_i \leq u_i \\ w_i \leq v_i \\ w_i \geq 0 \\ w_i \geq u_i + v_i - 1 \end{array}\right\} \begin{array}{l} \text{Conservative constraints} \\ \text{for } i = 1, 2, ..., 10 \end{array}$$

$$\left.\begin{array}{l} u_i + u_{11-i} = 1 \\ v_i + v_{11-i} = 1 \end{array}\right\} \text{Logical constraints for } i = 1, 2, ..., 5$$

$$\left.\begin{array}{l} \sum_{(l,s)\in SA_i} z_{l,s} + M(a_i - 1) \leq Z(a_i) \leq \sum_{(l,s)\in SA_i} z_{l,s} \\ 0 \leq Z(a_i) \leq M\, a_i \\ \sum_{(l,s)\in ST_i} z_{l,s} + M(t_i - 1) \leq Z(t_i) \leq \sum_{(l,s)\in ST_i} z_{l,s} \\ 0 \leq Z(t_i) \leq M\, t_i \\ \sum_{(l,s)\in SC_i} z_{l,s} + M(c_i - 1) \leq Z(c_i) \leq \sum_{(l,s)\in SC_i} z_{l,s} \\ 0 \leq Z(c_i) \leq M\, c_i \\ \sum_{(l,s)\in SG_i} z_{l,s} + M(g_i - 1) \leq Z(g_i) \leq \sum_{(l,s)\in SG_i} z_{l,s} \\ 0 \leq Z(g_i) \leq M\, g_i \end{array}\right\} \begin{array}{l} \text{Constraints for linearizing} \\ \text{product terms} \end{array}$$

$$u_i, v_i \in \{0, 1\} \quad \text{for } i = 1, 2, ..., 5$$

$$0 \leq u_i, v_i \leq 1 \quad \text{for } i = 6, 7, ..., 10$$

$$0 \leq a_i, t_i, c_i, g_i \leq 1 \quad \text{for } i = 1, 2, ..., 10$$

$z_{l,s}$ ’s are treated as non-negative continuous variables for l =1,2, … ,18 and s =1,2, … ,90 where M can be any value greater than or equal to 18.

In Program 2, since u_i and v_i are binary variables, a_i, t_i, c_i, and g_i should have binary values following (3). Although $z_{l,s}$ are treated as continuous variables, the values of $z_{l,s}$ should be 0 or 1. This is because the optimal solution of a linear program should be a vertex point satisfying $\sum_s z_{l,s} = 1$ for all l.

Consider the following proposition.

Proposition 4 Let the optimal solution of Program 2 be $x^* = (Z^*, u^*, v^*)$ and $\sum_s z_{l,s} = 1$. Assume that a sequence *l* contains sites $s_1, s_2, ..., s_k$ such that $0 < z^*_{l,s_j} < 1$ for *j*=1, 2, … *k*, then,

$$\sum_i \theta^i_{l,s_1} = \sum_i \theta^i_{l,s_2} = ... = \sum_i \theta^i_{l,s_k} = \max\{\sum_i \theta^i_{l,s}\},$$

where θ^i_{l,s_j} are specified in (6).

Proof For $\sum_s z_{l,s} = 1$, if $s_p, s_q \in \{s_1, s_2, ...s_k\}$ where $\sum_i \theta^i_{l,s_p} > \sum_i \theta^i_{l,s_q}$, then to maximize $Score_l = \sum_{l,j} z_{l,s_j} \sum_i \theta^i_{l,s_j}$ requires $z_{l,s_q} = 0$. This conflicts with the observation that $0 < z_{l,s_q} < 1$, therefore $\sum_i \theta^i_{l,s_1} = \sum_i \theta^i_{l,s_2} = ... = \sum_i \theta^i_{l,s_k}$.

After solving Program 2 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:

$$z_{1,64} = z_{2,58} = z_{3,79} = z_{4,66} = z_{5,53} = z_{6,63} = z_{7,27} = z_{8,42} = z_{9,12} = z_{10,17}$$
$$= z_{11,64} = z_{12,44} = z_{13,51} = z_{14,74} = z_{15,20} = z_{16,56} = z_{17,87} = z_{18,81} = 1$$

All other $z_{l,s}$'s have value 0.

In Program 2 the total number of 0-1 variables is 2*m* and the total number of the continuous variables is 20*m*+$|l| * |s|$. Since the number of 0-1 variables is independent of the lengths of *l* and *s*, a CSI problem with many long sequences can be solved effectively.

Suboptimal common sites

Program 2 can find the exact global optimum solution. Sometimes the second best and the third best solution may also be useful. It is very convenient for the proposed method to find a complete set of common sites by adding some extra constraints. For instance, the second best solution of Program 2 can be obtained conveniently by solving the following program:

$$\text{Maximize} \quad \sum_{l=1}^{18} Score_l \qquad (16)$$

subject to (i) The same constraints in Model 1

(ii) $t_1 + g_2 + t_3 + g_4 + a_5 + t_6 + c_7 + a_8 + c_9 + a_{10} \le 9$ (new constraint)

The new constraint is used to force the program to find a new solution different from the solution of Program 2. The found second best common site is "TTTGA□□□□□□ TCAAA" with score 129. Similarly we can find another solution by adding following constraint into (16).

$$t_1 + t_2 + t_3 + g_4 + a_5 + t_6 + c_7 + a_8 + a_9 + a_{10} \le 9$$

The found third best common site is “AAATT□□□□□□AATTT” with score 129.

六、結果與討論

Several experiments are tested here, using the example in the Appendix, to analyze the effect of sequence length and number of sequences on the computational time. All examples are solved by LINGO (Schrage, 1999), a widely used optimization software, on a personal computer with a Pentium 4 2.0G CPU. A software package named “Global Site Seer” is developed based on Program 2 for solving CSI problems. This software is available from http://www.iim.nctu.edu.tw/~cjfu/gss.htm.

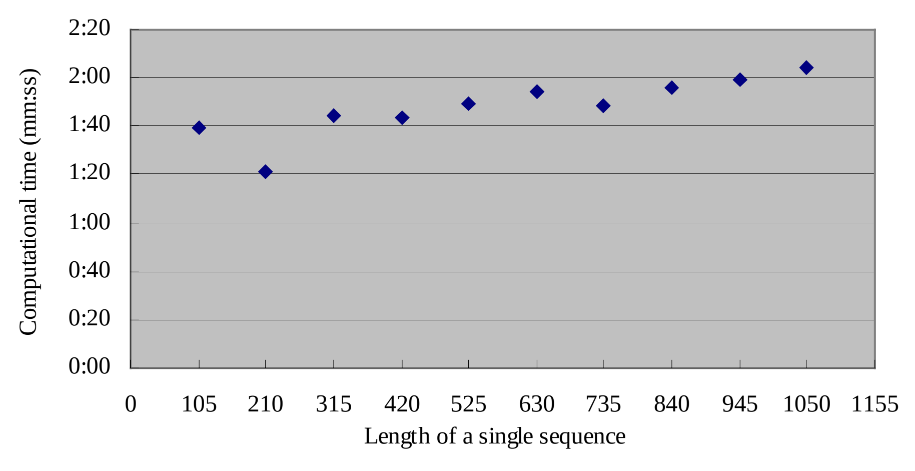
Fig. 2 illustrates the experimental results for analyzing the time complexity. Fig. 2(a) is the computational time given various sequence lengths, where the number of sequences is fixed at 18. The results show that the computational time changes slightly even if the sequence length is increased from 105 to 1050. Fig. 2(b) 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 CSI problems with large sequence length and a large number of sequence number. Fig. 2(c) shows that the computational time rises exponentially as the number of independent positions increases.

This study proposes a linear mixed 0-1 programming approach for solving CSI problems. Comparing with the widely used maximum likelihood methods, the proposed method can reach a global optimum rather than finding a local optimum or a feasible solution. Additionally, by utilizing binary variables some logical constraints can be embedded into the models. It is also convenient to find the complete set of the second, third, etc. best common sites. Since the number of binary variables is fully independent of the number of sequences and the length of a sequence, the proposed method can treat a large CSI problem with many long sequences. For treating a CSI problem with many independent positions in an acceptable time, this study also proposes a method for distributed computing.

Two issues remaining for further study. The first is to extend this method to treat various practical CSI problems. The second is to develop a more refined distributed algorithm to solve some CSI problems by numerous computers via internet.

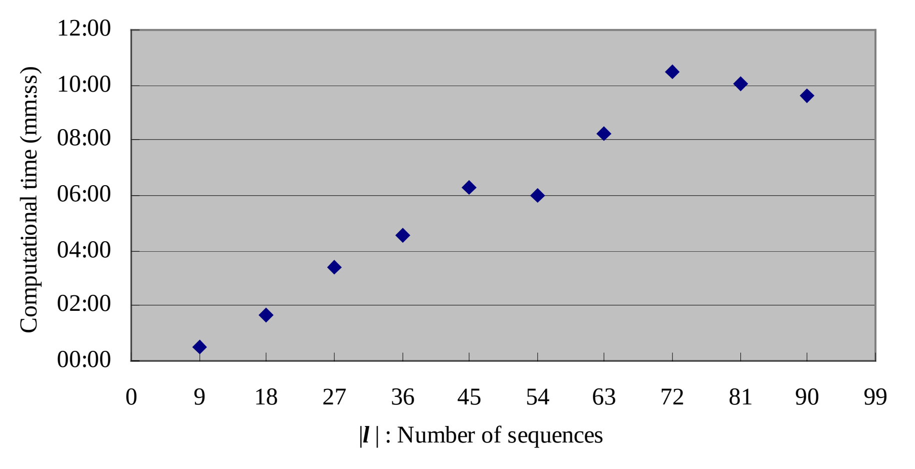
(a) Computational time versus sequence length

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

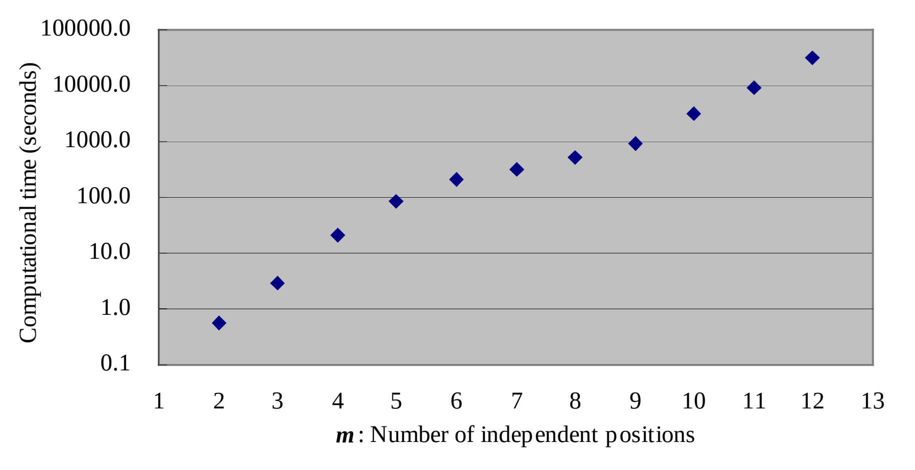


Fig. 2. The relationship between computational time and various factors involved in a CSI problem. This figure illustrates the computational time of solving Program 2 with (a) various sequences sizes; (b) various number of sequences and (c) various independent positions.