EFFICIENT COMPUTATION OF TEMPLATE MATRICES

T. Asano (JAIST), G. Valiente (UPC), F. Rosselló (UIB)

V Jornadas de Matemática Discreta y Algorítmica
Soria, July 12-14 2006

Template matrices

Definition

The template matrix of a matrix M over Σ for the template $ab \in \Sigma^2$, is the matrix $\mathcal{M}_{ab}$ defined by

- $\mathcal{M}_{ab}[j,k] = 1$ if there exists some row i such that $M[i,j] = a$ and $M[i,k] = b$
- $\mathcal{M}_{ab}[j,k] = 0$ otherwise

Template matrices

Definition

The template matrix of a matrix M over Σ for the template $ab \in \Sigma^2$, is the matrix $\mathcal{M}_{ab}$ defined by

- $\mathcal{M}_{ab}[j,k] = 1$ if there exists some row i such that $M[i,j] = a$ and $M[i,k] = b$
- $\mathcal{M}_{ab}[j,k] = 0$ otherwise

Example

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \mathcal{M}_{10} = \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}$$

Template matrices

Definition

The template matrix of a matrix M over Σ for the template $ab \in \Sigma^2$, is the matrix $\mathcal{M}_{ab}$ defined by

- $\mathcal{M}_{ab}[j,k] = 1$ if there exists some row i such that $M[i,j] = a$ and $M[i,k] = b$
- $\mathcal{M}_{ab}[j,k] = 0$ otherwise

Example

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \mathcal{M}_{10} = \begin{pmatrix} 0 & 1 & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}$$

Template matrices

Definition

The template matrix of a matrix M over Σ for the template $ab \in \Sigma^2$, is the matrix $\mathcal{M}_{ab}$ defined by

- $\mathcal{M}_{ab}[j,k] = 1$ if there exists some row i such that $M[i,j] = a$ and $M[i,k] = b$
- $\mathcal{M}_{ab}[j,k] = 0$ otherwise

Example

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathcal{M}_{10} = \begin{pmatrix} 0 & 1 & 1 & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}$$

Template matrices

Definition

The template matrix of a matrix M over Σ for the template $ab \in \Sigma^2$, is the matrix $\mathcal{M}_{ab}$ defined by

- $\mathcal{M}_{ab}[j,k] = 1$ if there exists some row i such that $M[i,j] = a$ and $M[i,k] = b$
- $\mathcal{M}_{ab}[j,k] = 0$ otherwise

Example

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathcal{M}_{10} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}$$

Template matrices

Definition

The template matrix of a matrix M over Σ for the template $ab \in \Sigma^2$, is the matrix $\mathcal{M}_{ab}$ defined by

- $\mathcal{M}_{ab}[j,k] = 1$ if there exists some row i such that $M[i,j] = a$ and $M[i,k] = b$
- $\mathcal{M}_{ab}[j,k] = 0$ otherwise

Example

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathcal{M}_{10} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}$$

Template matrices

Definition

The template matrix of a matrix M over Σ for the template $ab \in \Sigma^2$, is the matrix $\mathcal{M}_{ab}$ defined by

- $\mathcal{M}_{ab}[j,k] = 1$ if there exists some row i such that $M[i,j] = a$ and $M[i,k] = b$
- $\mathcal{M}_{ab}[j,k] = 0$ otherwise

Example

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathcal{M}_{10} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & \\ & & 0 & \\ & & & 0 \end{pmatrix}$$

Template matrices

Definition

The template matrix of a matrix M over Σ for the template $ab \in \Sigma^2$, is the matrix $\mathcal{M}_{ab}$ defined by

- $\mathcal{M}_{ab}[j,k] = 1$ if there exists some row i such that $M[i,j] = a$ and $M[i,k] = b$
- $\mathcal{M}_{ab}[j,k] = 0$ otherwise

Example

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \mathcal{M}_{10} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

And now, the abstract

Computation of template matrices is an intermediate step in the solution of some important problems:

- Perfect phylogeny haplotyping
- Site consistency in perfect phylogeny
- Product of boolean matrices

Efficient computation of template matrices would entail simple efficient solutions to these problems.

And now, the abstract

Computation of template matrices is an intermediate step in the solution of some important problems:

- Perfect phylogeny haplotyping
- Site consistency in perfect phylogeny
- Product of boolean matrices

Efficient computation of template matrices would entail simple efficient solutions to these problems.

The straightforward algorithm computes a template matrix of an $n \times m$ matrix in time $O(nm^2)$.

And now, the abstract

Computation of template matrices is an intermediate step in the solution of some important problems:

- Perfect phylogeny haplotyping
- Site consistency in perfect phylogeny
- Product of boolean matrices

Efficient computation of template matrices would entail simple efficient solutions to these problems.

The straightforward algorithm computes a template matrix of an $n \times m$ matrix in time $O(nm^2)$.

We present an $O(nm^2/\log(n))$ algorithm and we propose a conjecture that would entail an $O(nm + m^2)$ algorithm.

(In)efficient computation of template matrices

Lemma

If M is $n \times m$, then $\mathcal{M}_{ab}$ can be computed in $O(nm^2)$ time.

(In)efficient computation of template matrices

Lemma

If M is $n \times m$, then $\mathcal{M}_{ab}$ can be computed in $O(nm^2)$ time.

Proof.

Use the obvious algorithm:
Traverse each pair j, k of columns until either finding a row i such that $M[i, j] = a$ and $M[i, k] = b$, in which case $\mathcal{M}_{ab}[j, k] = 1$, or exhausting the rows, in which case $\mathcal{M}_{ab}[j, k] = 0$. □

The "exact" cost of this algorithm is the template number for M and ab.

Definition

Let M be an $n \times m$ matrix.

For every $1 \leqslant j, k \leqslant m$ with $j \neq k$ and for every $a, b \in \Sigma$, let

- $T_{ab}^{(j,k)}(M) = \min\{i \mid 1 \leqslant i \leqslant n, M[i,j] = a, M[i,k] = b\}$ if it exists,
- $T_{ab}^{(j,k)}(M) = n$ otherwise.

The template number for M and ab is

$$T_{ab}(M) = \sum_{\substack{1 \leqslant j,k \leqslant m \\ j \neq k}} T_{ab}^{(j,k)}(M).$$

Remark

$T_{ab}(M)$ need no be in $O(mn + m^2)$ (the size of M plus the size of $\mathcal{M}_{ab}$).

Example

Let Id_m be the $m \times m$ diagonal matrix. Since there are no $1 \leqslant i, j, k \leqslant m$ with $j \neq k$ such that $\mathrm{Id}_m[i, j] = \mathrm{Id}_m[i, k] = 1$, we have that

$$T_{11}^{(j,k)} = m \text{ if } j \neq k$$

and then

$$T_{11}(\mathrm{Id}_m) = m(m-1)m \in O(m^3).$$

A more efficient computation of template matrices

Theorem

Let M be an $n \times m$ matrix over Σ. Every template matrix $\mathcal{M}_{ab}$ can be computed in $O(nm^2/\log n)$ time.

The proof uses the four russians algorithm strategy (Arlazarov et al, 1970)

We only consider here the case $a \neq b$. The case $a = b$ is similar.

1 Rename the elements of M: $a \mapsto 0$, $b \mapsto 1$, others $\mapsto 2$ and remove repeated rows

Example

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \qquad \mathcal{M}_{10}?$$

1 Rename the elements of M: $a \mapsto 0$, $b \mapsto 1$, others $\mapsto 2$ and remove repeated rows

Example

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \overset{\text{replacement}}{\Longrightarrow}$$

1 Rename the elements of M: $a \mapsto 0$, $b \mapsto 1$, others $\mapsto 2$ and remove repeated rows

Example

$$M = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Cost: $O(nm)$

2 Define a compressed matrix M^c_{ab} of size $\lceil n/L \rceil \times m$, where $L = \lceil (\log_3 n)/2 \rceil$, by

$$M^c_{ab}[i,j] = \sum_{\ell=0}^{L-1} 3^\ell M_{ab}[iL+\ell, j]$$

Example

$$M = \left(\begin{array}{cccccccccc} 0&1&1&0&0&0&1&0&1&1 \\ 0&1&1&1&0&1&0&1&1&1 \\ \hline 1&0&0&1&1&1&1&1&0&0 \\ 1&1&1&1&0&1&1&1&1&1 \\ \hline 1&1&1&0&0&0&1&1&1&1 \\ 1&0&1&1&0&1&0&1&1&1 \\ \hline 1&1&1&1&0&1&0&1&1&1 \end{array}\right) \xrightarrow{\text{compress}}$$

2 Define a compressed matrix M^c_{ab} of size $\lceil n/L \rceil \times m$, where $L = \lceil (\log_3 n)/2 \rceil$, by

$$M^c_{ab}[i,j] = \sum_{\ell=0}^{L-1} 3^\ell M_{ab}[iL+\ell, j]$$

Example

$$M^c_{10} = \begin{pmatrix} 0 & 4 & 4 & 3 & 0 & 3 & 1 & 3 & 4 & 4 \\ 4 & 3 & 3 & 4 & 1 & 4 & 4 & 4 & 3 & 3 \\ 4 & 1 & 4 & 3 & 0 & 3 & 1 & 4 & 4 & 4 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Cost: $O(nm)$

3 Define a matrix R of size $3^L \times 3^L$ by $R[p,q] := b_0 + 2b_1$, where

$$b_0 = \begin{cases} 1 & \text{if } I_{p-1}[i] = 0 \text{ and } I_{q-1}[i] = 1 \text{ for some } i \\ 0 & \text{otherwise} \end{cases}$$

$$b_1 = \begin{cases} 1 & \text{if } I_{p-1}[i] = 1 \text{ and } I_{q-1}[i] = 0 \text{ for some } i \\ 0 & \text{otherwise} \end{cases}$$

Example

$$R = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 2 & 0 & 0 & 3 & 1 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 3 & 2 & 0 & 1 & 0 & 0 & 1 & 0 \\ 2 & 2 & 2 & 2 & 0 & 0 & 2 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Cost: $O(L3^{2L}) = O(n \log(n))$

4 Compute $\mathcal{M}_{ab}$ from M^c_{ab} and R by: For $1 \leqslant i < j \leqslant m$

- If $R(M^c_{ab}[k,i], M^c_{ab}[k,j]) = 1$ or 3 for some k, then $\mathcal{M}_{ab}[i,j] = 1$
- If $R(M^c_{ab}[k,i], M^c_{ab}[k,j]) = 2$ or 3 for some k, then $\mathcal{M}_{ab}[j,i] = 1$
- In all other cases $\mathcal{M}_{ab}$ entries are 0

Example

$$\mathcal{M}_{10} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{pmatrix}$$

Cost: $O(nm^2/L) = O(nm^2/\log(n))$

A conjectured efficient computation of template matrices

Definition

Given an $n \times m$ matrix M over an alphabet Σ, for every $1 \leqslant j \leqslant m$ and for every $a \in \Sigma$, let $M_{j,a}$ be the ordered set of maximal row intervals $[i_1, i_2]$ such that $M[i_1, j] = \cdots = M[i_2, j] = a$.

For every $[i_1, i_2] \in M_{j,a}$, let $\rho^j_a(i_1, i_2)$ be the position of $[i_1, i_2]$ in $M_{j,a}$.

Definition (continued)

For every $1 \leqslant j, k \leqslant m$ *with* $j \neq k$, *for every* $a, b \in \Sigma$, *let*

$$R_{ab}^{(j,k)}(M) = \begin{cases} \min\Big\{ \rho_a^j(i_1, i_2) + \rho_b^k(i_3, i_4) \mid \\ \qquad [i_1, i_2] \in M_{j,a}, [i_3, i_4] \in M_{k,b}, \\ \qquad [i_1, i_2] \cap [i_3, i_4] \neq \emptyset \Big\} & \text{if it exists} \\ |M_{j,a}| + |M_{k,b}| & \text{otherwise} \end{cases}$$

The reduced template number for M *and* ab *is*

$$R_{ab}(M) = \sum_{\substack{1 \leqslant j,k \leqslant m \\ j \neq k}} R_{ab}^{(j,k)}(M).$$

Theorem

A template matrix $\mathcal{M}_{ab}$ derived from M can be computed in time $R_{ab}(M) + nm$.

Proof.

The lists of intervals $(M_{j,a})_{1 \leqslant j \leqslant m}$, $(M_{j,b})_{1 \leqslant j \leqslant m}$ can be constructed in $O(nm)$ time, by traversing M in column order.
Once they are available, we enumerate the set

$$P_{ab} = \{(j,k) \mid M[i,j] = a,\ M[i,k] = b \text{ for some } i\}$$

by performing, for each j, k, a simultaneous traversal of $M_{j,a}$ and $M_{k,b}$ until two intervals intersect.
This takes time $R_{ab}^{(j,k)}(M)$ for every j, k. □

Theorem

A template matrix $\mathcal{M}_{ab}$ derived from M can be computed in time $R_{ab}(M) + nm$.

Example

$$M = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathcal{M}_{10}[1,2] = ?$$

Theorem

A template matrix $\mathcal{M}_{ab}$ derived from M can be computed in time $R_{ab}(M) + nm$.

Example

$$M = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathcal{M}_{10}[1,2] = ?$$

Theorem

A template matrix $\mathcal{M}_{ab}$ derived from M can be computed in time $R_{ab}(M) + nm$.

Example

$$M = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathcal{M}_{10}[1,2] = ?$$

Theorem

A template matrix $\mathcal{M}_{ab}$ derived from M can be computed in time $R_{ab}(M) + nm$.

Example

$$M = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathcal{M}_{10}[1,2] = ?$$

Theorem

A template matrix $\mathcal{M}_{ab}$ derived from M can be computed in time $R_{ab}(M) + nm$.

Example

$$M = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathcal{M}_{10}[1,2] = 1$$

Definition

An $n \times m$ matrix M over an alphabet Σ is canonical if it has no duplicate rows or columns and its rows are sorted in lexicographical order.

Remark

Any matrix M over an alphabet Σ can be resolved into a canonical matrix M', by removing duplicate rows and columns and sorting the rows in lexicographical order, in $O(nm)$ time by radix sorting techniques.

If the original matrix M does not have any duplicate columns, $\mathcal{M}_{ab} = \mathcal{M}'_{ab}$ for all $a, b \in \Sigma$.

Conjecture

For every $n \times m$ canonical matrix M over an alphabet Σ and for every $a, b \in \Sigma$, $R_{ab}(M) \in O(nm + m^2)$.

Evidence:

- We have computed explicitly $R_{ab}(M)$ for several families of matrices with $T_{ab}(M)$ in $O(nm^2)$, and they always grow in $O(nm + m^2)$.

Conjecture

For every $n \times m$ canonical matrix M over an alphabet Σ and for every $a, b \in \Sigma$, $R_{ab}(M) \in O(nm + m^2)$.

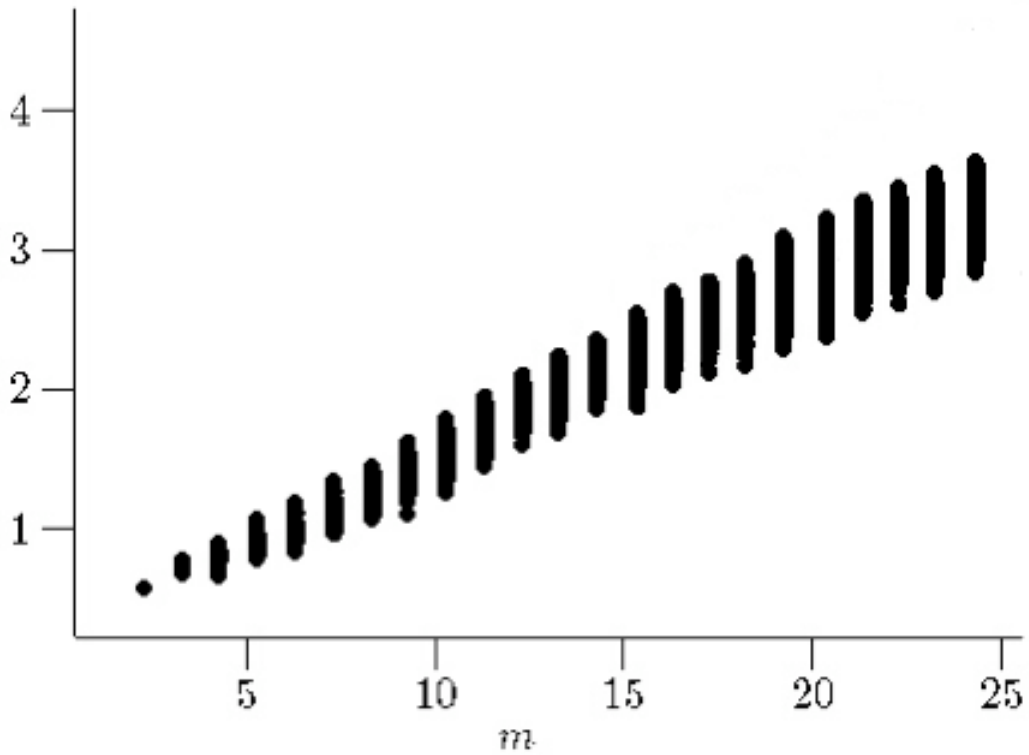
Evidence:

- We have computed the reduced template numbers for large set of random binary matrices, sorting them in lexicographical order by rows before computing the reduced template numbers. Their reduced template numbers are bounded by $O(nm + m^2)$, with a multiplicative constant in the asymptotic upper bound that grows very slowly relative to the input size.

Conjecture

For every $n \times m$ canonical matrix M over an alphabet Σ and for every $a, b \in \Sigma$, $R_{ab}(M) \in O(nm + m^2)$.

Typical growth of $R_{ab}(M)/(nm + m^2)$ in these experiments



Conjecture

For every $n \times m$ canonical matrix M over an alphabet Σ and for every $a, b \in \Sigma$, $R_{ab}(M) \in O(nm + m^2)$.

No clue about how to attack this problem. Any idea?

Applications of template matrices

Computation of template matrices is the bottleneck of simple algorithms for

- Perfect phylogeny haplotyping
- Site consistency in perfect phylogeny
- Product of boolean matrices

Perfect Phylogeny Haplotyping (PPH)

Definition

A genotype matrix is an $n \times m$ matrix M over the alphabet $\{0, 1, 2\}$, where the ith row $M[i, *]$ describes the genotype of species s_i, each column $M[*, j]$ represents a polymorphic locus, and each column j for which $M[i, j] = 2$ is a polymorphic site.

Problem

Given a genotype matrix M, find a binary $2n \times m$ haplotype matrix where each genotype row $M[i,*]$ expands to two rows $M'[i,*]$ and $M'[i',*]$, in such a way that:

- $M'[i,j] = M'[i',j'] = M[i,j]$ *if* $M[i,j] \in \{0,1\}$;
- $M'[i,j] \neq M'[i',j']$ *if* j $M[i,j] = 2$;
- M' *admits a perfect phylogeny: there do not exist columns* j_1, j_2 *such that* $M'[*, \{j_1, j_2\}]$ *contains each one of the rows* $00, 01, 10, 11$.

or decide that it does not exist.

Let M be a genotype matrix.

- Columns j, k are companion columns if there is a companion row i such that $M[i, j] = M[i, k] = 2$.
- Two companion columns j, k are forced in-phase if the expansion of the companion row in $M'[*, \{j, k\}]$ contains $\{00, 11\}$ and out-of-phase if it contains $\{01, 10\}$.
- The genotype graph for M has one node for each column in M, one edge in E_f for each pair of companion columns of M that are either forced in-phase or out-of-phase, and one edge in E_n for each pair of companion columns of M that are not forced in-phase or out-of-phase.

From the genotype graph, the haplotype matrix of a $n \times m$ genotype matrix M can be computed (or decided that it does not exist) in time $O(nm + m^2)$. (Bafna et al, JCB 04)

Lemma

The forced edges E_f in the genotype graph for M are given by

$$\mathcal{M}_{22} \cap (((\mathcal{M}_{00} \cup \mathcal{M}_{20} \cup \mathcal{M}_{02}) \cap (\mathcal{M}_{11} \cup \mathcal{M}_{21} \cup \mathcal{M}_{12})) \cup ((\mathcal{M}_{10} \cup \mathcal{M}_{20}) \cap (\mathcal{M}_{01} \cup \mathcal{M}_{02})))$$

and the non-forced edges E_n are given by $\mathcal{M}_{22} \setminus E_f$.

Corollary

Efficient computation of template matrices would entail a simple efficient solution for PPH.

An alternative efficient solution was proposed recently (Ding et al, RECOMB05)

Site consistency in perfect phylogeny (SCPP)

Problem

Resolve a given genomic (binary) matrix M into another that admits a perfect phylogeny by removing the least number of (segregating) columns.

Definition

The conflict graph for an $n \times m$ genomic matrix M over the alphabet $\Sigma = \{0, 1\}$ has one node for each column in M and one edge $\{j, k\}$ for each pair of columns j and k such that there is a row i with $M[i, j] = M[i, k] = 0$, a row i with $M[i, j] = 0$ and $M[i, k] = 1$, a row i with $M[i, j] = 1$ and $M[i, k] = 0$, and a row i with $M[i, j] = M[i, k] = 1$.

From the conflict graph, the SCPP problem for a $n \times m$ genomic matrix M that can be derived in a galley tree can be solved in time $O(nm + m^2)$. (Asano et al, 04)

(In general, it is NP-hard)

From the conflict graph, the SCPP problem for a $n \times m$ genomic matrix M that can be derived in a galley tree can be solved in time $O(nm + m^2)$. (Asano et al, 04)
(In general, it is NP-hard)

Lemma

The edges in the conflict graph for a given genomic matrix M are given by $\mathcal{M}_{01} \cap \mathcal{M}_{10} \cap \mathcal{M}_{11}$.

Products of boolean matrices

Proposition

Let $T(n, m)$ be the time needed to construct $\mathcal{M}_{ab}$ derived from an $n \times m$ matrix M.
Two boolean matrices with dimensions $m_1 \times n$ and $n \times m_2$, respectively, can be multiplied in $O(T(n, m_1 + m_2))$ time.

Products of boolean matrices

Proposition

Let $T(n, m)$ be the time needed to construct $\mathcal{M}_{ab}$ derived from an $n \times m$ matrix M.
Two boolean matrices with dimensions $m_1 \times n$ and $n \times m_2$, respectively, can be multiplied in $O(T(n, m_1 + m_2))$ time.

Proof.

Given two boolean matrices X and Y with dimensions $m_1 \times n$ and $n \times m_2$, respectively, let A be the matrix obtained from X by replacing each 1 by a, and let B be the matrix obtained from Y by replacing each 1 by b. Let also M be the $n \times (m_1 + m_2)$ matrix $M = [A^T B]$ and let Z be the $m_1 \times m_2$ matrix $Z = X \cdot Y$. Then,

$$\begin{aligned} Z[j,k] = 1 \quad &\text{iff} \quad X[j,i] = 1 \text{ and } Y[i,k] = 1 \text{ for some } 1 \leqslant i \leqslant n \\ &\text{iff} \quad A[j,i] = a \text{ and } B[i,k] = b \\ &\text{iff} \quad M[i,j] = a \text{ and } M[i, m_1 + k] = b \\ &\text{iff} \quad \mathcal{M}_{ab}[j, m_1 + k] = 1\,. \end{aligned}$$

Navigation