Discrete Mathematics and Theoretical Computer Science **(subm.)**, by the authors, 1–rev

Analysis of Biclusters with Applications to Gene Expression Data[†]

Gahyun Park[1] and Wojciech Szpankowski[1]

[1]*Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA.*

For a given matrix of size $n \times m$ over a finite alphabet $\mathcal{A}$, a bicluster is a submatrix composed of selected columns and rows satisfying a certain property. In microarrays analysis one searches for largest biclusters in which selected rows constitute the same string (pattern); in another formulation of the problem one tries to find a maximally dense submatrix. In a conceptually similar problem, namely the bipartite clique problem on graphs, one looks for the largest binary submatrix with all '1'. In this paper, we assume that the original matrix is generated by a memoryless source over a finite alphabet $\mathcal{A}$. We first consider the case where the selected biclusters are square submatrices and prove that with high probability (whp) the largest (square) bicluster having the same row-pattern is of size $\log_Q^2 nm$ where Q^{-1} is the (largest) probability of a symbol. We observe, however, that when we consider *any* submatrices (not just *square* submatrices), then the largest area of a bicluster jumps to An (whp) where A is an explicitly computable constant. These findings complete some recent results concerning maximal biclusters and maximum balanced bicliques for random bipartite graphs.

Keywords: Random matrix, two-dimensional patterns, bicluster, microarray data, biclique.

---

1 Introduction

Clustering is a well known algorithmic technique that partitions a set of input data (vectors) into subsets such that data in the same subset are close to one another in some metric. Recent developments in computational biology, in particular, microarray analysis, have commenced a resurgence of interest in the problem of finding hidden structures in large matrices (cf. Wang et al. (2002); Tanay et al. (2002)). In particular, one is interested in determining the similarity in a *subset* of higher dimensions (subset that has to be determined as well). This problem is known as *biclustering*. More formally, it can be defined as the problem of finding a partition of the vectors and a subset of the dimensions such that the projections along those directions of the vectors in each cluster are close to one another.

In many applications often the problem of interest is to find a bicluster (satisfying some additional property) with the largest area. For example, in gene expression data one searches for the largest submatrix with the property that all rows have the same pattern (i.e., are the same). This problem is known to be NP-complete. Therefore, there is interest in finding efficient heuristics to find large biclusters (cf. Lonardi

[†]This research was supported by NSF Grant CCR-0208709, the NIH grant R01 GM068959-01, and AFOSR Grant FA8655-04-1-3074.

et al. (2004); Q. Sheng and Moor (2003); Tanay et al. (2002); Wu et al. (2004)). Once a bicluster is found, one must assess its (statistical) significance, which can be accomplished by comparing it to an atypical bicluster found in a reference model. In this paper, we assume that the reference model is a probabilistic model in which a matrix $n \times m$ over a finite alphabet $\mathcal{A}$ of size V is generated by a memoryless source. Our goal is to determine the size of a *typical* largest submatrix (bicluster) consisting of the same rows. We shall argue then that any bicluster found in the *real* model (e.g., microarray) is statistically insignificant if its size is comparable to the typical size of the largest bicluster found in the reference (probabilistic) model.

Before we present our main results, let us further motivate our study. Biclustering has important applications in several areas, such as data mining, machine learning, computational biology, and pattern recognition. Data arising from text analysis, market-basket data analysis, web logs, etc., is usually arranged in a contingency table or co-occurrence table, such as, a word-document table, a product-user table, a CPU-job table or a webpage-user table. In computational biology, this problem is associated with the analysis of gene expression data obtained from microarray experiments. Biclustering of gene expression data is a promising methodology for identification of groups of genes that exhibit a coherent pattern across a subset of conditions. A number of algorithms have been introduced for extracting biclusters in microarray data (cf. Lonardi et al. (2004); Q. Sheng and Moor (2003); Wu et al. (2004)). Once the gene expression matrix is discretized into a matrix over a finite alphabet, finding similar gene behavior across a subset of conditions resembles the problem of finding subsequences having the same alphabetic expressions in sequence data. Therefore, the biclustering problem reduces to the problem of finding a subset of the rows and a subset of columns such that the submatrix induced has the property that each row reads the same string.

Now, we present our main results. As mentioned above, we consider biclustering in a probabilistic framework in which a memoryless source generates an $n \times m$ random matrix over a finite alphabet $\mathcal{A}$. We denote by p_i the probability of generating symbol $i \in \mathcal{A}$, where $1 \le i \le V = |\mathcal{A}|$. Let also $p_{\max} = \max\{p_1, \ldots, p_V\}$ and $Q = p_{\max}^{-1}$. We first prove that the largest *square* submatrix having the same rows (i.e., every row has the same pattern, for example *ababb*) is of size $\log_Q^2 nm$ with high probability (whp). When we relax the restriction of square submatrices, then we observe an interesting phenomenon. The largest biclusters are either of area An or of area Bm when one side of a bicluster is $O(n)$ (or $O(m)$) and the other $O(1)$. In this case, the largest submatrix is very "skinny" and in fact achieves the largest possible area among all biclusters. These results complete the study of Lonardi et al. (2004) who only were able to prove an upper bound of the size of largest subclusters. In fact, our results can be viewed as a direct generalization of Darling and Waterman (1985) who analyzed largest consecutive subblocks in a random matrix.

Moreover, our results directly imply a typical behavior of the maximum biclique in a random bipartite graph and complete the study of Dawande et al. (2001). Indeed, Dawande et al. (2001) studied the maximum biclique problem in a bipartite graph. The adjacency matrix of a bipartite graph $G = (V_1, V_2, E)$ with $|V_1| = n$ and $|V_2| = m$ is a matrix $M \in \{0,1\}^{n \times m}$. An edge $(i, j) \in E$ connects node $i \in V_1$ to node $j \in V_2$ if $M_{i,j} = 1$. Thus, a submatrix of 1's in M corresponds to a subgraph of G which is completely connected. A *biclique* $C = U_1 \cup U_2$, $U_1 \subseteq V_1, U_2 \subseteq V_2$ is such a subgraph. The largest bicluster problem with $\mathcal{A} = \{0,1\}$ is called *maximum edge biclique problem*, which is known to be NP-complete (cf. Peters (2000)). The problem requires finding a biclique having the maximum number of edges. A restricted version of this problem, where there is an additional requirement that $|U_1| = |U_2|$, is called the *maximum balanced biclique problem* (MBBP), which is also NP-complete (cf. Garey and Johnson (1979)). Finding the largest square biclique with a binary alphabet reduces to the maximum balanced biclique problem (MBBP). Dawande et al. (2001) proved an upper bound on the size of the maximum balance biclique when the probability of generating an edge is p. In this paper, as a side result, we establish that the size of such a maximum balance biclique is $\log nm / \log(p^{-1})$ with high probability.

The paper is organized as follows. In the next section we present our main results and their consequences. Proofs are delayed till the last section.

2 Main Results

Throughout, we assume that a memoryless source generates an $n \times m$ matrix $M = \{m_{ij}\}_{i,j \in \mathcal{A}}$ over a finite alphabet $\mathcal{A}$ and that $m = O(n^k)$ and $n = O(m^k)$ for a constant $k > 0$. The probability of generating symbol $i \in \mathcal{A}$ is equal to p_i. By M_{IJ} we denote a submatrix of M consisting of a rows $I = \{i_1, i_2, \ldots, i_a\}$ such that $1 \le i_1 < \ldots < i_a \le n$ and b columns $J = \{j_1, \ldots, j_b\}$ such that $1 \le j_1 < \ldots < j_b \le m$. We say a submatrix M_{IJ} is *matching* if each row of M_{IJ} is the same, that is, each row consists of identical strings over $\mathcal{A}$ (see also (1) below). In particular, a column vector is matching if it consists of the same symbol. This is illustrated in the next example.

Example 2.1 *The* 5×5 *matrix* M *over* $\mathcal{A} = \{x, y, z, w\}$ *is*

$$\begin{bmatrix} x & x & y & w & z \\ w & z & y & y & w \\ w & z & y & y & w \\ w & z & x & y & x \\ x & z & w & y & w \end{bmatrix}.$$

The following two (square) submatrices

$$M_{2,3,5;2,4,5} = \begin{bmatrix} z & y & w \\ z & y & w \\ z & y & w \end{bmatrix},$$

and

$$M_{2,3,4;1,2,4} = \begin{bmatrix} w & z & y \\ w & z & y \\ w & z & y \end{bmatrix}$$

are composed of three identical rows. In fact, they are the largest square submatrices satisfying this property. But these square submatrices are not the largest (general) submatrices with this property. For example,

$$M_{2,3;1,2,3,4,5} = \begin{bmatrix} w & z & y & y & w \\ w & z & y & y & w \end{bmatrix}$$

is the largest bicluster, with the area ten.

Let M_{IJ} be a matching submatrix of $M_{n,m} \in \mathcal{A}^{n \times m}$. The problem addressed in this paper is defined as follows.

LARGEST BICLUSTER(f): problem **Instance:** A matrix $M \in \mathcal{A}^{n\times m}$ over the alphabet $\mathcal{A}$. **Question:** Find a row selection I and a column selection J such that the rows of M_{IJ} are identical strings and an objective function $f(M_{IJ})$ is maximized.

Applications determine what objective functions are to be considered. In this paper, following the microarrays approach we assume that $f(M_{IJ}) = |I||J|$ with or without any restriction on the shape of the submatrix. That is, we maximize the area of matching submatrices.

More formally, our goal is to find

$$W_{n,m} = \max_{I,J}\{|I||J| = a\cdot b:\ m_{i_1,j_k} = \ldots = m_{i_a,j_k},\ \ 1\le k\le b,\ \ m_{ij}\in\mathcal{A}\}. \tag{1}$$

Now we are in the position to formulate our main results. We first consider *square* submatrices M_{IJ} ($|I| = |J|$). The next theorem will be proved in the next section.

Theorem 2.1 *Let $M \in \mathcal{A}^{n\times m}$ be a random matrix generated by a memoryless source. Let $P_a = p_1^a + \cdots + p_V^a$ and $Q = p_{max}^{-1}$, where $p_{max} = \max\{p_1,\ldots,p_V\}$. The area of the largest square matching submatrix M_{IJ} is (whp)*

$$W_{n,m} = \log_Q^2 nm \quad (pr.).$$

More precisely, for any $\varepsilon > 0$,

$$\Pr\left\{(1-\varepsilon)\log_Q^2 nm \le W_{n,m} \le (1+\varepsilon)\log_Q^2 nm\right\} = 1 - O\left(\frac{\log^6 nm}{nm}\right)$$

for sufficiently large n and m.

Interestingly enough, if we drop the restriction of the shape of the largest submatrix (i.e., $|I| = |J|$), then the largest biclusters are of order $O(m)$ (or $O(n)$) and very skinny, as observed in practice by Lonardi et al. (2004).

Theorem 2.2 *We adopt the same assumption as in Theorem 2.1. The area of the largest matching submatrix M_{IJ} is whp either*

$$W_{n,m} = a^* P_{a^*} m \tag{2}$$

or

$$W_{n,m} = b^* p_{\max}^{b^*} n \tag{3}$$

where a^* is the integer r that maximizes $rP_r = r(p_1^r + \cdots + p_V^r)$ and $b^* = \max\{1, \lfloor 1/\log Q\rfloor\}$. If (2) holds, then M_{IJ} consists of a^* rows and $\lfloor P_{a^*} m\rfloor$ columns; if (3) holds, then M_{IJ} consists of symbols having the maximum emitting probability, and $|I| = p_{\max}^{b^*} n$ while $|J| = b^*$.

Remark: In the maximum biclique problem on random graphs, as discussed in Dawande et al. (2001), it is assumed that an edge is generated with probability p. The corresponding adjacency matrix M is over a binary alphabet $\{0,1\}$ with 1 indicating an edge. Our previous results are directly applicable to the maximum biclique problem with Q replaced by $1/p$.

3 Analysis

In this section we prove our main results. Throughout we use the first and the second moment methods (cf. Szpankowski (2001)).

3.1 Proof of Theorem 2.1

We now prove Theorem 2.1. To simplify our presentation, we first consider a square matrix M, that is, we assume $m = n$. We search for the largest square submatrix.

We first establish an upper bound by proving that the largest area $W_{n,n}$ is at most $(2(1+\varepsilon)\log_Q n)^2$ with high probability for any $\varepsilon > 0$. Let $a = \lceil 2(1+\varepsilon)\log_Q n \rceil$. Using Boole's inequality we find

$$\Pr\{W_{n,n} \geq a^2\} \leq \binom{n}{a}^2 P_a^a.$$

Since $\binom{n}{a} \leq n^a$ and $P_a^a \leq (V p_{max}^a)^a$, we have

$$\begin{aligned}\Pr\{W_{n,n} \geq a^2\} &\leq V^a \exp(2a\log n - a^2 \log Q)\\ &\leq V^a \exp[a(2\log n - 2(1+\varepsilon)\log n)]\\ &\leq V^a \exp(-2\varepsilon a \log n) = \left(\frac{V}{n^{2\varepsilon}}\right)^a \to 0\end{aligned}$$

Now, we establish the matching lower bound. To begin with we let $a = \lfloor 2(1-\varepsilon)\log_Q n \rfloor$ and define a random variable X_{IJ} to be $|I||J|$ if M_{IJ} is a matching submatrix of M and 0 otherwise. Then we can write $W_{n,m} = \max_{I,J}\{X_{IJ}\}$. Throughout, we use the following notations.

$$\begin{aligned}A_{IJ} &= \{X_{IJ} = a^2\},\\ S_1 &= \sum_{I,J} \Pr\{A_{IJ}\},\\ S_2 &= \sum_{(I,J)\neq(L,M)} \Pr\{A_{IJ} \cap A_{LM}\}.\end{aligned}$$

Note that

$$\{W_{n,n} \geq a^2\} = \bigcup_{|I|=|J|=a} A_{IJ}.$$

Here we utilize Chung and Erdős second moment method (cf. Szpankowski (2001)) and obtain

$$\Pr\{W_{n,n} \geq a^2\} \geq \frac{S_1^2}{S_1 + S_2}.$$

We begin by estimating S_1. Observe that

$$\binom{n}{a} = \left(\frac{n}{2\pi a(n-a)}\right)^{1/2} \exp\left[nH\left(\frac{a}{n}\right) + O\left(\frac{1}{a} + \frac{1}{n-a}\right)\right], \tag{4}$$

where $H(x) = -x\log x - (1-x)\log(1-x)$ is the binary entropy. Therefore, we have

$$\binom{n}{a} \geq \frac{C_1}{\sqrt{a}}\exp\left(a\log\frac{n}{a}\right),$$

for some constant $C_1 > 0$. Since $P_a^a \geq p_{max}^{a^2}$, we obtain

$$\begin{aligned} S_1 &= \binom{n}{a}^2 P_a^a \\ &\geq \frac{C_1^2}{a}\exp[2a\log\frac{n}{a} - a^2\log Q] \\ &= \frac{C_1^2}{a}\exp[a(2\log\frac{n}{a} - a\log Q)] \\ &\geq \frac{C_1^2}{a}\exp\left\{a\left(2\log\frac{n}{2\log_Q n} - 2(1-\varepsilon)\log n\right)\right\} \geq C_2 n^{\varepsilon a} \to \infty \end{aligned}$$

for some constant C_2.

Next, we show that $\frac{S_2}{S_1^2} \to 1$. We split S_2 into two terms $S_2^{'}$ and $S_2^{''}$; $S_2^{'}$ sums over such (I,J) and (L,M) that $M_{IJ} \cap M_{LM} = \emptyset$, and $S_2^{''}$ covers the rest. Notice that M_{IJ} and M_{LM} are independent when the indices are in $S_2^{'}$. It is easy to estimate that

$$S_2^{'} \leq S_1^2.$$

More work is needed to assess $S_2^{''}$. We define (r,c)*-overlap* when r rows overlap between I and L rows, and c columns overlap between J and M columns . Define $N(r,c)$ and $P(r,c)$ as the number of (r,c)-overlaps and the probability of an (r,c)-overlap, respectively. Then

$$S_2^{''} = \sum_{\substack{1\leq r,c\leq a \\ (r,c)\neq(a,a)}} N(r,c)P(r,c).$$

Using the fact that $\binom{n-a}{a-r} \leq \binom{n}{a}\left(\frac{a}{n}\right)^r$ and $\binom{a}{r} \leq a^r$, we find

$$\begin{aligned} N(r,c) &= \binom{n}{a}\binom{n-a}{a-r}\binom{a}{r}\cdot\binom{n}{a}\binom{n-a}{a-c}\binom{a}{c} \\ &\leq \binom{n}{a}^4 n^{-r-c}a^{2r+2c}. \end{aligned}$$

Let us now estimate the probability $P(r,c)$. Observe that in the (r,c) overlap, there are $2a-r$ rows of c common columns that contribute P_{2a-r}^c, and then there are two subsets of $a-c$ columns contributing P_a^{a-c} (cf. Figure 1) . Thus

$$P(r,c) = P_{2a-r}^c P_a^{a-c} P_a^{a-c}.$$

Using the known inequality $P_b^{1/b} \leq P_a^{1/a}$ for $b \geq a > 0$ (Lemma 4.4 of Szpankowski (2001)) we arrive at

$$\begin{aligned} P(r,c) &= P_{2a-r}^c P_a^{a-c} P_a^{a-c} \\ &\leq P_a^{\frac{(2a-r)c}{a}} P_a^{2a-2c} = P_a^{2a-\frac{rc}{a}} \end{aligned}$$

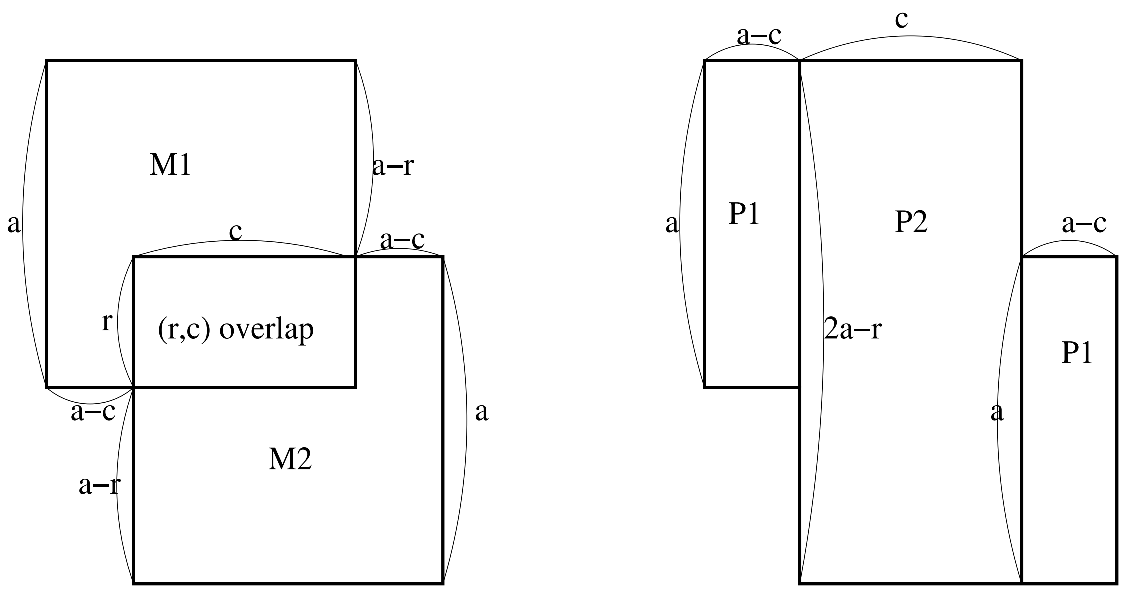

Fig. 1: An (r,c)-overlap of two $a \times a$ sub-matrices $M1$ and $M2$ (left figure); $P(r,c) = P1P2P1$ where $P1 = P_a^{a-c}$ and $P2 = P_{2a-r}^{c}$ (right figure).

Therefore, we have

$$\frac{S_2''}{S_1^2} \leq \sum_{1 \leq r,c \leq a} \frac{a^{2r+2c}}{n^{r+c}} P_a^{-\frac{rc}{a}} \leq \sum_{1 \leq r,c \leq a} \frac{a^{2r+2c}}{n^{r+c}} Q^{rc}. \tag{5}$$

Since $\frac{a^{2r+2c}}{n^{r+c}} Q^{rc} \leq \frac{a^4}{n^2} Q$, for $1 \leq r,c \leq a$, we have

$$\frac{S_2''}{S_1^2} \leq \sum_{1 \leq r,c \leq a} \frac{a^4 Q}{n^2} = \frac{a^6 Q}{n^2} = O\left(\frac{\log^6 n}{n^2}\right) \tag{6}$$

In summary,

$$\begin{aligned}
\Pr\{W_{n,n} \geq a * a\} &\geq \frac{S_1^2}{S_1 + S_2' + S_2''} \\
&= \frac{1}{\frac{1}{S_1} + \frac{S_2'}{S_1^2} + \frac{S_2''}{S_1^2}} \\
&\geq \frac{1}{O(n^{-\varepsilon a}) + 1 + O(n^{-2}\log^6 n)} \\
&= 1 - O\left(\frac{\log^6 n}{n^2}\right).
\end{aligned}$$

This complete the proof for $n = m$.

Finally, we extend our analysis to $m \neq n$ case. In this case we can easily modify the previous method and show that the optimal size of a square submatrix is $a \cdot a$, where $a = \log_Q nm$.

3.2 Proof of Theorem 2.2

Here we prove Theorem 2.2. Since a proof of (3) is similar to (2), we only present a detailed proof of (2).

Proof of (2): First we let $b = \beta m$ for some $0 < \beta < 1$ and $a = O(1)$. First we prove that $\Pr\{W_{n,m} \geq a \cdot b\} \to 0$ for sufficiently large m. We assume that the largest area consists of at least a rows and at least b columns. Observe that, in general, $\{W \geq a \cdot b\}$ does not imply that the number of rows is at least a and the number of columns is at least b. In the appendix, we prove that the same result holds without this assumption. By Boole's inequality one can find

$$\Pr\{W_{n,m} \geq a \cdot b\} \leq \binom{n}{a} \sum_{k=b}^{m} \binom{m}{k} P_a^k (1-P_a)^{m-k}. \tag{7}$$

The upper bound in (7) can be derived from the following observation. The number of matching columns among the fixed a rows has a binomial distribution with parameters m and P_a, i.e. the probability that the number of such columns is equal to k is $\binom{m}{k} P_a^k (1-P_a)^{m-k}$.

Suppose that $\beta > P_a$. Let k^* be the k that maximizes $\Delta(k) := \binom{m}{k} P_a^k (1-P_a)^{m-k}$. It is easy to see that $k^* = \lfloor P_a m \rfloor$. We note that $\Delta(k)$ is increasing for $k < k^*$ and decreasing when $k > k^*$. Since $b > k^*$, each $\Delta(k)$ in the summation (7) can be replaced with $\Delta(b)$, which still preserves the inequality. That is,

$$\Pr\{W_{n,m} \geq a \cdot b\} \leq \binom{n}{a} m \binom{m}{b} P_a^b (1-P_a)^{m-b}.$$

Since

$$\binom{m}{b} \leq \frac{C}{\sqrt{2\pi m(1-\beta)\beta}} \exp(mH(\beta)),$$

where $H(\beta) = \beta \log \beta^{-1} + (1-\beta)\log(1-\beta)^{-1}$ is the entropy and C is a constant (omitting the constant C) we arrive at

$$\begin{aligned}\Pr\{W_{n,m} \geq a \cdot b\} &\leq \frac{\binom{n}{a} m}{\sqrt{2\pi m(1-\beta)\beta}} \exp(mH(\beta) - b \log P_a^{-1} - (m-b)\log(1-P_a)^{-1}) \\ &= \frac{\binom{n}{a} m}{\sqrt{2\pi m(1-\beta)\beta}} \exp(mf(\beta)),\end{aligned}$$

where $f(\beta) = H(\beta) - \beta \log P_a^{-1} - (1-\beta)\log(1-P_a)^{-1}$. Note that $f'(\beta) = \log\left(\frac{(1-\beta)P_a}{\beta(1-P_a)}\right) < 0$, for $\beta > P_a$. Therefore, we can write $f(\beta) = -\gamma$ for some constant $\gamma > 0$. Finally, we have

$$\Pr\{W_{n,m} \geq a \cdot b\} \leq \frac{\binom{n}{a}(m-b)}{\sqrt{2\pi n(1-\beta)\beta}} \exp(-\gamma m) \to 0,$$

since we are assuming that $n = O(m^k)$ for k constant. We have proved that $W_{n,m} \leq a \cdot \beta m$ whp for $a = O(1)$ and for any $\beta > P_a$.

For the lower bound, we observe that for fixed a, $W_{n,m} \geq \mathbf{E}[X_{IJ}] = a \cdot P_a m$, since $P_a m$ is the expected number of matching columns. Therefore $W_{n,m} = aP_a m$ whp. We further can optimize it by finding a^* that maximizes $a^* P_{a^*}$. This completes the proof.

Proof of (3): We let $a = \alpha n$ for some $0 < \alpha < 1$ and fix b columns for some $b > 0$. Let also R be a row vector consisting of (any) b symbols. We are concerned with the number of row vector R occurred in the submatrix consisting of those b columns.

Let $P(b)$ be the probability of emitting all elements of R (e.g., for $b = 4$ and $R = (1,2,1,3)$, $P(b) = p_1^2 p_2 p_3$). The rest of proof is similar to the previous one and one can obtain that $W_{n,m} = bP(b)n$ whp. The final step is to find b that maximizes $bP(b)$. We note that $P(b) \leq p_{max}^b$, and if $P(b) = p_{max}^b$ the largest submatrix found consists of the symbols that have the maximum emitting probability. We further optimize bp_{max}^b and find that bp_{max}^b reaches the maximum when $b = 1/\log Q$.

References

R. Darling and M. Waterman. Matching rectangles in d-dimensions: Algorithms and laws of large numbers. *Adv. Math.*, 55, 1985.

M. Dawande, P. Keskinocak, J. M. Saminathan, and S. Tayur. On bipartite and multipartite clique problems. *Journal of Algorithms*, 41:388–403, 2001.

M. S. Garey and D. Johnson. *Computers and Intractibility: A Guide to NP-Completeness*. Freeman, New York, 1979.

S. Lonardi, W. Szpankowski, and Q. Yang. Finding biclusters in gene expression data by random projections. In *15th Annual Combinatorial Pattern Matching Symp.*, 2004.

R. Peters. The maximum edge biclique problem is np-complete. Technical report, Tilburg University Department of Econometrics research memorandum, 2000.

Y. M. Q. Sheng and B. Moor. Biclustering microarray data by Gibbs sampling. *Bioinformatics*, 19, 2003.

W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley, New York, 2001.

A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. pages 136–144, 2002.

H. Wang, W. Wang, J. Yang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data (SIGMOD-02)*, pages 394–405. ACM Press, New York, 2002.

C.-J. Wu, Y. Fu, T. M. Murali, and S. Kasif. Gene espression module discovery using Gibbs sampling. *Genome Informatics*, 16(1):239–248, 2004.

Appendix: Complete Proof of Theorem 2.2

Here we drop the assumption that the largest area consists of at least a rows and at least b columns where $a = O(1)$ and $b = \beta m$, with $0 < \beta < 1$. To begin with, we define random variables R (and C) as the number of rows (and columns) in the largest matching submatrix. Observe that for any fixed $w > 0$, we have

$$Pr\{W_{n,m} \geq w\} = \sum_{r=1}^{w} \Pr\{R = r,\, C \geq \lceil w/r \rceil\}$$

In the rest, we omit the symbols $\lceil\ \rceil$ or $\lfloor\ \rfloor$ for simplicity. Without loss of generality, we may assume that $C \geq R$. Otherwise, we end up with the result (3) in Theorem 2.2

Therefore, we can write $Pr\{W_{n,m} \geq w\}$ as

$$Pr\{W_{n,m} \geq w\} = \sum_{r=1}^{w} \sum_{c=r^*}^{m} \binom{n}{r} \binom{m}{c} P_r^c (1-P_r)^{m-c},$$

where $r^* = \max\{\lceil w/r \rceil, r\}$. We re-write it as

$$Pr\{W_{n,m} \geq w\} = \sum_{r=1}^{\sqrt{w}} \binom{n}{r} \sum_{c=w/r}^{m} \Delta(r,c) + \sum_{r=\sqrt{w}+1}^{w} \binom{n}{r} \sum_{c=r}^{m} \Delta(r,c),$$

where $\Delta(r,c) = \binom{m}{c} P_r^c (1-P_r)^{m-c}$.

The remaining part of the proof follows in the same footsteps as in Section 3.2. Let a^* be the integer r that maximizes $rP_r = r(p_1^r + \cdots + p_V^r)$ and let $w = (1+\varepsilon)a^* P_{a^*} m$, for some $\varepsilon > 0$. Since $w/r > mP_r$, for any fixed r, $\Delta(r,c) \leq \Delta(r, w/r)$ holds for $c \geq w/r$. When $r \geq \sqrt{w}$, then $\Delta(r,c) \leq \Delta(r,r)$ holds for $c \geq r$. Therefore, we obtain, for sufficiently large m and n

$$\begin{aligned}
\Pr\{W_{n,m} \geq w\} &\leq m \sum_{r=1}^{\sqrt{w}} \binom{n}{r} \Delta(r, w/r) + m \sum_{r=\sqrt{w}+1}^{w} \binom{n}{r} \Delta(r,r) \\
&\leq m\sqrt{w} n^{\sqrt{w}} e^{-\gamma_1 m} + m \sum_{r=\sqrt{w}+1}^{w} n^r m^r V^r p_{max}^{r^2} \\
&\leq m\sqrt{w} n^{\sqrt{w}} e^{-\gamma_1 m} + mwn^{\sqrt{w}} m^{\sqrt{w}} V^{\sqrt{w}} p_{max}^{w} \\
&\leq e^{-\gamma_1 m} + e^{-\gamma_2 m},
\end{aligned}$$

where γ_1 and γ_2 are positive constants that may change from line to line. We have shown that $W_{n,m} \leq (1+\varepsilon)a^* P_{a^*} m$ whp, as desired.