© 2013 FIZ Karlsruhe

io-port 05050587

Park, Gahyun; Szpankowski, Wojciech

Analysis of biclusters with applications to gene expression data.

Martínez, Conrado (ed.), 2005 international conference on analysis of algorithms. Papers from the conference, Barcelona, Spain, June 6–10, 2005. Nancy: The Association Discrete Mathematics & Theoretical Science (DMTCS). Discrete Mathematics and Theoretical Computer Science Proceedings AD, 267-277, electronic (2005).

Summary: For a given matrix of size $n \times m$ over a finite alphabet 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 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