Bulletin of Mathematical Biology Vol. 48, No. 2, pp. 189–195, 1986.
Printed in Great Britain.

0092-8240/86$3.00 + 0.00
Pergamon Press Ltd.

INTERVAL GRAPHS AND MAPS OF DNA

■ MICHAEL S. WATERMAN*
Departments of Mathematics and of Biological Sciences,
University of Southern California,
Los Angeles, CA 90089-1113, U.S.A.

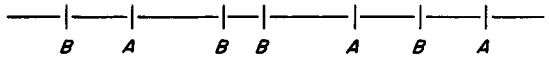
■ JERROLD R. GRIGGS†
Department of Mathematics and Statistics,
University of South Carolina, SC 29208, U.S.A.

A special class of interval graphs is defined and characterized, and an algorithm is given for their construction. These graphs are motivated by an important representation of DNA called restriction maps by molecular biologists. Circular restriction maps are easily included.

The study of interval graphs has its origin in a paper of Benzer (1959) who was studying the structure of bacterial genes. At that time it was not known whether or not the collection of DNA composing a bacterial gene was linear. It is now well-known that such genes are linear words over a four-letter alphabet, and Benzer's work was basic in establishing this fact. Essentially, he obtained data on the overlap of fragments of the gene and showed the data consistent with linearity.

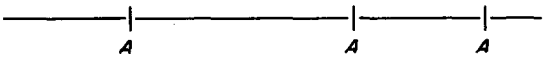
Of course there is no longer active interest in Benzer's problem, but we take this opportunity to draw attention to a special class of interval graphs central to the modern practice of molecular biology. These graphs arise in connection with restriction maps which show the location of certain sites (short specific sequences) on a specific DNA. Danna *et al.* (1973) first sketched the principles of restriction mapping, which utilizes a property of restriction enzymes. These enzymes, found in various bacteria, cleave or cut the DNA at all occurrences of short specific sequences.

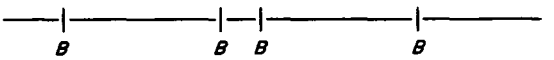
To make these ideas specific, we graphically present a restriction map, which we will refer to as an $A \cdot B$ map, with three occurrences of restriction site A and four occurrences of restriction site B.



* Supported by a grant from the System Development Foundation.
† Supported by an NSF grant and through a grant from the System Development Foundation.

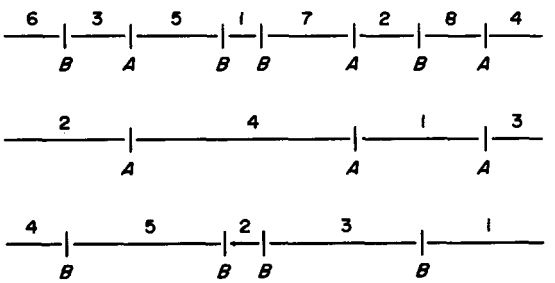
Next we show the maps for A and B separately and refer to them as the A map and B map, respectively





Biologists, when constructing restriction maps, can identify individual intervals between sites, but they cannot directly observe the order of these intervals. Instead they establish whether or not an A interval overlaps a B interval and, from this overlap data, construct the map. Frequently this overlap data comes from determining which A and B intervals contain the various $A \cdot B$ intervals. In fact, frequently the most difficult aspect of restriction map construction is determining the overlap data. This difficult problem is not pursued further here.

Next the intervals are arbitrarily labelled:



Label the components of the A map by $A_1, A_2, \ldots$ and of the B map by $B_1, B_2, \ldots$. Define the incidence matrix $I(A, B)$ whose (i, j)th entry is 1 if $A_i \cap B_j \neq \phi$ and is 0 otherwise. For the above example,

$$I(A, B) = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

$$I(A, A \cdot B) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

$$I(B, A\cdot B) = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

As mentioned above $I(A, A\cdot B)$ and $I(B, A\cdot B)$ are frequently known while $I(A, B)$ is desired. The next proposition relates these matrices.

PROPOSITION. *For the incidence matrices defined above*

$$I(A, B) = I(A, A\cdot B)I'(B, A\cdot B), \tag{1}$$

where $I'(B, A\cdot B)$ is the transpose of $I(B, A\cdot B)$.

Proof. Notice that the (i, j) element of this matrix product equals the number of $A\cdot B$ intervals in both the ith A interval and the jth B interval. But the $A\cdot B$ intervals are formed by intersection of A intervals with B intervals so that the entries in the matrix product must be 0 or 1.

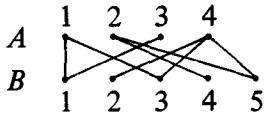
Having shown that $I(A, B)$ is easily obtained from $I(A, A\cdot B)$ and $I(B, A\cdot B)$, we now turn to characterizing $I(A, B)$ and then present an algorithm for constructing restriction maps from $I(A, B)$. Two equivalent characterizations are discussed and then collected in Theorem 1.

The matrix $I(A, B)$ tells us when an A interval and a B interval have an $A\cdot B$ interval in common, or, equivalently, when the interiors of the A interval and B interval intersect. Thus, constructing a restriction map from $I(A, B)$ is equivalent to finding an interval representation for a certain graph $G(A, B)$ which is obtained in the following natural way: the vertex set $V(A, B)$ of $G(A, B)$ consists of the union of the set of A intervals and the set of B intervals, and the edge set $E(A, B)$ consists of the unordered pairs $\{A_i, B_j\}$ for each A intervals, A_i, and B-interval, B_j, which overlap.

If $G(A, B)$ arises from a restriction map we need only delete the endpoints of the pieces in the A map and the B map to obtain an (open) interval representation of the graph $G(A, B)$. Thus $G(A, B)$ is an interval graph. Since the interiors of the A intervals are disjoint, the interiors of the B intervals are disjoint, and the interior of every A interval (respectively, B interval) overlaps the interior of some B interval (respectively, A interval), it follows that $G(A, B)$ is bipartite with no isolated vertices.

Conversely, one may construct a restriction map for any bipartite interval graph G without isolated vertices by drawing together the intervals representing vertices in each of the two parts until the A and B intervals correspond to A and B maps.

Here is the graph for our example above:



We next observe that $G(A, B)$ is a connected graph unless the A and B restriction sites coincide. In general, if the A and B restriction sites coincide k times (not counting the ends), then $G(A, B)$ will have $k + 1$ connected components.

To construct a restriction map from $I(A, B)$, that is, an interval representation of $G(A, B)$, we need only linearly order the edges of $G(A, B)$ in such a way that for every vertex v, all edges containing v occur consecutively. This is a special case of the ordering of the maximal cliques of arbitrary interval graphs (Gilmore and Hoffman, 1964). We shall obtain below such an ordering of the edges for our example. Another property of the graphs $G(A, B)$ which can arise is that $G(A, B)$ contains no isolated vertices, no cycles C_n, $n \geqslant 3$ (induced or not), and that no vertex has three or more vertices each of degree $\geqslant 2$ adjacent to it. That is, $G(A, B)$ is a forest with no isolated vertices. This is a special case of a characterization of arbitrary interval graphs (Lekkerkerker and Boland, 1962).

There is also a 0–1 matrix formulation characterizing interval graphs (Fulkerson and Gross, 1964). In our special case it can be shown that for the graphs $G(A, B)$ that can arise, the associated 0–1 matrix $I(A, B)$ can be put into a particularly nice form. Specifically, if the rows and columns of $I(A, B)$ are permuted in accordance with an ordering of the edges of $G(A, B)$ which corresponds to a restriction map (interval representation), then the 1s in $I(A, B)$ will belong to one of a collection of $k + 1$ staircase shapes going from top left to lower right, where k is the number of components of $G(A, B)$. Each row or column has its 1s consecutive and meets precisely one of the staircases. There can be no two by two all ones submatrix. Here is such a permutation for our example.

$$
\begin{array}{cc} & \begin{array}{ccccc} & & B & & \\ 4 & 5 & 2 & 3 & 1 \end{array} \\ A \begin{array}{c} 2 \\ 4 \\ 1 \\ 3 \end{array} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}
$$

These characterizations of restriction maps are now collected in the following theorem.

THEOREM 1. *The following statements are equivalent:*

(i) *The bipartite graph $G(A, B)$ is the graph constructed from some restriction mapping.*

(ii) *$G(A, B)$ is a bipartite interval graph with no isolated vertices.*

(iii) *$I(A, B)$ can be transformed by row and column permutations into staircase form with each row or column having 1s in precisely one of these staircases.*

When $G(A, B)$ arises from a restriction map, that is, when $G(A, B)$ is a bipartite graph without isolated points,

$$\tfrac{1}{2}|V(A, B)| \leqslant |E(A, B)| \leqslant |V(A, B)| - 1.$$

The lower bound on $|E(A, B)|$ follows since there are no isolated points, so that every vertex is on some edge. The upper bound follows by considering an interval representation of such a $G(A, B)$ and noting that in going from left to right, every new vertex encountered, except the first one, accounts for at most one new edge. Here we are ordering the intervals according to their left endpoints and breaking ties arbitrarily. Thus $|E(A, B)|$ and $|V(A, B)|$ are of the same order. For more information on interval graphs in general, see Golumbic (1980).

Interval graphs in general can be recognized and their representations can be found in time which is linear in the size of the graph, by an algorithm due to Booth and Leuker (1976). For the class of graphs considered here we can provide a recognition and representation algorithm which is particularly simple. Like the algorithm for the general problem, it requires only linear time and storage.

To save storage, we work with the edges rather than $I(A, B)$. Let E be the set of all edges e and V the set of all vertices v.

Algorithm

1. For all $v \in V$ find $\deg(v)$.
2. Set $L_0 = L = \{v: \deg(v) \geqslant 2\}$,
   $L(v) = \{u: \{u, v\} \in E \text{ and } u \in L\}$.
3. Find $v \in L_0$ with exactly one neighbor u satisfying $u \in L_0$.
   If no v exists, go to 5.
   Set $v_1 = v$ and $v_2 = u$.
4. Find the maximal list $\{v_1, v_2\}\{v_2, v_3\} \ldots \{v_{r-1}, v_r\}$
   where $v_t \in L_0$ and $v_t \neq v_{t+1}$.
   Delete $v_1, v_2, \ldots, v_r$ from L_0.
   Go to 3.

5. For $v \in L \sim L_0$ and $u \in L^c$ insert $\{v, u\}$ between $\{v_t, v\}$ and $\{v, v_{t+2}\}$.
6. Add, to the lists already obtained, $\{v, u_1\}\{v, u_2\} \ldots \{v, u_r\}$ where $v \in L_0, u_t \in L^c$.
7. Add $\{u, v\}$ where $u, v \in L^c$.
8. Output all lists. Stop.

It is easy to see that the algorithm runs in linear time. The key is step 4 which depends on the fact noted earlier that each vertex has at most two neighbors in L. The algorithm can be modified to provide recognition of these graphs.

Consider the graph from our example. Suppose the vertices are listed $A_1, A_2, A_3, A_4, B_1, B_2, B_3, B_4, B_5$, and edges are listed $\{A_1, B_1\}, \{A_1, B_3\}, \{A_2, B_4\}, \{A_2, B_5\}, \{A_3, B_1\}, \{A_4, B_2\}, \{A_4, B_3\}, \{A_4, B_5\}$. The algorithm steps 1 and 2 give

vertex v	A_1	A_2	A_3	A_4	B_1	B_2	B_3	B_4	B_5
degree deg(v)	2	2	1	3	2	1	2	1	2
$L(v)$	B_1, B_3	B_5	B_1	B_3, B_5	A_1	A_4	A_1, A_4	A_2	A_2, A_4.

In steps 3 and 4, we encounter the end A_2 which leads to long vertices B_5, A_4, B_3, A_1, B_1. No other ends are to be found. So far this gives this edge ordering:

$$\{A_2, B_5\}, \{B_5, A_4\}, \{A_4, B_3\}, \{B_3, A_1\}, \{A_1, B_1\}.$$

Now we go through the edge list and insert each edge which contains a short vertex and one of the long vertices previously involved in the ordering. The new edge must be adjacent to a previous edge involving the long vertex. We insert it between previous ones, if there is a choice, or else on the left or right if the long vertex was a left or right end, respectively. It happens that in our example we now obtain a complete ordering:

$$\{A_2, B_4\}, \{A_2, B_5\}, \{A_4, B_5\}, \{A_4, B_2\}, \{A_4, B_3\}, \{A_1, B_3\}, \{A_1, B_1\}, \{A_3, B_1\},$$

The Booth-Leuker algorithm for general interval graphs as described in Golumbic (1980) uses an appropriate data structure. It also describes there how to obtain all possible edge orderings, although we can describe it simply for our restricted problem: the order of the components can be permuted, the order of the edges involving a fixed long vertex and any short vertices can be permuted, and components containing an end can have the order of their edges reversed (but not arbitrarily permuted).

In conclusion we note that Jungck *et al.* (1982) discuss the connections between interval graphs and protein sequencing. They give a historical account and credit Fox with the problem of determining linear order by

means of overlap data. In addition, Fitch *et al.* (1983) discuss mapping DNA and give a brief description of determining linear order. This paper is concerned with graph theoretical structures and not with practical methods for mapping DNA. Finally, we note that DNA is in many cases circular and that minor modifications of our results suffice for the circular case. To begin the algorithm simply chose any v such that $\deg(v) \geqslant 2$.

THEOREM 2. (i) *A bipartite graph $G(A, B)$ is consistent with a circular restriction map if and only if the graph obtained by removal of any vertex of degree 2 is a bipartite interval graph with no isolated vertices.* (ii) *A graph $G(A, B)$ satisfying* (i) *must be mapped as circular if and only if all $v_0 \in L = \{v: \deg(v) \geqslant 2\}$ have card $\{u: \{u, v_0\} \in E$ and $u \in L\} \geqslant 2$.*

We close by mentioning two problems. The first is to devise algorithms and to characterize these graphs for higher dimensional sets, rectangles say, in Euclidean space. The second is to devise algorithms for the minimum number of edge removals to convert $G(A, B)$ into a restriction map when it is not intially one.

The first author (M.S.W.) wishes to thank G. C. Rota and W. M. Fitch for valuable discussions. The algorithms were programmed in C by Mark Eggert.

LITERATURE

Benzer, S. 1959. "On the topology of genetic fine structure." *Proc. natn. Acad. Sci. U.S.A.* **45**, 1607–1620.

Booth, K. S. and G. S. Leuker. 1976. "Testing for the consecutive ones property, interval graphs, and graph planarity using PO algorithms." *J. Comput. Syst. Sci.* **13**, 335–379.

Danna, K. J., G. H. Sack and D. Nathans. 1973. "Studies of Simian Virus 40 DNA. VII. A cleavage map of the SV40 genome." *J. molec. Biol.* **78**, 363–376.

Fitch, W. M., T. F. Smith, and W. W. Ralph. 1983. "Mapping the order of DNA restriction fragments." *Gene* **22**, 19–29.

Fulkerson, D. R. and O. A. Gross. 1964. "Incidence matrices with the consecutive 1s property." *Bull. Am. math. Soc.* **70**, 681–684.

Gilmore, P. C. and A. J. Hoffman. 1964. "A characterization of comparability graphs and of interval graphs." *Can. J. Math.* **16**, 539–548.

Golumbic, M. C. 1980. *Algorithmic Graph Theory and Perfect Graphs*. New York: Academic Press.

Jungck, J. R., G. Dick and A. G. Dick. 1982. "Computer-assisted sequencing, interval graphs, and molecular evolution." *Bio-Syst.* **15**, 259–273.

Lekkerkerker, C. G. and J. C. Boland. 1962. "Representation of a finite graph by a set of intervals on the real line." *Fund. Math.* **51**, 45–64.

RECEIVED 1-28-85
REVISED 10-7-85