BIOINFORMATICS

Vol. 17 no. 2 2001
Pages 174–179

The art of pedigree drawing: algorithmic aspects

Frédéric Tores[1] *and Emmanuel Barillot*[1, 2,*]

[1]*GIS Infobiogen, 7 rue Guy Môquet, BP 8, 94801 Villejuif cedex and* [2]*Généthon, 1 bis rue de l'Internationale, 91000 Évry, France*

Received on July 13, 2000; revised on September 7, 2000; accepted on September 9, 2000

ABSTRACT

Motivation: Giving a meaningful representation of a pedigree is not obvious when it includes consanguinity loops, individuals with multiple mates or several related families.

Results: We show that finding a perfectly meaningful representation of a pedigree is equivalent to the interval graph sandwich problem and we propose an algorithm for drawing pedigrees.

Contact: Emmanuel.Barillot@infobiogen.fr

INTRODUCTION

The improvements of molecular biology have made large and complex pedigrees a usual resource for epidemiologists. Pedigrees under study can now reach several hundreds of individuals for humans, and several thousands for plants or animals. They generally include consanguinity loops, individuals with multiple mates or several related families. This poses the problem of representing graphically this information in a meaningful form, so that its internal organization be obvious to understand. Traditionally, there is a vertical sorting of the individuals by generation. Secondly, the individuals are ordered horizontally. If the vertical sorting seems trivial, the positioning of individuals in the generation may be difficult. Misplacing individuals will create line-crossing and the pedigree representation may become cryptic as the number of line-crossing grows. To our knowledge, none of the existing pedigree drawing softwares addresses these points satisfactorily in all possible cases. They are often commercial products that integrate several functionalities (for instance: Cyrillic, Cherwell Scientific, 2000; Progeny, Progeny Corp, 2000; GAP, Epicenter Software, 2000); their drawing algorithms are most of the time not public, but it can be inferred from their results that they are very different from our approach.

The main points addressed in this paper are the definition of a perfect representation of a pedigree, the demonstration of equivalence between the pedigree drawing problem and the interval graph sandwich problem and the proposition of an algorithm to find a perfect representation of a pedigree.

In the first section, we formulate precisely the problem by defining a perfect representation of a pedigree. Then interval graphs are introduced in the second section and the equivalence between perfect representations and the interval graph sandwich problem is shown. Finally, an algorithm for deriving a perfect representation from a pedigree is presented in the last section.

POSITION OF THE PROBLEM

Definitions about pedigrees

A pedigree $\mathcal{P}$ is a set of individuals that are related by four types of relations: mate, parent, child and sib. We suppose in the following that the pedigree is fully connected: each pair of individuals in $\mathcal{P}$ can be joined by a chain of related individuals.

We define a family as the descendants of a mating with no ascendant in $\mathcal{P}$. Thus a pedigree may consist of one or several families and if it has several families, they intersect somehow.

We define an orphan mate as a mate with no other family than his/her mate relatives, that is with no parent, no sib and no other mate.

Each individual is given a generation rank, with the common meaning of generation. But an ambiguity may arise when there is a loop in a pedigree, such as a consanguinity loop (an uncle mated with his niece for example). To cope with this problem and clear up the ambiguity, one can consider that the generation rank attribution starts with an arbitrary individual in $\mathcal{P}$ and is then propagated according to a sib–first, parent–second, child–third and mate–last strategy. Of course, the generation rank is incremented for children, decremented for parents and left unchanged for sibs and mates and one never passes twice through the same individual. This ensures that all sibs will have the same generation rank (which gives the definition of the generation rank of a sibship) and at least one parent has a generation rank immediately lower than the child rank. The generation rank(s) of a mating is defined as the generation rank of the mates in case of equality, and as the range they define if they differ.

*To whom correspondence should be addressed.

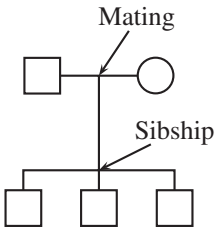

Fig. 1. Example of a simple pedigree.

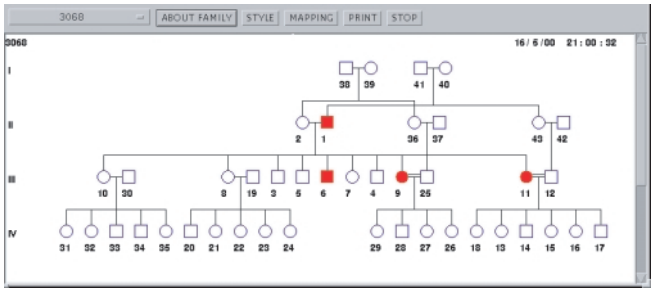

Fig. 2. Example of a non-trivial PDP pedigree. It includes 2 consanguinity loops and 2 line crossings on this drawing.

Rules of representation of a pedigree

The common practice and some standardization efforts (Bennet *et al.*, 1995) have defined some rules of representation of a pedigree that should make clear its internal structure. The base unit is the individual who is depicted as a circle for a woman, and as a square for a man. The first degree familial relationships, and only them, are represented through links between individuals:

(1) mates are linked and create a mating unit;

(2) sibs are linked to a sibship unit;

(3) the sibship is linked to its parental mating.

This is summarized in the Figure 1 with a nuclear family. Some rules for positioning individuals and drawing links have also been defined:

(4) people from the same generation are drawn on the same horizontal line, and the older generations are at the top;

(5) links between mates are as much as possible horizontal;

(6) sibships are materialized by tooth-down horizontal combs that connect all their members;

(7) sibship-to-parents links are as much as possible vertical.

It is easy to produce a pedigree drawing that respects these seven rules but its readability may be seriously spoiled by many link crossings, or by the fact that individuals in a mating or in a sib, although connected, may not appear as adjacent. Such a case is illustrated Figure 2.

Rules of readability of a pedigree

We therefore define some rules of readability to whose a pedigree drawing should conform to be perfectly meaningful:

(a) No overlap is allowed between individuals. This is easily realized if individuals are from different generations because there is no other constraints on the vertical dimension. This is less obvious if individuals are from the same generation, because they are also bound by the other constraints that follow.

(b) Mates must be adjacent.

(c) Sibs must be adjacent, but orphan spouses may be inserted within a sibship. We define as 'extended sibship' a sibship with all the orphan spouses of the sibship members.

(d) Parents are above their child sibship (the span of parent mating representation and the one of child sibship must have a common vertical).

(e) There are no link crossing.

A pedigree for which a drawing verifying these five rules exists is said to be a perfectly drawable pedigree (PDP). Our problem is to find:

- a method for establishing whether a pedigree is PDP; and
- an algorithm to compute the perfect drawing of a PDP.

Problematic cases

Simple pedigrees as the one in Figure 1 are obviously PDP and their drawing is easily computable without any sophisticated algorithm. But it may not be trivial to position individuals on the drawing as shown on Figure 2 when the pedigree includes:

- consanguinity loops;
- inter-generation matings;
- individual(s) with multiple mates;
- several individuals in a family mated to non-orphan individuals;
- several related families.

In fact, it can be easily shown that a pedigree cannot be PDP when it includes:

- individual(s) with more than two mates; or
- cycle(s) of related families (for instance, an individual in family A is mated to an individual in family B, another in B to a person in family C, and another in C to a person in A); or
- more than two individuals from the same generation and family with non-orphan mates.

In the next section, we propose a method to test whether a pedigree is PDP or not.

INTERVAL GRAPH AND PEDIGREE DRAWING

What is an interval graph?

A graph G is defined as the couple $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \ldots, v_n\}$ is a set of elements called vertices or nodes and $\mathcal{E} = \{e_1, \ldots, e_n\}$ is a set of edges that connect pairs of elements from $\mathcal{V}$ ($e_i = \{p_{i_1}, p_{i_2}\}$). Two elements v and v' from $\mathcal{V}$ that are connected by an edge are said adjacent, which is noted $v \leftrightarrow v'$, while $v \nleftrightarrow v'$ denotes no connection.

A graph is an interval graph if, and only if, there exists for every vertex an interval on the real line such that two vertices are adjacent if and only if their corresponding intervals intersect. It can be shown that a graph is an interval graph if, and only if, it contains no chordless cycles bigger than three (it is said to be chordal) and no asteroidal triplets (an asteroidal triplet is a triplet of non-adjacent vertices for which there exists between each two of them a path that has no vertex adjacent to the third). Examples of a 4-cycle and an asteroidal triplet are given in Figure 3. Also Rose *et al.* (1976) showed that testing if a graph is an interval graph can be done in a linear time.

Why to use interval graphs in pedigree drawing optimization?

Intuitively it seems that the rules of readability exposed above could be expressed in terms of intervals and interval overlaps. For instance if an interval is associated to each individual, rule (a) means that no two intervals from the same generation can be connected; rule (b) means that the union of the two intervals of the individuals in a mating must be an interval; also the union of the intervals of the individuals in a sibship must be an interval to abide by rule (c); and rule (d) amounts to an overlap between the interval of the sibship and the ones of the parental mating.

These remarks suggest to define the following vertices :

- a vertex per individual;
- a vertex per mating;

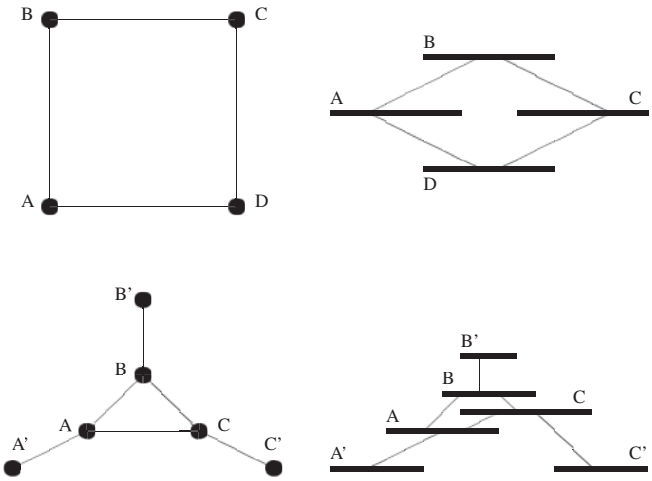

Fig. 3. Example of four-cycle (top) and asteroidal triplet (bottom) and illustration of why they are forbidden in interval graphs: for example on the right side, an edge between B and D is missing in the cycle, and edges from B' to A, and from B' to C are missing in the asteroid.

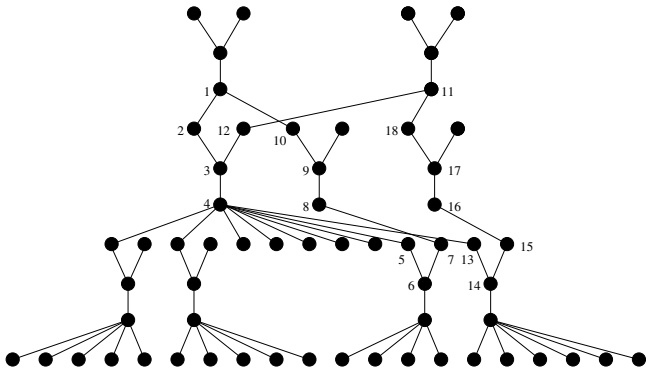

Fig. 4. This graph is derived from pedigree of Figure 2 using the vertices defined above (a vertex per individual, a vertex per mating and a vertex per extended sibship).

- a vertex per extended sibship.

and the following sets of vertices:

- $\mathcal{V}_i$, the set of vertices corresponding to individuals;
- $\mathcal{V}_m$, the set of vertices corresponding to matings;
- $\mathcal{V}_s$, the set of vertices corresponding to sibships; and
- $\mathcal{V} = \mathcal{V}_i \cup \mathcal{V}_m \cup \mathcal{V}_s$.

and advocate the use of interval graphs defined on this set of vertices $\mathcal{V}$. Figure 4 shows the natural graph derived from the pedigree of Figure 2 using the set of vertices

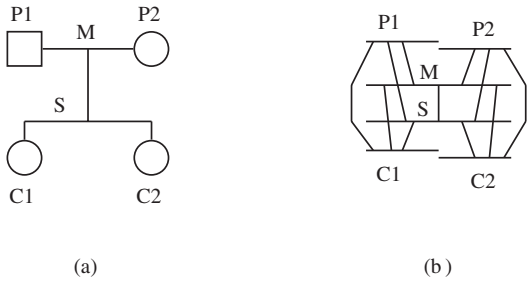

Fig. 5. Example of pedigree (a) and its associated interval graph (b). *M* stands for mating and *S* for sibship. Note that the span of an individual extends up to his/her neighbour.

mentioned above. As one can see, it also includes cycles and its drawing presents line-crossings.

Equivalence between PDP and the interval graph sandwich problem

We now remark that the perfect drawing of a pedigree can be seen as an interval graph on the set of vertices $\mathcal{V}$ defined above. If we associate to each individual, mating and sibship the interval it spans in the horizontal dimension as in Figure 5, one can build an interval graph by connecting the intervals that intersect. The drawing corresponds simply to an interval realization consistent with the interval graph.

So the rest of the section is devoted to characterizing the interval graphs on $\mathcal{V}$ whose realizations correspond in the sense we have defined to a perfect drawing of the pedigree.

Denote:

- $\mathcal{I}(\mathcal{V}, \mathcal{E})$ such an interval graph;
- $p(v)$ the set of individual(s) that compose(s) the mating, sibship or individual associated to the vertex $v, \forall v \in \mathcal{V}$,
- $g(v)$ the set of generation rank(s) of $p(v), \forall v \in \mathcal{V}$,
- $c(v)$ the set of individual(s) that is (are) child(ren) of $p(v), \forall v \in \mathcal{V}_i \cup \mathcal{V}_m$.

By definition, $\mathcal{I}(\mathcal{V}, \mathcal{E})$ has a realization that corresponds to a perfect drawing if, and only if, the rules of readability exposed in the first section are respected:

Conditions arising from rule (a). To respect rule (a), individuals from the same generation must not intersect:

$$\forall v_1, v_2 \in \mathcal{V}_i | g(v_1) = g(v_2), \qquad v_1 \nleftrightarrow v_2.$$

If we denote $\mathcal{E}_a^- = \{\{v_1, v_2\} | v_1, v_2 \in \mathcal{V}_i \text{ and } g(v_1) = g(v_2)\}$, we can formulate a first condition:

$$\mathcal{E}_a^- \cap \mathcal{E} = \emptyset.$$

Conditions arising from rule (b). To respect rule (b), $\mathcal{I}$ should have edges connecting individuals to their mating(s), and no other individuals from the same generation should be in between. It can be shown easily that one can require that the vertices corresponding to other individuals from the same generation should not be adjacent to the vertex corresponding to the mating; although this is a more restrictive requirement, it will not affect the *existence* of realizations.

$$\forall v_i \in \mathcal{V}_i, v_m \in \mathcal{V}_m, \quad (p(v_i) \subset p(v_m)) \Rightarrow v_i \leftrightarrow v_m$$
$$(g(v_i) = g(v_m) \text{ and } p(v_i) \not\subset p(v_m)) \Rightarrow v_i \nleftrightarrow v_m.$$

If we denote $\mathcal{E}_b^+ = \{\{v_i, v_m\} | v_i \in \mathcal{V}_i, v_m \in \mathcal{V}_m \text{ and } p(v_i) \subset p(v_m)\}$ and $\mathcal{E}_b^- = \{\{v_i, v_m\} | v_i \in \mathcal{V}_i, v_m \in \mathcal{V}_m \text{ and } g(v_i) = g(v_m) \text{ and } p(v_i) \not\subset p(v_m)\}$, we can formulate a second condition:

$$\mathcal{E}_b^- \cap \mathcal{E} = \emptyset \qquad \text{and} \qquad \mathcal{E}_b^+ \subset \mathcal{E}.$$

Conditions arising from rule (c). Rule (c) can be treated similarly to rule (a) with sibships playing the role of matings:

$$\forall v_i \in \mathcal{V}_i, v_s \in \mathcal{V}_s, \quad (p(v_i) \subset p(v_s)) \Rightarrow v_i \leftrightarrow v_s$$
$$(g(v_i) = g(v_m) \text{ and } p(v_i) \not\subset p(v_s)) \Rightarrow v_i \nleftrightarrow v_s.$$

If we denote $\mathcal{E}_c^+ = \{\{v_i, v_s\} | v_i \in \mathcal{V}_i, v_s \in \mathcal{V}_s \text{ and } p(v_i) \subset p(v_s)\}$ and $\mathcal{E}_c^- = \{\{v_i, v_s\} | v_i \in \mathcal{V}_i, v_s \in \mathcal{V}_s \text{ and } g(v_i) = g(v_m) \text{ and } p(v_i) \not\subset p(v_s)\}$, we can formulate a third condition:

$$\mathcal{E}_c^- \cap \mathcal{E} = \emptyset \qquad \text{and} \qquad \mathcal{E}_c^+ \subset \mathcal{E}.$$

Conditions arising from rule (d). Rule (d) implies that the vertex corresponding to a sibship should be adjacent to the parental mating vertex:

$$\forall v_m \in \mathcal{V}_m, v_s \in \mathcal{V}_s | c(v_m) = p(v_s), v_m \leftrightarrow v_s.$$

If we denote $\mathcal{E}_d^+ = \{\{v_m, v_s\} | v_m \in \mathcal{V}_m, v_s \in \mathcal{V}_s \text{ and } p(v_m) = p(v_s)\}$, we can formulate a fourth condition:

$$\mathcal{E}_d^+ \subset \mathcal{E}.$$

Conditions arising from rule (e). Rule (e) requires no link crossing which is more problematic. In a pedigree drawing, there are three types of lines:

- horizontal lines for matings and sibships;
- vertical lines for linking a mating to its child sibship;
- oblique lines for inter-generation matings.

To ensure that no two horizontal lines will intersect, one has to require that matings and sibships from the

same generation should not intersect, except if they have an individual in common. In such a case, they have to intersect only for this individual, but this is already taken into account by our treatment of rules (b) and (c).

Vertical lines cannot cross horizontal lines nor other vertical lines because they link a mating to its child sibship one generation underneath.

Oblique lines can cross any type of lines, so edges that connect the vertex corresponding to the inter-generation mating to any vertex from a generation strictly comprised in between the generations of the mates should be forbidden.

This leads to the fifth condition:

$$\forall v_m \in \mathcal{V}_m, v \in \mathcal{V}_m \cup \mathcal{V}_s | g(v_m) \cap g(v_s) \neq \emptyset \quad \text{and} \quad p(v_m) \cap p(v_s) = \emptyset, \quad \text{then} \quad v_m \nleftrightarrow v_s.$$

If we denote $\mathcal{E}_e^- = \{\{v_m, v\} | v_m \in \mathcal{V}_m, v \in \mathcal{V}_m \cup \mathcal{V}_s$ and $g(v_m) \cap g(v_s) \neq \emptyset$ and $p(v_m) \cap p(v_s) = \emptyset\}$, we can formulate the fifth condition as:

$$\mathcal{E}_e^- \cap \mathcal{E} = \emptyset.$$

Summary of conditions. At this point we can summarize our conditions to produce two sets of edges $\mathcal{E}^+ = \mathcal{E}_b^+ \cup \mathcal{E}_c^+ \cup \mathcal{E}_d^+$ and $\mathcal{E}^- = \mathcal{E}_a^- \cup \mathcal{E}_b^- \cup \mathcal{E}_c^- \cup \mathcal{E}_e^-$. If we denote $\mathcal{E}_{\text{all}} = \mathcal{V} \times \mathcal{V}$ the set of all possible edges on $\mathcal{V}$, the graph $\mathcal{I}(\mathcal{V}, \mathcal{E})$ must verify:

$$\mathcal{E}^+ \subset \mathcal{E} \subset \mathcal{E}_{\text{all}} - \mathcal{E}^-. \tag{1}$$

So far we have proven that a graph associated to the perfect drawing of a pedigree is an interval graph that respects equation (1).

Equivalence. Reciprocally if one finds an interval graph that respects equation (1), any of its interval realizations provides a means of obtaining a perfect drawing of the pedigree: indeed, the abscissae of an individual span on the drawing are given by those of its corresponding interval in the realization, while its ordinate depends only on its generation.

Therefore the PDP problem is equivalent to the problem of finding an interval graph whose set of edges is comprised between two nested sets. This is known as the interval graph sandwich (IGS) problem and has already been investigated elsewhere; in our case, a polynomial solution exists for which we refer the reader to (Belfer and Golumbic, 1990; Golumbic and Shamir, 1992; Belfer and Golumbic, 1993).

ALGORITHM FOR OPTIMIZING THE PEDIGREE DRAWING

We now use the equivalence between interval graphs and PDP to propose an algorithm for optimizing the pedigree drawing. Our algorithm proceeds in three steps as follows:

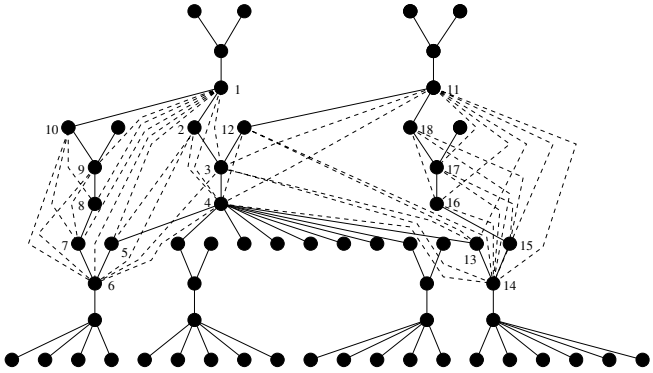

Fig. 6. This graph is derived from pedigree of Figure 2, as the graph in Figure 4, but has also been augmented with new edges to give an interval graph verifying equation (1). For the sake of readability, not all the new edges are represented: only those concerning the two cycles have been drawn.

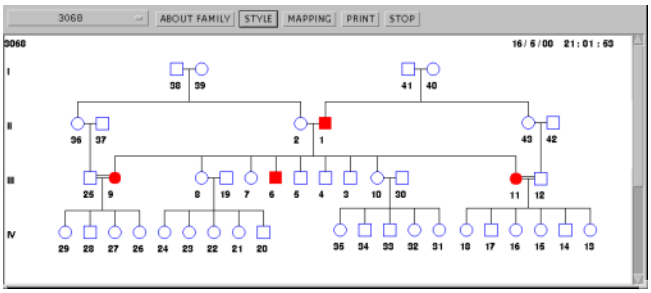

Fig. 7. Drawing of the same pedigree as in Figure 2, now deduced from the IG realization of Figure 6.

(1) derive $\mathcal{V}$, $\mathcal{E}^+$ and $\mathcal{E}^-$ from the pedigree to draw;

(2) solve the graph sandwich problem to find an interval graph $\mathcal{I}(\mathcal{V}, \mathcal{E})$ with $\mathcal{E}^+ \subset \mathcal{E} \subset \mathcal{E}_{\text{all}} - \mathcal{E}^-$;

(3) if such an interval graph exists, deduce a perfect drawing of the pedigree.

Step 1 is detailed in the subsection 'Equivalence between PDP and IGS' where references on how to carry out step 2 are also given. The Figure 6 shows the graph from Figure 4 as it would look like after nodes are ordered by the IGS algorithm from step 2.

To achieve step 3, one generates a set of intervals consistent with the interval graph (Booth and Lueker, 1976; Rose *et al.*, 1976; Lueker, 1975), for example by ordering the maximal cliques (a clique is a set of completely connected vertices) of the graph. Then the position of each interval associated to an individual gives the position of this individual, which allows an easy drawing of the pedigree.

CONCLUSION AND PROSPECTS

The problem of optimizing the drawing of a pedigree is not trivial. The existence of a perfect solution (with no link crossing) can be tested in a linear time. When a perfect solution exists, it can be found with the algorithm introduced in this paper which is based on the interval graph sandwich theory. We are now implementing such a method in the CoPE software for pedigree drawing (Brun-Samarcq *et al.*, 1999) and we are also studying the case of pedigrees that are not PDP.

If the IGS problem with $\mathcal{V}$, $\mathcal{E}^+$ and $\mathcal{E}_{\text{all}} - \mathcal{E}^-$ has no solution, the pedigree is not PDP and one can look for a solution that will minimize the number of link crossings. Several strategies exist. One can view the pedigree as a directed graph from top to bottom, and thus use one of the heuristics of line-crossing elimination (Di Batista *et al.*, 1998). In a work in progress, we are exploring two heuristics. The first one relies on the line-crossing elimination in directed graph, and genetic algorithms are used to find the correct node ordering in each layer. The second heuristic relies on the interval graph sandwich problem and is based on the augmentation and/or diminution of graph (which has already been implemented for another problem in biology; Whittaker *et al.*, 1993).

ACKNOWLEDGEMENTS

We thank Frédéric Guyon for stimulating discussions on the problem of pedigree drawing optimization and Laurence Brun-Samarcq for sharing her expertise on pedigree drawing. We are also grateful to Philippe Dessen and Guy Vaysseix for their help in this work.

REFERENCES

Belfer,A. and Golumbic,M.C. (1990) A combinatorial approach to temporal reasoning. In *Proceedings of the Fifth Jerusalem Conference on Information Technology*. IEEE Computer Society Press, pp. 774–780.

Belfer,A. and Golumbic,M.C. (1993) Counting endpoint sequences for interval orders and interval graphs. *Discrete Math.*, **114**, 23–29.

Bennett,R.L., Steinhaus,K.A., Uhrich,S.B., O'Sullivan,C.K., Resta,R.G., Lochner-Doyle,D., Markel,D.S., Vincent,V. and Hamanishi,J. (1995) Recommendations for standardized human pedigree nomenclature. *Am. J. Hum. Genet.*, **56**, 745–752.

Booth,K.S. and Lueker,G.S. (1976) Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. Comput. Syst. Sci.*, **13**, 335–379.

Brun-Samarcq,L., Gallina,S., Philippi,A., Demenais,F., Vaysseix,G. and Barillot,E. (1999) Cope: a collaborative pedigree drawing environment. *Bioinformatics*, **15**, 345–346.

Cherwell Scientific (2000) Cyrillic software. *Technical Report* http://www.cyrillicsoftware.com.

Di Batista,G., Eades,P., Tamassia,R. and Tollis,I. (1998) *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice-Hall, Upper Saddle River.

Epicenter Software (2000) Gap, Genetic analysis package. *Technical Report* http://icarcus2.hsc.usc.edu/epicenter.

Golumbic,M.C. and Shamir,R. (1992) Interval graphs, interval orders and the consistency of temporal events. *Lecture Notes in Computer Science, 601*. Springer-Verlag, New York.

Lueker,G.S. (1975) Efficient algorithms for chordal graphs and interval graphs, *PhD Thesis*, Program in Applied Mathematics and the Department of Electrical Engineering, Princeton University, Princeton, NJ.

Progeny Corp (2000) Progeny software. *Technical Report* http://www.progency2000.com.

Rose,D.J., Tarjan,R.E. and Lueker,G.S. (1976) Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.*, **5**, 266–283.

Whittaker,C.C., Mundt,M.O., Faber,V., Balding,D.J., Dougherty,R.L., Stallings,R.L., White,S.W. and Torney,D.C. (1993) Computations for mapping genomes with clones. *Int. J. Genome Res.*, **1**, 195–226.