Inference of Order in Genetic Systems *

John N. Guidi and Thomas H. Roderick
The Jackson Laboratory
600 Main Street
Bar Harbor, Maine 04609-1500
{jng,thr}@jax.org

Abstract

We survey and discuss issues required of intelligent systems to support research efforts in locus mapping. In particular we focus on the issues of order, on how one can automate the reasoning processes of ordering, and the database structures required to support orders, including ambiguity and uncertainty. We conclude with a summary of work to be done.

Introduction

The organization of genes in chromosomes in eukaryotes was one of the major discoveries in this century. The delineation of the ordering within species has led to insights into genetic regulation, genetic interaction (e.g., position effects), developmental mechanisms (e.g., X-inactivation and genomic imprinting), and genetic evolution (e.g., tandem duplication, polyploidy, and paralogous chromosomes and segments). It has also provided markers to follow other genetic defects (e.g., Huntington Disease) as well and tools for understanding the nature of defective genes and their clinical effects. In general the discovery of linkage has provided a fundamental basis for delineating, arranging and organizing all eukaryotic genes and DNA sequences as they are discovered and described. Furthermore, the conservation of linkage and sequences among even distantly related species has led to understanding of chromosomal evolution providing medical researchers with a vast number of nonhuman, mammalian, genetically homologous models that have major value in understanding the human problem, possibly leading to opportunities for gene therapy.

Mapping genes in humans and other mammals is now a major international objective in biomedical research. This has led to an exponentially increasing amount of linkage and sequence information. The storage and access of this vast amount of complex information is an essential component of this effort (Robbins 1992; Honda et al. 1993).

*This work was supported in part by National Institutes of Health grant HG00330.

A convenient summary of this information is in the form of physical or genetic maps. The entry of raw linkage or physical data needs to be planned carefully but is generally a trivial problem other than being labor intensive. The construction of consensus maps from numerous diverse studies with a variety of types of raw data is the more difficult problem. Experts generally agree on which kinds of data are the most important and which construction rules are relevant and the relative emphasis that should be assigned to each rule. Computer programs based on these rules and weighted relative to their importance are now essential to aid in building at least "first pass" consensus maps.

A number of techniques are available for determining order at different levels of resolution. Cytological methods can place a locus within a chromosome band. Linkage analysis of multi-point crosses can provide order and distance measurements of recombination frequency in centiMorgans (cM). Overlapping sets of clones of various sizes (e.g., YACs, cosmids, phages) can provide order by piecing together these overlapping intervals. The highest level of resolution is the DNA sequence itself.

Determining order is a fundamental problem in mapping. The orders and relative distances of loci may not be in total agreement when combining different experimental techniques or data sets. Except for the possibility of naturally occurring chromosomal inversions, we expect to find a single order within a species. With additional data, and with the certain introduction of new techniques to augment current mapping techniques, the order should eventually be resolved but differences in distances may still remain. For example, in linkage analyses, there are known chromosomal segments in several species that show unusually high recombination frequencies. Also, there exist regions in the genome with varying numbers of repetitive elements.

Consider the following example, which is illustrated in Figure 1. Given a single 2 point cross, one obtains a distance statistic that places one locus a, relative to another locus b, but order can not be determined with respect to the centromere or any other marker. For

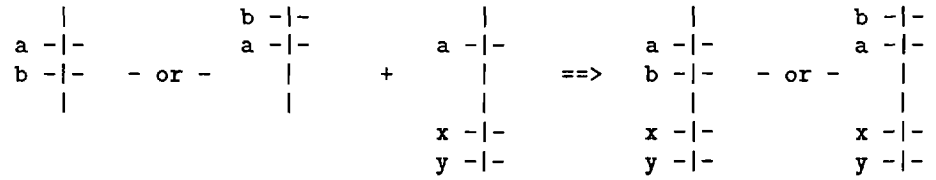


Figure 1: Deduction with Ambiguity

species with acrocentric chromosomes (e.g., mouse), the convention is that the centromere is placed at the top of diagrams. This information, although ambiguous with respect to order, can be combined with data from other experiments that determined the order and distance among loci a, x, and y, to deduce that $(b \prec x)$, and $(b \prec y)$.

As another example, consider the case of combining multiple distance measurements. It is often possible to infer order based solely on distances. Given the following genetic distance measurements, together with their uncertainties, satisfaction of the triangle inequality yields a partial order which is unique, except the orientation is unknown.

Given $|\ e-f\ |$ = 21.7±2.7 cM,
$|\ f-g\ |$ = 5.3±1.7 cM,
$|\ e-g\ |$ = 15.5±2.3 cM,

then the order must be either:
$(e \prec g \prec f)$ or
$(f \prec g \prec e)$.

It is imperative that representations of distances and order accommodate uncertainty and ambiguity. Given a technique, there can be experimental, as well as attribute specific differences. For example, with respect to genetic linkage, sex can have an effect on distance measurements. Sampling sizes and other experimental conditions may contribute to different laboratories making different, and sometimes contradictory, observations. All of these situations need to be represented directly, and in a manner which facilitates answering queries concerning order and distance.

The rate of locus mapping in the mouse has been increasing exponentially and the number of laboratories involved in mapping activities, either directly or indirectly, is increasing as well. The number of loci mapped in the mouse genome is over 4500 and the comparative map has 933 mapped mouse loci that have human homologs (Hillyard et al. 1992). Estimates are that there may be up to 100,000 genes in the mouse and human genomes. Even considering all the laboratory data, the volume of information involved in genomics is not in the terabyte per day range. Nevertheless, there are clearly significant amounts of mouse genome data that are yet to be collected, and the sizing issues are significant.

Issues

Research in genetics contributes in a fundamental way to our understanding of biology and nature, and the field is exploding with respect to the amount of information and the introduction of new experimental techniques. It is vital that intelligent systems provide natural representations for objects found in genetics. Limitations of current systems include lack of support for concepts of distance and order, inability to accommodate uncertainty of measurement and ambiguity of results, and poor support for handling derived data.

The concepts of order and distance are fundamental to genetics. The genetic objects involved can be considered as either atomic or composite, depending on context. For example, in cytological and linkage mapping, the resolution is such that a locus can be considered as an atomic entity. In sequencing experiments, however, a gene locus needs to be considered as a composite. At the molecular level, a gene is an aggregate of regulatory regions, introns and exons. Different levels of resolution influence how an object needs to be represented.

Distance is a n-ary metric that consists of application specific, and application independent properties. The semantics of the objects, the measurements, and the corresponding uncertainties are all application specific. However, classes of applications exhibit similar domain independent properties. For example, applications that observe the triangle inequality property can have their measurements manipulated algebraically in identical ways, regardless of the application area. Also, depending on context, it may or may not be meaningful to convert measurements from one metric to another. For example, conversion between centiMorgans and base pairs is problematic.

Order is a binary relation that consists of domain specific and domain independent properties. The relation partial order, with notation $\preceq$, on the set P has the the following properties (Davey & Priestly 1990):

reflexive - $(x \preceq x)$ for all x in P

anti-symmetric - $(x \preceq y)$ and $(y \preceq x)$ imply $(x = y)$ for all x, y in P

transitive - $(x \preceq y)$ and $(y \preceq z)$ imply $(x \preceq z)$ for all x, y, z in P

The set P is called a partially ordered set, or poset. The semantics of the operator $\preceq$ are domain specific. One semantic interpretation in genetics, which is useful for $\preceq$, is order with the centromere as origin and the poset P being the loci that have been mapped. For a set of loci for which not all orders are extensionally represented, the transitive property can be used to infer order.

A poset is said to be totally ordered, or a chain, if for all x, y either $(x \preceq y)$ or $(y \preceq x)$. A naive view would have the genomic maps of the mouse represented by a single chain where, for any two loci x, y, one of $(x = y)$, $(x \prec y)$, or $(y \prec x)$ would hold. Synonyms would account for $(x = y)$, otherwise, for every pair of loci, one would always either precede or succeed the other. This idealized situation does not occur. In many cases, the order among loci is unknown, can not be inferred, or considering the experimental error, can not be determined accurately.

Intervals ordered relative to one another can be used to model genetic loci, but there are some biological anomalies that need to be accommodated (Robbins 1993). Some examples in the mouse include loci that involve alternative splicing (*Evi-1* (Bordereaux et al. 1990), *Lyt-2* (recently renamed *Cd8a*) (Zamoyska et al. 1985)), nested genes (*Fim-2* spans *Csfmr* (Sola et al. 1988), U14 snRNA (renamed Rna14) is in hsc70 (renamed *Hsp70*) (Liu & Maxwell 1990)), and possibilities of a locus having more than one chromosomal site (Howe et al. 1979; Wettstein & Colombo 1987).

In addition, this is a field of tremendous activity, and as such the data are fluid. Because of this activity, and the fact that knowledge about genetics continually changes, it is possible for one laboratory to produce, say, an order which is contradictory with that obtained at another laboratory. Over time, depending on the loci in question, it is expected that with additional data, or with clarification of existing data, the differing orders will be resolved. But it is critical that these contradictions be made available in the database systems supporting this work.

Order theory provides a powerful tool to explore problems of order and distance in genetics, and existing work can be built upon to accommodate the situations outlined above. It is important that the tools developed are generic and can be used by genomes other than mouse, including non-linear genomes (e.g., the mitochondria, *E. coli*). Support for partial orders in database systems is required, and attention must be paid to finding data structures to efficiently support ordered data, taking into consideration uncertainty and ambiguity. The transitive property of the genetic posets permits inference. Even with uncertainty and ambiguity, useful inferences are still possible. Thus, in conjunction with work on storage structures, efforts on how best to compute transitive closures with these properties need to be done in parallel.

The integration of orders based on differing powers of resolution is required. Molecular biologists have recognized this problem and have made an effort to use Sequence Tag Sites (STS) to provide anchor points to connect physical orders obtained with, and within, different techniques (Olson et al. 1989). In addition, methods have been described to integrate information from genetic, cytological, and molecular maps (Collins et al. 1992). The ability to construct composite objects by integrating data that are similar, but obtained with different techniques, needs to be addressed.

Issues with respect to summary data are common in statistical and scientific data. Imprecise summaries are often acceptable when, for example, the volume of base data is too large, or the complexity of the base data prevents easy comprehension. The consensus map of the mouse produced by experts (Hillyard et al. 1992) is summary data, and in some cases, the mouse chromosome committees (Committees 1992) provide summary data as well. There are mouse chromosome committees that do not provide base data to support their summary maps. Also, the coupling of a scientific database system with analytical tools is a common situation. A number of analytical tools are available that will generate derived information with respect to order and distance. Thus, systems which support genetic orders need to address the issues of integration of base, summary and derived data.

Examples

An intelligent system for genetic researchers needs to answer a variety of questions. For example, with genetic information available from mouse and other species, the following queries are reasonable:

Q1: **Find all loci within 10cM of locus *a*.**
First, all base linkage data that are within 10cM of locus *a*, with appropriate uncertainties are obtained. Automated deduction, using transitivity, uncertainty, and distance metrics, derives additional loci from linkage information. Different levels of resolution are consulted. For example, the linkage data may support that locus *b* is within 10cM, and there may be physical evidence that gene *c* is close enough to *b* to warrant inclusion. Also, the summarized information from the chromosome committee may state that locus *d* is within 10cM of *a* without providing the base data to support this. Additionally, the potential that there are ambiguous data needs to be taken into account. The result is a list of possible loci with the likelihoods that they solve the query.

Q2: **What is the order for all loci between *a* and *b*?**
Likely orders with supporting evidence are displayed, in addition to less likely supported orders. The user must have the opportunity to include, or exclude, any data desired. As examples, it may be desired to use only base data for a specific sex,

or data derived from a particular analytical tool may need to be excluded.

Q3: **Given the region between a and b in the mouse, where do segments appear to be conserved in human?**
This query requires that all possible loci between a and b in the mouse are obtained, and used to generate a list of any known homologies in human. Possible partial orders of homologous human loci are then presented. Defining what is a homology is a complex scientific issue, and the user must be able to control the generation of the homology list based on the data, information about the experimental techniques involved, and personal preferences.

Q4: **I have determined order $a \prec b \prec c$. Is there any experimental, or derived information that contradicts this order?**
This requires generating possible partial orders from the database, and then determining if there are any that would contradict this assertion. This is a very common query in practice.

Related Work

Order is a fundamental concept in a variety of application areas. The representation, searching, and sorting of ordered structures is a core concept in computer science (Knuth 1973). Applications in process systems (Bergstra & Klop 1991; Poguntke 1986), decision theory (Fishburn 1989), planning and scheduling (Conway et al. 1967; Fishburn 1989; Degano et al. 1990), historical and temporal systems (Soo 1991; Tansel et al. 1993), and spatial systems (Egenhofer & Franzosa 1991; Egenhofer 1991; Frank 1991; Randell et al. 1992) are among the areas where considerable research has been done on order. The use of partial orders and lattices is natural in object oriented systems to represent class inheritance (Heuer & Sander 1991). A temporal logic approach to support lists (ordered sequences) in an object oriented data model has been described (Richardson 1992).

Fishburn discussed the mathematics of interval orders extensively (Fishburn 1985a; Fishburn 1985b). Ordered intervals have been used in a variety of applications. Interval graphs were used to represent DNA restriction maps (Waterman & Griggs 1986). Intervals are ubiquitous in database and reasoning systems that involve historical or temporal concepts (Allen 1983; Allen 1984; van Benthem 1989; Ladkin 1986a; Ladkin 1986b; McKenzie & Snodgrass 1991).

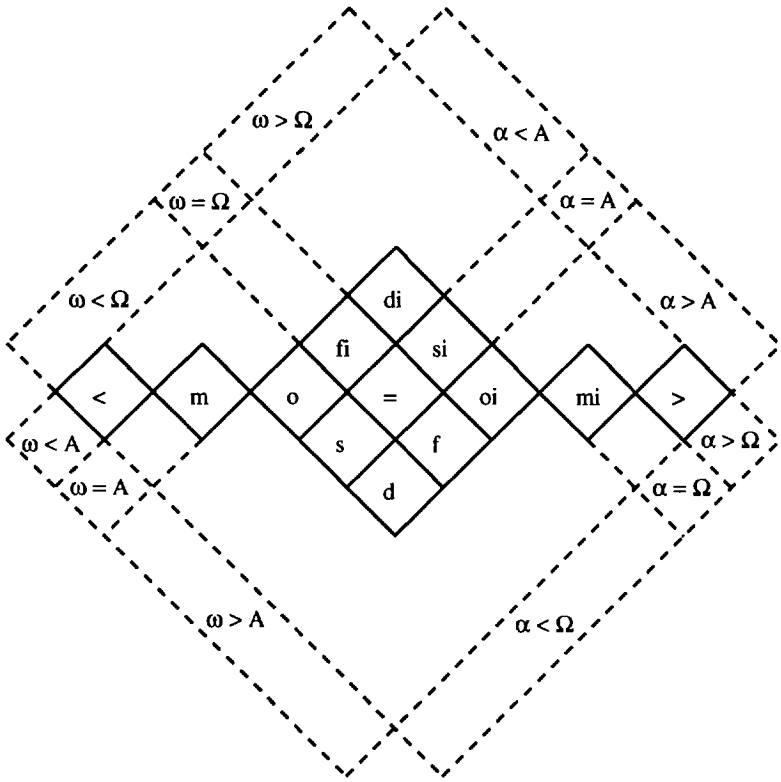
The transitive property of temporal intervals permits the computation of intensional orderings. This was explored in detail in a seminal paper (Allen 1983), that enumerated the transitivity operations, and described a constraint propagation algorithm for an interval based temporal logic. Building on this work, a calculus of generalized intervals, which are made of finite sequences of points in a linear order, was described (Ligozat 1991). Semi-intervals were described that examined neighborhoods of relationships based on beginnings and endings, and explored issues of compactness, symmetry, and redundancy (Freksa 1992). Dean and McDermott addressed nonmonotonic reasoning about time with incomplete information in a time map management system (Dean & McDermott 1987). Yang presented a conflict algebra to be used with constraint satisfaction problems (Yang 1990). van Beek and Cohen discussed approximate reasoning with temporal intervals (van Beek & Cohen 1990). Williams and Kong considered temporal incompleteness in a deductive database (Williams & Kong 1991).

It has been shown that the problem of finding a total ordering of a finite set of elements and a set of ordered triples of these elements is NP-complete (Opatrny 1979). Efforts to infer what is true over certain intervals of time, where the order of events is indeterminate, was shown to be NP-complete, but an inexact, though useful, answer can be obtained in polynomial time with an algorithm that computes sets of events preceding a given event (Dean & Boddy 1988). The intractability of Allen's interval algebra is discussed in (Vilian & Kautz 1986).

Recursive queries have been addressed by database researchers, including the use of transitive closure algorithms. A number of algorithms have been described to handle recursion in logic queries in databases (Bancilhon & Ramakrishnan 1986). An effective recursive query processing technique for deductive databases was offered by (Vieille 1989). Issues concerning query language support for graph traversal problems were discussed in (Mannino & Shapiro 1990). A number of efforts examined efficient computation of transitive closure (Ioannidis & Ramakrishnan 1988; Agrawal et al. 1990; Jagadish 1990). The access of recursively defined complex objects was addressed in (Shoning 1990).

Researchers in genetics have devised computer programs for specific applications to generate linkage maps from defined types of data. These include the work of our colleague J. H. Edwards whose program OM produced from all linkage data the distances between loci based on the maximum-likelihood method (Edwards 1987; Morton 1988; Edwards 1989; Edwards et al. 1993). Ott and colleagues provided a general purpose program called LINKAGE aimed at ordering loci from multi-locus crosses as well as handling specific traits from specific pedigrees (Lathrop et al. 1984; Lathrop & Lalouel 1988). The MAPMAKER program was produced and can be used for codominant, dominant or recessive traits in F2-type pedigrees and codominant traits in CEPH-type pedigrees (Lander et al. 1987).

A number of efforts have provided computer support for representation of ordered physical and linkage maps in genetics. A survey of linguistic approaches for sequence comparisons is available (Myers 1991). De-



symbol	relation
<	before
>	after
d	during
di	contains
o	overlaps
oi	overlapped by
m	meets
mi	met by
s	starts
si	started by
f	finishes
fi	finished by
=	equals

Figure 2: Interval Relations (after Freksa 1992)

ductive databases have been used to maintain DNA maps in partial digestion experiments (Tsur et al. 1990). Constraint propagation techniques have been used with order and distance data to identify conflicts, and for unambiguous data sets, to provide order (Letovsky & Berlyn 1992). A grammar has been produced that focused on producing a syntax to describe relationships among genetic objects (Pearson 1992).

Theory, Methods and Data

Work in temporal databases provides a starting point for an algebra that describes the properties for orders and distances of genetic objects. An analogy can be made between temporal and chromosome intervals and ordering within these. Points are modeled as intervals with the same lower and upper bound for end points. Figure 2, from (Allen 1983) and (Freksa 1992), illustrates the relationships between two intervals X, Y where X begins with α and ends with ω, and Y with A and Ω.

In has been shown that these relationships can all be described in terms of the *meets* relation (Allen & Hayes 1985). For example, consider two loci (X, Y) that are ordered $(X \prec Y)$. The satisfaction of $(X$ *before* $Y)$ is equivalent to stating that there exists some interval K such that $(X$ *meets* $K)$ and $(K$ *meets* $Y)$.

It is often the case that direct observation between two loci is not available. However, the transitive properties of partial orders can be used to deduce order. Allen describes in detail the computation of transitive closures using all combinations of the aforementioned interval relationships in figure 2 (Allen 1983) . Freksa details, using beginnings and endings of intervals, fine and coarse reasoning with incomplete knowledge (Freksa 1992) .

Besides transitive closure of ordering, additional inference rules have been documented that make use of order, distance, and uncertainty measurements (Letovsky & Berlyn 1992)). All these rules are sufficient for data sets where the stated and the derived orders have no contradictions. However, one absolutely must accommodate ambiguous orders. Even if there is overwhelming evidence for a particular order, evidence to the contrary needs to be retained. Mouse genetics is a particularly active field of investigation, with aggressive efforts to identify and map loci, and to develop new methods to optimally map the mouse genome. The data are thus fluid. In addition to experimental error, there may be valid scientific reasons why order ambiguities may exist (e.g., a rare inversion or transposition event may have occurred).

Requirements for storage structures demand that they: have an innate ability to present order relation-

ship between nodes; are able to accommodate ambiguity; include distance metrics, where available; include uncertainty measurements; have an indexing mechanism that supports the concept of "near" neighbors. As an example, with certain assumptions, PQ-trees provide an excellent data structure for representing sets with restricted permutations (Booth & Lueker 1976). Figure 3 illustrates a PQ-tree for a region of 8 genetic loci, where the neighbors within two localized subregions are known, but not the orientation (e.g., either $(C \prec G \prec H)$ or $(H \prec G \prec C)$, and either $(E \prec F)$ or $(F \prec E)$). P nodes are illustrated as circles and represent permutations. Q nodes are illustrated with bars and represent subsequences whose neighbors are known, but whose orientations are unknown. PQ-trees are limited in that they do not accommodate the concepts of distance, or likelihood. Also, there is no mechanism to represent ambiguous orders.

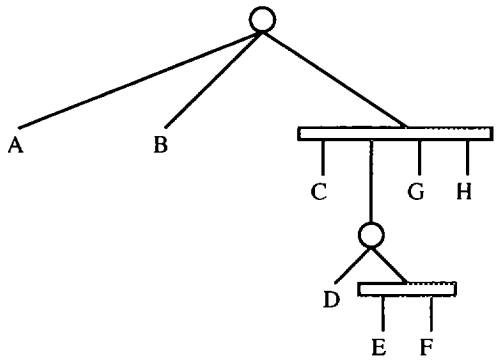


Figure 3: PQ-tree

It is anticipated that tools will be developed that implement the order algebra using efficient storage structures to accommodate ambiguity and uncertainty. Primary sources of data to be used with these tools are the Genomic Database of the Mouse (GBASE) and eventually the Mouse Genome Database (MGD), when MGD becomes available. GBASE information is comprised of the following major classes of data for each published work: In addition to a complete citation, there is a brief abstract of the paper, listing salient points concentrating on mapping information. The mapping data are entered for each pair of loci for each publication and include: chromosomal location of the loci, the type of cross (i.e. backcross, intercross, etc.), whether the data are from recombinant inbred lines, whether it is parasexual data for a given locus, and whether data concern in situ hybridization or molecular sequencing data for a given locus. The phase of the cross is given (i.e., whether in coupling or repulsion), the sex of the segregating parent, the number of recombinants for all two-loci linkages, the total number observed, percent recombination calculated appropriately for the type of cross, the standard error of the percent recombination, the inverse of the variance, and the recombination rate times the inverse of the variance (used in combining data for a given pair of loci over all studies).

Additional information from The Jackson Laboratory's GBASE and other mouse databases will provide the foundation for the Mouse Genome Database (MGD), which subsumes the others. In addition, MGD will provide expanded coverage. The current MGD schema also provides for: data from seven major categories of experiments (e.g., experimental crosses, recombinant inbred strains, congenic strains, somatic cell hybridization (including somatic cell, microcell and radiation hybrids), in situ hybridization, chromosomal aberrations, and preliminary data).

Research in Progress

We are producing an algebra that describes ordered objects and the operations permitted on these objects. This will provide a formal description of our concepts of order and distance, and act as the foundation for our future efforts. In essence, our algebra will define our query language. We generalize the objects we are concerned with as continuous intervals, which can themselves be aggregates of intervals (van Benthem 1983). As a starting point, our algebra is based on interval relationships and the computation of incomplete inferences (Allen 1983; Freksa 1992; Shoham 1988). We will support additional inference rules over these intervals that take into account order, distance metrics, and uncertainty (Letovsky & Berlyn 1992), but importantly, we will define the rules so they support ambiguity. We are providing a number of operators in our algebra, including transitive closure of order, precedence, distance manipulations, neighborhood, and position. All of these operators will be defined to take into consideration uncertainty and ambiguity. The operators will also support the various biological anomalies of the genetic objects being modeled that prevent use of existing temporal models (Robbins 1993). We will support the concepts of orientation and polarity, both among and within objects, with these operators.

References

Agrawal, R.; Dar, S.; and Jagadish, H. V. 1990. Direct Transitive Closure Algorithms: Design and Performance Evaluation. *ACM Transactions on Database Systems* 15(3):427-458.

Allen, J.F. 1983. Maintaining Knowledge about Temporal Intervals. *Communications of the Association of Computing Machinery* 26(11):832-843.

Allen, J. F. 1984. Towards a General Theory of Action and Time. *Artificial Intelligence* 23(2):123-154.

Allen, J.F.; and Hayes, P.J. 1985. A Common-Sense Theory of Time. In Proceedings of the International Joint Conference on Artificial Intelligence, 528-531. Los Angeles, CA: Morgan Kaufmann.

Bancilhon, F.; and Ramakrishnan, R. 1986. An Amateur's Introduction to Recursive Query Processing Strategies. In Proceedings of the ACM SIGMOD '86, 16–52.

Bergstra, J. A.; and Klop, J. W. 1991. An Introduction to Process Algebra. In *Applications of Process Algebra*, 1-21. Vol. 17 of Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Booth, K.S.; and Lueker, G.S. 1976. Testing for the Consecutive Ones Property, Interval Graphs, and Graph Planarity Using P-Q Tree Algorithms. *Journal of Computer and System Sciences* 13:335-379.

Bordereaux, D.; Fichelson, S.; Tambourin, P.; and Gisselbrecht, S. 1990. Alternative splicing of the Evi-1 zinc finger gene generates mRNAs which differ by the number of zinc finger motifs. *Oncogene* 5:925-927.

Collins, A.; Keats, B. J.; Dracopoli, N.; Shields, D. C.; and Morton, N. E. 1992. Integration of gene maps: Chromosome 1. *Proceedings of the National Academy of Sciences USA* 89:4598-4602.

Committees, Mouse Chromosome 1992. Encyclopedia of the Mouse Genome II. *Mammalian Genome* 3:Special Edition.

Conway, R. W., Maxwell, W. L., and Miller, L. W. 1967. *Theory of Scheduling.* Addison-Wesley.

Davey, B. A., and Priestly, H. A. 1990. *Introduction to Lattices and Order.* Cambridge University Press.

Dean, T.; and McDermott, D. V. 1987. Temporal Data Base Management. *Artificial Intelligence* 32:1-55.

Dean, T.; and Boddy, M. 1988. Reasoning about Partially Ordered Events. *Artificial Intelligence* 36:375-399.

Degano, P.; de Nicola, R.; and Montanari, U. 1990. A partial ordering semantics for CCS. *Theoretical Computer Science* 75:223-262.

Edwards, J. H. 1987. The locus ordering problem. *Annals of Human Genetics* 51:251-258.

Edwards, J. H. 1989. The locus positioning problem. *Annals of Human Genetics* 53:271-275.

Edwards, J. H., Hillyard, A. L. and Roderick, T. H. 1993. The OM Linkage Program. Forthcoming.

Egenhofer, M. 1991. Reasoning about Binary Topological Relations. In Advances in Spatial Databases. 2nd Symposium, SSD'91, 143-160. Zurich, Switzerland: Springer-Verlag.

Egenhofer, M. J.; and Franzosa, R. D. 1991. Point-set topological spatial relations. *International Journal of Geographical Informations Systems* 5(2):161-174.

Fishburn, P. C. 1985a. Interval graphs and interval orders. *Discrete Mathematics* 55:135–149.

Fishburn, P. C. 1985b. *Interval orders and interval graphs.* John Wiley & Sons, Inc.

Fishburn, P.C. 1989. Human decision making and ordered sets. In Proceedings of the NATO Advanced Study Institute on Algorithms and Order, 437–465. Vol. 255 of NATO ASI series. Series C, Mathematical and Physical Sciences. Ottawa, Canada: Kluwer Academic Publishers.

Frank, A. U. 1991. Qualitative Spatial Reasoning about Cardinal Directions. In Technical Papers 1991 ACSM-ASPRS Annual Convention. Auto-Carto 10, 148-167.

Freksa, C. 1992. Temporal reasoning based on semi-intervals. *Artificial Intelligence* 54:199-227.

Heuer, A.; and Sander, P. 1991. Preserving and Generating Objects in the LIVING IN A LATTICE Rule Language. In Proceedings Seventh International Conference on Data Engineering, 562-569.

Hillyard, A. L.; Doolittle, D. P.; Davisson, M. T. and Roderick, T. H. 1992. Locus Map of Mouse with Comparative Map Points of Human on Mouse. The Jackson Laboratory, Bar Harbor, ME 04609.

Honda, S.; Parrott, N. W.; Smith, R.; and Lawrence, C. 1993. An Object Model for Genome Information at All Levels of Resolution. In Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences, 564-573. IEEE Computer Society Press.

Howe, R. C.; Ahmed, A.; Faldetta, T. J.; Byrnes, J. E.; Rogan, K. M.; Dorf, M. E.; Taylor, B. A.; and Humphreys, R. E. 1979. Mapping of the Lyb-4 Gene to Different Chromosomes in DBA/2J and C3H/HeJ Mice. *Immunogenetics* 9:221-232.

Ioannidis, Y. E.; and Ramakrishnan, R. 1988. Efficient Transitive Closure Algorithms. In Proceedings of the Conference on Very Large Databases, 382-394. Los Angeles, California: Morgan Kaufmann.

Jagadish, H. V. 1990. A Compression Technique to Materialize Transitive Closure. *ACM Transactions on Database Systems* 15(4):558–598.

Knuth, D. E. 1973. *The Art of Computer Programming, Vol. 3: Sorting and Searching.* Addison-Wesley.

Ladkin, P. 1986a. Primitives and Units for Time Specification. In Proceedings AAAI-86 5th National Conference on Artificial Intelligence, 354-359. Philadelphia, PA: Morgan Kaufmann.

Ladkin, P. 1986b. Time Representation: A Taxonomy of Interval Relations. In Proceedings AAAI-86 5th National Conference on Artificial Intelligence, 360-366. Philadelphia, PA: Morgan Kaufmann.

Lander, E. S.; Green, P.; Abrahamson, J.; Barlow, A.; Daly, M. J.; Lincoln, S. E.; and Newburg, L. 1987. MAPMAKER: An Interactive Computer Package for Constructing Primary Genetic Linkage Maps of Experimental and Natural Populations. *Genomics* 1:174–181.

Lathrop, G. M.; Lalouel, J. M.; Julier, C.; and Ott, J. 1984. Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences USA* 81:3443–3446.

Lathrop, G. M.; and Lalouel, J. M. 1988. Efficient computations in multilocus linkage analysis. *American Journal of Human Genetics* 42:498–505.

Letovsky, S.; and Berlyn, M. B. 1992. CPROP: A rule-based program for constructing genetic maps. *Genomics* 12:435–446.

Ligozat, G. 1991. On generalized interval calculi. In Proceedings, Ninth National Conference on Artificial Intelligence, 234–240. American Association for Artificial Intelligence. AAAI Press/The MIT Press.

Liu, J.; and Maxwell, E. S. 1990. Mouse U14 snRNA is encoded in an intron of the mouse cognate hsc70 heat shock gene. *Nucleic Acids Research* 18(22):6565-6571.

Mannino, M. V.; and Shapiro, L. D. 1990. Extensions to Query Languages for Graph Traversal Problems. *IEEE Transactions on Knowledge and Data Engineering* 2(3):353-363.

McKenzie, L.E., Jr.; and Snodgrass, R.T. 1991. Evaluation of Relational Algebras Incorporating the Time Dimension in Databases. *ACM Transactions on Database Systems* 23(4):501-543.

Morton, N. E. 1988. Multipoint mapping and the emperor's clothes. *Annals of Human Genetics* 52:309–318.

Myers, E. W. 1991. An Overview of Sequence Comparison Algorithms in Molecular Biology, Technical Report, TR 91-29. The University of Arizona.

Olson, M.; Hood, L.; Cantor, C.; and Botstein, D. 1989. A Common Language for Physical Mapping of the Human Genome. *Science* 245:1434-1435.

Opatrny, J. 1979. Total Ordering Problem. *SIAM Journal of Computing* 8(1):111-114.

Pearson, P. L. 1992. GDB Map Grammar - Present and Future (Personal Communication - May 17).

Poguntke, W. 1986. Order-Theoretic Aspects of Scheduling. In Combinatorics and Ordered Sets, 1-32. Vol. 57 of Contemporary Mathematics. American Mathematical Society.

Randell, D. A.; Chui, Z.; and Cohn, A. G. 1992. A Spatial Logic based on Regions and Connection. In Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning, 165-176. Cambridge, MA: Morgan Kaufmann.

Richardson, J. 1992. Supporting Lists in a Data Model (A Timely Approach). In Proceedings of the 18th International Conference on Very Large Data Bases, 127-138. Vancouver, British Columbia: Morgan Kaufmann.

Robbins, R. J. 1992. Challenges in the Human Genome Project. *IEEE Engineering in Medicine and Biology* 11(1):25-34.

Robbins, R. J. 1993. Representing Genomic Maps in a Relational Database. In *Computational Methods in Genome Research*, Plenum Publishers. Forthcoming.

Shoham, Y. 1988. *Reasoning about Change*. The MIT Press.

Shoning, H. 1990. Integrating Complex Objects and Recursion. In Deductive and Object-Oriented Databases. Proceedings of the First International Conference on Deductive and Object-Oriented Databases (DOOD89). December 4-6, 1989, 573-592. Kyoto Research Park, Kyoto, Japan: North-Holland.

Sola, B.; Simon, D.; Mattéi, M.-G.; Fichelson, S.; Bordereaux, D.; Tambourin, P. E.; Guenet, J.-L.; and Gisselbrecht, S. 1988. *Fim-1*, *Fim-2/c-fms*, and *Fim-3*, Three Common Integration Sites of Friend Murine Leukemia Virus in Myeloblastic Leukemias, Map to Mouse Chromosomes 13, 18, and 3, Respectively. *Journal of Virology* 62(11):3973-3978.

Soo, M. D. 1991. Bibliography on Temporal Databases. *SIGMOD Record* 20(1):14-23 (this is also available via anonymous FTP from cs.arizona.edu in the bibliographies in bib/*).

Tansel, A. U., Clifford, J., Gadia, S. K., Jajodia, S., Segev, A., and Snodgrass, R. 1993. *Temporal Databases: Theory, Design, and Implementation*. The Benjamin/Cummings Publishing Company.

Tsur, S.; Olken, F.; and Naor, D. 1990. Deductive Databases for Genomic Mapping (Extended Abstract). In North American Conference on Logic Programming (NACLP) - Workshop on Applications of Deductive Databases. Austin, TX: The MIT Press.

van Beek, P.; and Cohen, R. 1990. Exact and approximate reasoning about temporal intervals. *Computational Intelligence* 6(3):132-144.

van Benthem, J. F. A. K. 1983. *The Logic of Time.* Vol. 156 of Synthese Library. D. Reidel Publishing Company.

van Benthem, J. 1989. Time, Logic, and Computation. In Linear Time, Branching Time, and Partial Order in Logics and Models for Concurrency, 1-49. Vol. 354 of Lecture Notes in Computer Science. Springer-Verlag.

Vieille, L. 1989. Recursive Query Processing: The Power of Logic. *Theoretical Computer Science* 69:1-53.

Vilian, M.; and Kautz, H. 1986. Constraint Propagation Algorithms for Temporal Reasoning. In Proceedings AAAI-86 5th National Conference on Artificial Intelligence, 377-382. Philadelphia, PA: Morgan Kaufmann.

Waterman, M. S.; and Griggs, J. R. 1986. Interval Graphs and Maps of DNA. *Bulletin of Mathematical Biology* 48(2):189–195.

Wettstein, P. J.; and Colombo, M. P. 1987. Immunodominance in the T Cell Response to Multiple Non-H-2 Histocompatibility Antigens. IV. Partial Tissue Distribution and Mapping of Immunodominant Antigens. *The Journal of Immunology* 139(7):2166-2171.

Williams, M. H.; and Kong, Q. 1991. Time and Incompleteness in a Deductive Database. In *Uncertainty in Knowledge Bases. Proceedings of the 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 443-455. Vol. 521 of Lecture Notes in Computer Science. Paris, France: Springer-Verlag.

Yang, Q. 1990. An Algebraic Approach to Conflict Resolution in Planning. In Proceedings Eight National Conference on Artificial Intelligence, 40-45. St. Paul, MN: Morgan Kaufmann.

Zamoyska, R.; Vollmer, A. C.; Sizer, K. C.; Liaw, C. W.; and Parnes, J. R. 1985. Two Lyt-2 Polypeptides Arise from a Single Gene by Alternative Splicing Patterns of mRNA. *Cell* 43:153-163.