Uniform Integration of Genome Mapping Data using Intersection Graphs

Eric Harley, *Department of Computer Science, University of Toronto, Toronto, Ont, Canada M5S 1A4*

Anthony Bonner, *Department of Computer Science, University of Toronto, Toronto, Ont, Canada M5S 1A4*

Nathan Goodman, *Bioinformatics Consultant, Brookline, MA 02445-2115*

Running Head: Uniform Integration of Mapping Data

Key words: genome, physical map, overlap data, landmark

Abstract

Motivation: The methods for analyzing overlap data are distinct from those for analyzing probe data, making integration of the two forms awkward. Conversion of overlap data to probe-like data elements would facilitate comparison and uniform integration of overlap data and probe data using software developed for analysis of STS data.
Results: We show that overlap data can be effectively converted to probe-like data elements by extracting maximal sets of mutually overlapping clones. We call these sets *virtual probes*, since each set determines a site in the genome corresponding to the region which is common among the clones of the set. Finding the virtual probes is equivalent to finding the maximal cliques of a graph. We modify a known maximal-clique algorithm such that it finds all virtual probes in a large dataset within minutes. We illustrate the algorithm by converting fingerprint and Alu-PCR overlap data to virtual probes. The virtual probes are then analyzed using double-linkage intersection graphs and structure graphs to show that methods designed for STS data are also applicable to overlap data represented as virtual probes. Next we show that virtual probes can produce a uniform integration of different kinds of mapping data, in particular STS probe data and fingerprint and Alu-PCR overlap data. The integrated virtual probes produce longer double-linkage contigs than STS probes alone, and in conjunction with structure graphs they facilitate the identification and elimination of anomalies. Thus, the virtual-probe technique provides (*i*) a new way to examine overlap data, (*ii*) a basis on which to compare overlap data and probe data using the same systems and standards, and (*iii*) a unique and useful way to uniformly integrate overlap data with probe data.

Availability: Freely available on request.
Contact: Eric Harley `eharley@cs.toronto.edu`

Introduction

Strategies for constructing contig maps of genomes can be classified as probe-based or overlap-based. In probe-based methods, typified by STS-content mapping, the elements being mapped are of two sorts – clones and probes – and the mapping data indicate which clones contain which probes. In overlap-based methods, typified by restriction digest fingerprinting, the only elements being mapped are clones, and the mapping data simply indicate which pairs of clones overlap. The techniques for analyzing these two kinds of mapping data are quite different. Systems for analyzing probe data include Mott et al., 1993; Cuticchia et al., 1993; Wang et al., 1994; Soderlund and Dunham, 1995; Green and Green, 1991; Magness and Green, 1996; Nadkarni et al., 1996; Harley et al., 1998. Systems for overlap data include Soderlund et al., 1997; Gillett et al., 1996; Fonstein and Haselkorn, 1995; Whittaker et al., 1993. In projects where probe and overlap data are combined, they are typically examined separately, as in the map-assembly methods used by the Centre d'Etude du Polymorphisme Humain and Généthon (CEPH/ Généthon) and the Whitehead Institute/MIT Center for Genome Research (WI/MIT) in their seminal human mapping efforts (Chumakov et al., 1995; Hudson et al., 1995; Schuler et al., 1996).

In this paper, we describe a uniform method for analyzing probe-based and overlap-based data using the concept of virtual probe. A virtual probe is the region of the genome shared by a maximal set of mutually overlapping clones. Finding virtual probes is equivalent to finding the maximal cliques in a graph whose nodes represent clones and whose edges indicate which pairs of clones overlap. Clique-finding is a classic problem in graph theory. It is well-known that enumeration of maximal cliques is exponential in the size of the graph (because a graph may contain an exponential number of maximal cliques (Moon and Moser, 1965)); the related problem of finding the largest clique in a graph is NP-complete (Garey and Johnson, 1979). We use a variant of the Bron-Kerbosch algorithm for finding maximal cliques (Bron and Kerbosch, 1973) which we adapt to run efficiently on the large, sparse graphs that arise in contig mapping.

Our method is intended to help human experts cope with the numerous errors encountered in typical mapping datasets. The method suppresses local detail to help analysts focus on the global structure of the data. We create a view of the data in which good data appear as straight paths, while errors show up as branched structures. The result is that data can quickly be inspected for anomalies before applying algorithms or software packages whose primary function is to find the most likely order of probes and a corresponding placement of the clones. We have previously described these methods for probe data produced through STS-content mapping (Harley et al., 1998). Here we adapt the methods for overlap data produced through restriction digest fingerprinting and similar mapping procedures.

We illustrate the method using data from the CEPH/ Généthon and WI/MIT mapping projects mentioned above. We convert the CEPH/ Généthon fingerprint and Alu-PCR overlap data into virtual probes using our variant of the Bron-Kerbosch algorithm. Next we analyze the virtual probe data using double-linkage intersection graphs and structure graphs. Double-linkage is a filter which diminishes the branching effects caused by chimeras and false positives, while structure graphs tend to suppress local non-linearities caused by false negatives. These quality controls were designed for analysis of STS-content data, but they apply equally well to overlap data that have been converted into virtual probes. Finally, we integrate the STS-content data from these projects with the virtual probe data and show that the combined dataset produces longer double-linkage contigs than STS data alone.

System and Methods

Source of data: Our primary source of STS data is Release 10 (May, 1996) of the WI/MIT dataset available by anonymous ftp (genome.wi.mit.edu, directory /pub/ human_STS_releases/may96). Our source for fingerprint and Alu-PCR data along with additional STS data is the March 1995 Release of the CEPH/ Généthon dataset available by anonymous ftp (ceph-genethon-map.cephb.fr, /pub/ceph-genethon-map/STS/ 29MAR95.DAT). These datasets were used by their respective organizations to construct maps of the human genome (Hudson et al., 1995; Chumakov et al., 1995; and Schuler et al., 1996).

Though these datasets have been supplanted for most practical purposes by the more recent BAC map produced by Washington University (http://genome.wustl.edu/ gsc/index.shtml), they remain invaluable resources for development of map construction methods. Key advantages include: 1) The earlier datasets are well characterized. 2) They are large enough to reveal important problems, but are much smaller than the BAC dataset. 3) They combine data from multiple mapping methods, with enough data from each method to allow investigation of methods individually and in various combinations.

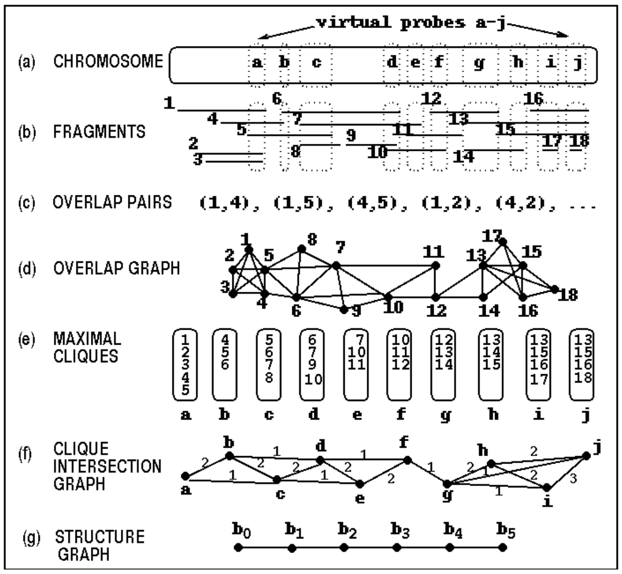

Figure 1: Clique model for analysis of overlap data.

Overlap: For the purposes of this paper, two clones A and B are inferred to overlap and form an unordered *overlap pair* based on fingerprint data when the CEPH/Généthon file `RELATIONS` lists clone B among the clones overlapping with clone A (and vice versa). The fingerprint data give rise to 49,383 overlap pairs among 31,392 YACS (yeast artificial chromosomes). Overlap of clones A and B is also inferred if one of the two clones, acting as an Alu-PCR probe, hybridizes with the other at a unique address, i.e., the hybridization is scored 'Unique' (U) by CEPH/Généthon. The CEPH/Généthon data show the results for experiments with 8,785 YACs used as Alu-PCR probes and 24,576 YACs used as targets. These Alu-PCR data produce 55,891 distinct overlap pairs.

A third method of inferring overlap of clones A and B is by observing that they both contain the same STS probe. Using STS probe data from WI/MIT (12,527 STSs and 21,051 YACs) and CEPH/Généthon (7,026 STSs and 18,298 YACs) we derived 364,536 overlap inferences, comprising 215,397 distinct overlap pairs. Overlaps inferred from the STS data corroborate 54% of the overlaps from the fingerprint data and 37% of the overlaps from the Alu-PCR data. The number of distinct overlap pairs in all three forms of data (STS, fingerprint, and Alu-PCR) combined is 271,257.

Overlap graph: This is a straightforward representation of overlap data, where nodes correspond to clones, and an edge means that the corresponding clones overlap. An example of a simple overlap graph is shown in Figure 1. Part (a) of this Figure shows a chromosome which is the source of a library of overlapping cloned fragments of DNA. Part (b) shows numbered cloned fragments directly under their region of origin on the chromosome. We assume that an experiment determines which fragments overlap. Some of the resulting overlap data are listed in part (c), which shows pairs of identifiers representing overlapping clones. For the purpose of this sketch, we assume that the experiment is perfect, i.e., all of the overlaps in the sketch result in observed overlap pairs, and there are no false-positive pairs. In this case, the sketch in (b) implies a unique set of overlaps in (c), although the converse is not true. Part (d) represents the overlap data as an overlap graph. Each edge (a,b) in this graph is in one-to-one correspondence with an overlap pair (a,b) in part (c).

Cliques: A *clique* is defined to be a subgraph in which each pair of nodes is joined by an edge, while a *maximal clique* is a clique which is not a proper subgraph of another clique (Harary, 1969). We will sometimes use the term clique to mean maximal clique, where the intended meaning is clear. Figure 1(e) shows groups of nodes which form maximal cliques in the example overlap graph. A number of heuristic algorithms for finding maximal cliques have been developed, and comparative reviews may be found in (Johnston, 1976; Pardalos and Xue, 1994; and Pardalos et al., 1999).

Experimental comparisons among these algorithms generally involve only a subset of the published algorithms, and are typically done on random graphs of less than 1000 nodes. In this work, we want to find all of the cliques in overlap graphs of about 30,000 nodes. Since these graphs are nonrandom and much larger than the graphs used in published studies comparing clique-listing algorithms, we implemented and compared several of the algorithms: the long-standing Bron and Kerbosch algorithm (BK) (Bron and Kerbosch, 1973), the simplified Bron and Kerbosch algorithm (SBK) (Johnston, 1976), an algorithm based on finding maximal independent sets (LTMIS) (Loukakis and Tsouros, 1981) and an algorithm (CN) with time complexity O(a(G)m) per clique, where a(G) is the arboricity of the graph G, and m is the number of edges (Chiba and Nishiszeki, 1985). We find that the time taken by LTMIS or CN grows too quickly with the size of the overlap graph, making these algorithms inappropriate for large sparse graphs of this type. The BK and SBK algorithms are practical (assuming we use an adjacency list rather than a square matrix to represent the graph) but take about an hour (99 and 49 minutes, respectively) of computation time on an overlap graph of 25,000 nodes, using a Sun Sparc Station 10. However, we modify the BK algorithm to the effect that the computation time on a graph of this size is reduced to

29 seconds. In essence, the modification applies the BK algorithm to find the maximal cliques in each subgraph S_i, $(1 <= i <= N)$, where S_i is composed of node i and its neighbors, and the maximal cliques involve only vertex i and its neighbors j such that $j > i$. (N denotes the number of nodes in the graph). C code for each of the algorithms mentioned is available from the authors on request.

Virtual-Probes: Notice that each maximal clique corresponds to a region of mutual overlap among its constituent clones, as illustrated by the dotted boxes superimposed on the clones sketched in Figure 1(b). In turn, each of these regions represents a site in the genome, as indicated by the placement of the clique labels on the chromosome in Figure 1(a). We call these regions *virtual probes* since (like STS probes) they mark sites in the genome, but (unlike STS probes) they are identified indirectly by groups of overlapping clones rather than by sequence. Each virtual probe has breadth equal to the width of the region of mutual overlap. In theory, the genomic region of one virtual probe cannot overlap with that of another, since this would imply that one maximal clique was a subset of another — a contradiction of terms.

Weighted intersection graphs: These are a well-studied class of graphs (Harary, 1969) where the nodes correspond to sets, and an edge of weight M between two nodes means that there are M elements in the intersection of the corresponding sets. We have previously discussed *probe intersection graphs*, where the node-set represents a group of clones which hybridize with a common probe (Harley et al., 1998). In this paper we focus on *clique intersection graphs*, where each node-set represents a maximal clique in the overlap graph, and a clique defines a virtual probe. Figure 1(f) shows the weighted intersection graph using labels a,b, ..., j for the cliques found in Figure 1(e). For example, the edge from node a to node b is shown with a weight of 2, since cliques a and b have two elements in common: fragments 4 and 5.

In all subsequent graphs of this paper, edges of unit weight are excluded to produce a *double-linkage* intersection graph. The double-linkage filter greatly reduces the number of edges attributable to false positives and the effects of chimerism (Arratia et al., 1991). Another filter for false positives is the size of cliques. A clique of size n in the overlap graph represents $\frac{n(n-1)}{2}$ overlapping pairs of clones, and each overlap corroborates the other overlaps. Cliques of just two clones of course lack this internal corroboration, and therefore we ignore them in order to filter out potential false overlaps.

Structure graphs: These provide a skeletal view of intersection graphs by compressing local complexity to reveal the underlying structure (Harley et al., 1998). The structure graph is formed on the basis of two breadth-first search (BFS) traversals of the intersection graph. The first BFS traversal starts at an arbitrary node s and identifies a node x which is any of the nodes farthest from s. The second BFS traversal starts at node x and partitions the nodes of the graph into layers according to their distance from x. Nodes which form a connected component within a given BFS layer are defined to form a *blob*, which becomes a node in the structure graph. Two nodes a, b in the structure graph, corresponding to blobs A, B in the intersection graph, are joined by an edge (a, b) if there is an edge (i, j) in the intersection graph, where $i \in A$ and $j \in B$.

For example, if the first BFS starts at node f in the graph of Figure 1(f), then we obtain $x = a$, as the node most distant from f. The second BFS then starts at node a and divides the graph into layers: $l_0 = \{a\}, l_1 = \{b, c\}, l_2 = \{d, e\}, l_3 = \{f\}, l_4 = \{g\}, l_5 = \{h, i, j\}$. Each layer is further subdivided into connected components called *blobs*. In this simple (noise-free) example each layer forms a single connected component, so that each layer is a blob: $b_i = l_i$. We form a structure graph by creating a node for each blob and an edge between pairs of blobs if in the intersection graph there is an edge between an element of one blob and an element of another. For example, there will be an edge between b_0 and b_1, since $a \in b_0$ is connected to $b \in b_1$ in Figure 1(f). The structure graph resulting from this example is shown in Figure 1(g). It is a simple path, as will be the case whenever the data is perfect (i.e., when the data can be modeled by a set of overlapping line segments), according to a theorem in (Harley et al., 1999). Note that in general, we form structure graphs from double linkage intersection graph, but for illustration purposes we did not remove the edges of unit weight in this example.

Graph visualization: We use the Hy+ data visualization system (Consens, 1994) to display intersection and structure graphs.

Chromosome assignment: We used the following rule to assign virtual probes to chromosomes: *Assign a virtual probe to Chromosome A if the number of its clones associated with Chromosome A is at least two and greater than the number associated with any other chromosome.* A clone is defined to be associated with chromosome c if it hybridizes with an STS assigned to chromosome c. Reliability and applicability of this rule were assessed on sets of clones defined by hybridization with STSs, since in that case the chromosome assignment of the STSs can be used as a standard. The assignment rule was applicable in 90% of the cases and

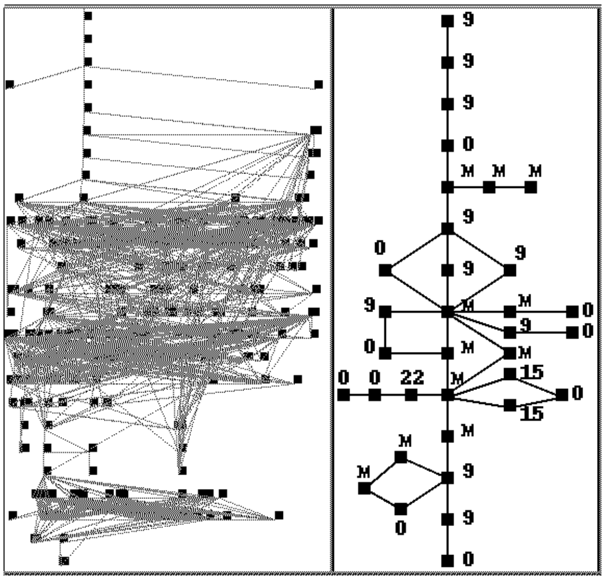

Figure 2: Alu-PCR virtual probes: *(left)* Double-linkage intersection graph (238 nodes, 1,805 edges); *(right)* structure graph (35 nodes, 40 edges).

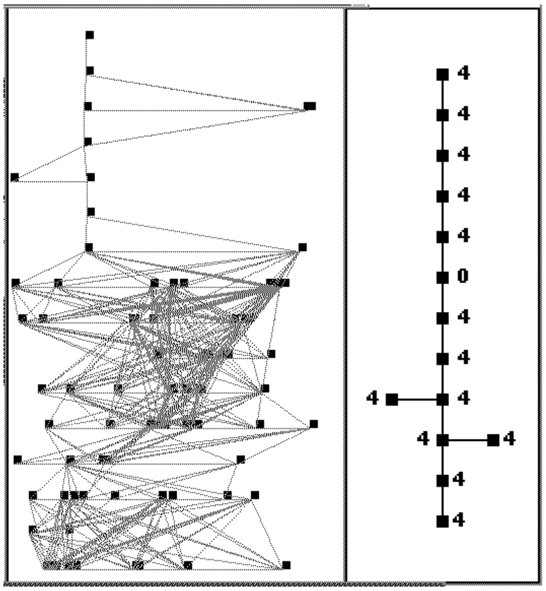

Figure 3: Fingerprint virtual probes: *(left)* Double-linkage intersection graph (82 nodes, 592 edges); *(right)* structure graph (15 nodes, 14 edges).

was correct 97% of the time (sample size 15,136).

Results

Alu-PCR data: We briefly examine CEPH/Généthon Alu-PCR data using the virtual-probe method to show that this method provides a useful and informative view of Alu-PCR data in isolation. We applied our modified BK algorithm to extract all of the maximal cliques from an overlap graph of Alu-PCR data. The total number of maximal cliques was 34,643, ranging in size from 2 to 14 YACs. Cliques of size two (20,285) were subsequently ignored as unreliable and not useful for double-linkage graphs. Cliques were assigned to chromosomes according to the rule described above. This resulted in unambiguous assignment of 87.7% of the cliques, while 3.6% of the cliques could not be assigned because they were equally associated with more than one chromosome, and 8.7% were not assigned because they did not have at least two YACs associated with a single chromosome.

We formed an intersection graph from the 14,358 cliques of size three or more. One huge, connected component comprised 96% of the nodes, an anomaly common in STS data at this level of analysis (i.e., single-linkage). This clearly indicates cross-linkage among virtual probes belonging to different chromosomes. Simple counting reveals that 19% of the unit-weight edges connect virtual probes assigned to different chromosomes (considering only edges between virtual probes that were successfully assigned to chromosomes). Only about 1.2% of the nonunit-weight edges are cross-links of this type, which leads us to examine only the double-linkage intersection graph.

The largest component of the double-linkage clique-intersection graph is shown in Figure 2, along with the corresponding trimmed structure graph. The nodes in the structure graph are labeled using the chromosome-assignment rule applied to blobs. Nodes which are not assigned a chromosome because they are equally associated with multiple chromosomes are labeled 'M', while blobs that do not have at least two clones associated with a single chromosome are labeled '0'. The labeled structure graph suggests that this component is a contig from Chromosome 9, along with some cross links to parts of chromosome 15 and 22. We do not analyze this component further, since the point here is only to show that the virtual probe technique makes possible new ways to examine overlap data. The information in this figure would be difficult to obtain by conventional means.

Fingerprint data: In this section, we use virtual probes to form a new view of the fingerprint data from CEPH/Généthon. We first extracted all of the maximal cliques from the overlap graph with the our modified BK algorithm. The total number of maximal cliques was 12,710, including 3,718 cliques of size two which were discarded as potential noise. Using

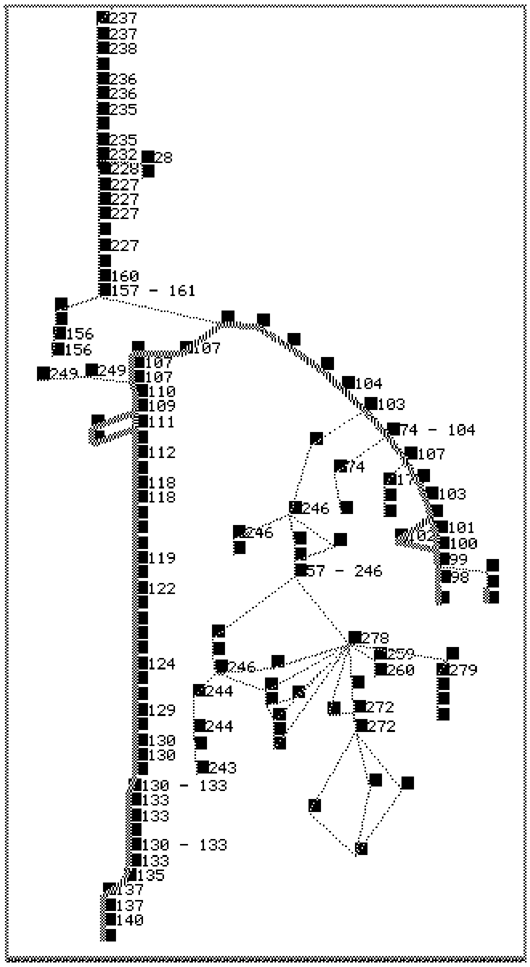

Figure 4: Structure graph (134 nodes, 165 edges) for the largest component (1,196 nodes, 5,839 edges) of the double-linkage intersection graph resulting from virtual-probe analysis of Chromosome 1. Blobs are labeled with genetic positions in cM. A proposed contig is outlined in bold.

the chromosome-assignment rule, 93% of the remaining cliques were unambiguously assigned to chromosomes. We formed an intersection graph based on cliques from fingerprint data, and the largest component comprised 66% of the nodes. This indicates considerable cross-linkage, and in fact 14% of the unit-weight edges link virtual probes belonging to different chromosomes. On the other hand, 0.8% of the nonunit-weight edges are cross-links, leading us as usual to use the double-linkage filter. Figure 3, shows the largest component of the double-linkage intersection graph and the corresponding structure graph. Nodes in the structure graph are labeled using the chromosome-assignment rule applied to blobs. This component appears to be a fairly clean and simple contig from Chromosome 4.

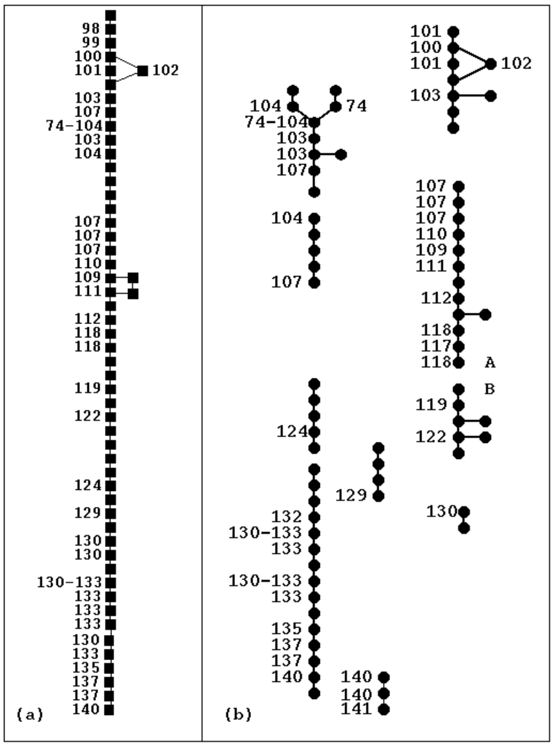

Figure 5: *(a)* Proposed contig extracted from integrated data; *(b)* corresponding STS contigs.

This example illustrates that the virtual probe technique makes possible simple noise filtration and analysis of fingerprint overlap data by methods developed for probe data.

Merged data: In this section we use virtual probes to integrate STS, Alu-PCR and fingerprint data. The STS data was first converted to overlap data by assuming that the clones hit by a particular STS mutually overlap, and then we combined the STS overlap data with the 'pure' overlap data. The total number of maximal cliques generated using the our modified BK algorithm on the integrated overlap data was 40,060, including 12,378 of size two that were discarded as potential noise. The number of cliques from the merged data decreases rapidly with clique size, though not as quickly as in the case of cliques from Alu-PCR or fingerprint data. Cliques of size 12 or more are primarily based on STS data. Application of the chromosome-assignment rule to cliques of size three or more resulted in 95% of the cliques being unambiguously assigned. The clique intersection graph is largely one component containing 99.5% of the nodes, indicating abundant cross-linkage. In fact, 58% of the unit-weight edges cross-link virtual probes from different chromosomes, while 2.6% of the nonunit- weight edges are cross-links.

Removing the unit-weight edges, we obtained a double-linkage clique intersection graph for the genome-wide data, but even this graph is almost entirely one component, comprising 24,304 of the 27,692 nodes. This reflects considerable cross linkage for edges with weight $M \geq 2$, a phenomenon we also see in STS data alone (Harley et al., 1998). This connected component can be broken up according to chromosome by assigning virtual probes to chromosomes (as explained in the Methods section) and then filtering out edges which link probes belonging to different chromosomes. This results in a number of connected components for each chromosome. The structure graph for the largest connected component formed from virtual probes assigned to Chromosome 1 or unassigned is shown in Figure 4. The corresponding intersection graph is omitted from this figure to save space. The structure graph is branched and looped, indicating that this component cannot be a single contig.

We have labeled nodes in the structure graph with genetic positions in order to help visualize where the contigs lie in this component. (Labeling with radiation hybrid positions produces similar results). A blob is given a range of positions as determined by the STSs that it contains. A blob is considered to contain an STS if the set of clones in the blob is a superset of the set of clones hit by the STS. The longest potential contig stretches from about 98 cM to 140 cM. This contig is shown in Figure 5(a). Figure 5(b) shows that many double-linkage contigs in STS data alone are required to cover this range of genetic positions. Similarly, the WI/MIT STS-based map of the human genome shows about a dozen double-linkage contigs in the range 98-140 cM (http://carbon.wi.mit.edu:8000/cgi-bin/contig/phys_map). In the region 80-140 cM of Chromosome 1, the *double-linkage* coverage in our integrated data is similar in extent to that of the *single-linkage* STS contigs in the WI/MIT map. Thus, data integration using virtual probes leads to longer double-linkage contigs.

Discussion

We have presented a new method to analyze and integrate overlap data with probe data. The method is based on the conversion of overlap data to maximal cliques. The cliques correspond to sites in the genome, and are therefore called virtual probes. Analysis of the overlap graph in terms of virtual probes offers the following advantages — (1) it provides simple ways to filter out many false positives and chimeric links: simply ignore cliques of size two and construct double-linkage intersection graphs; (2) software systems designed for analysis of probe data can be applied without modification to overlap data transformed to virtual probes; and (3) virtual probes make possible a seamless integration of numerous forms of overlap and probe data. Separate analysis of these two different data types by the same system presents a unique opportunity for comparing probe and overlap results based on a single standard. Integrated analysis of the two data types by the same system increases the amount of data available to the system from which to make inferences, resulting in larger contigs and possibly more accurate ordering of probes, cliques and clones.

We converted Alu-PCR and fingerprint overlap data (separately) to sets of virtual probes to illustrate that these forms of pure overlap data can then be analyzed by methods developed for probe data. Alu-PCR data is typically considered in conjunction with other kinds of data, and in a way which makes its contribution secondary. For example WI/MIT uses Alu-PCR data as auxiliary evidence of connection between double-linkage STS contigs, and CEPH/Généthon uses Alu-PCR data along with fingerprint evidence as one way to construct overlapping clone paths between STS markers. In either case, it is not clear to what extent the Alu-PCR data can be trusted, nor is it clear what the Alu-PCR data looks like as a whole by itself. The result in Figure 2 shows that noise filtration and graphical methods developed for probe data can be applied to Alu-PCR data to produce new and interesting views of the data. The results in Figures 2 and 3 together indicate that analysis based on virtual probes offers a way to compare different types of overlap data by a common measure.

We used virtual probes to provide a uniform integration of several types of data (Alu-PCR, fingerprint and STS data). Analysis of the integrated data in one region of Chromosome 1 showed that the resulting double-linkage contigs were much more extensive than those obtained using STS data alone. Detailed analysis (not shown) reveals that STS contigs are sometimes joined via double-linkage through cliques that are based on a mixture of data types. Thus, contigs are automatically extended as a consequence of our general approach, without introducing separate algorithms or rules for this purpose.

Extraction of these contigs from complex graphical components requires additional positional evidence such as genetic or radiation hybrid markers. Once contigs are extracted, an ordering algorithm can be applied to the virtual probes, although this is not done here. The resulting order should be more accurate when using virtual probes of integrated data than when using STS data alone, since more information is associated with each probe. We postulate that the method of virtual probes will also be useful in sequence assembly where the problem of forming contigs based on overlap data recurs.

We have used structure graphs and their visual representation to allow rapid inspection of the data. Branching in the structure graph reveals anomalies which may be the result of chimeric YACs, repeat regions or other causes. The biologist can isolate the data representing the few blobs in the region of the fork, and examine this data separately. It might turn out that further experimentation is necessary to determine which YACs or probes involved in the fork are the source of the anomaly, or it may be possible to make a deduction based on the graph alone.

Future work: Analysis of simulated data shows that the ratio of (a) the number of cliques that would result given a typical rate of false negatives in the observation of overlaps to (b) the number of cliques that would result given observations without false negatives grows exponentially with increasing coverage. We are currently developing algorithms to counter this effect.

Throughout this paper we use only double-linkage intersection graphs. This ignores a great many edges that have unit weight, leading one to wonder if we might be losing useful information. One of the benefits of integrating data by using virtual probes is that good single-linkage data from several sources may combine to form usable double-linkage data. Experiments (not shown) designed to extract proximity information from the remaining unit-weight edges were unsuccessful. (However, one can do the reverse — use known proximity information to extract good single-YAC links.) This line of research could be explored more in the future, and might prove fruitful on other datasets.

Acknowledgements:

We gratefully acknowledge the assistance and expert advice of Lincoln Stein at Cold Spring Harbor Laboratory and Alberto Mendelzon at the University of Toronto.

References

Arratia,R., Lander,E.S., Tavare,S., and Waterman,M.S. (1991) Genomic mapping by anchoring random clones: A mathematical analysis. *Genomics* 11: 806–827.

Bron,C. and Kerbosch,J. (1973) Finding All Cliques of an Undirected Graph. *Communications of the ACM*, 16: 575-577.

Chumakov,I.M., Rigault,P., Le Gall,I., Bellanné-Chantelot,C., Billault,A., Guillou,S., Soularue,P., Guasconi,G., Poullier,E., Gros,I., *et al.* (1995) A YAC contig map of the human genome. *Nature, Suppl.* 377: 175-298.

Chiba,N. and Nishizeki,T. (1985) Arboricity and Subgraph Listing Algorithms. *SIAM J. Comput.*, 14: 210-223.

Consens,M. (1994) *Creating and Filtering Structural Data Visualizations using Hygraph Patterns.* Ph.D. Thesis, University of Toronto.

Cuticchia,A.J.,

Arnold,J., and Timberlake,W.E. (1993) ODS: ordering DNA sequences — a physical mapping algorithm based on simulated annealing. *CABIOS* 2: 215-219.

Dib,C., Sabine,F., Fizames,C., Samson,D., Drouot N., Vignal,A., Millaseau,P., Marc,S., Hazan,J., Seboun,E., Lathrop,M., Gyapay,G., Morissette,J., and Weissenbach,J. (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380: 152-154.

Fonstein,M. and Haselkorn,R. (1995) Physical Mapping of Bacterial Genomes. *Journal of Bacteriology* 177: 3361-3369.

Garey,M.R. and Johnson,D.S. (1979) *Computers and Intractability.* Freeman and Company, San Francisco.

Gillett,W., Hanks,L., Wong,G.K., Yu,J., Lim,R. and Olson,M.V. (1996) Assembly of High-Resolution Restriction Maps Based on Multiple Complete Digests of a Redundant Set of Overlapping Clones. *Genomics* 33: 389-408.

Green,E.D., and Green,P. (1991) Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods Appl*, 2: 77-90.

Harary, F. (1969) *Graph Theory.* Addison-Wesley Publishing Company, Reading, Mass.

Harley,E., Bonner,A.J., and Goodman,N. (1996) Good Maps Are Straight. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, 88-97.

Harley,E., Bonner,A.J., and Goodman,N. (1998) Revealing Hidden Interval Graph Structure in STS-Content Data. Bioinformatics 15: 278-285.

Hudson,T.J., Stein,L.D., Gerety,S.S., Ma,J., Castle,A.B., Silva,J., Slonim,D.K., Baptista,R., Kruglyak,L., Xu,S.H., *et al.* (1995) An STS-Based Map of the Human Genome. *Science* 270: 1945-1954.

Johnston,H.C. (1976) Cliques of a Graph — Variations on the Bron-Kerbosch Algorithm. *International Journal of Computer and Information Sciences* 5: 209-238.

Loukakis,E. and Tsouros,C. (1981) A Depth First

Search Algorithm to Generate the Family of Maximal Independent Sets of a Graph Lexicographically. *Computing*, 27: 249-266.

Magness,C. and Green,P. (1996) SEGMAP User's Manual, Version 3.48. Copyright 1994-1996.

Moon,J.W. and Moser,L. (1965) On cliques in graphs. *Israel J. Math* 3: 23-28.

Mott,R., Grigoriev,A., Maier,E., Hoheisel,J., and Lehrach,H. (1993) Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of Schizosaccharomyces pombe. *Nucleic Acids Research* 8: 1965-1974.

Nadkarni,P.M., Banks,A., Montgomery,K., LeBlanc-Stracewski,J., Miller,P., and Krauter,K. (1996) CONTIG EXPLORER: Interactive Marker-Content Map Assembly. *Genomics* 31: 301-310.

Pardalos,P.M. and Xue,J. (1994) The maximum clique problem. *Journal of Global Optimization* 4: 301-328.

Pardalos,P.M., Bomze,I.M., Budinich,M. and Pelillo,M. (1999) The Maximum Clique Problem. in *Handbook of Combinatorial Optimization, Supplement, Vol. A* (Eds: Du,D. and Pardalos,P.M.), Kluwer Academic Publishers: 1-74.

Schuler,D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tome,P., Aggarwal,A., Bajorek,E., *et al.* (1996) Gene Map of the Human Genome. *Science* 274: 540-546.

Schuler,D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tome,P., Aggarwal,A., Bajorek,E., *et al.* (1996) Gene Map of the Human Genome. *Science* 274: 540-546.

Soderlund,C. and Dunham,I. (1995) SAM: A system for iteratively building marker maps. *CABIOS* 11:645-655.

Soderlund,C., Longden,I., and Mott,R. (1997) FPC: a system for building contigs from restriction fingerprinted clones. *CABIOS* 13:523-535.

Walter,M.A., Spillett,D.J., Thomas,P., Weissenbach,J., and Goodfellow,P.N. (1994) A method for constructing radiation hybrid maps of whole genomes. *Nature Genetics* 7: 22-26.

Wang,Y., Prade,R.A., Griffith,J., Timberlake,W.E., and Arnold,J. (1994) ODS_BOOTSTRAP: assessing the statistical reliability of physical maps by bootstrap resampling. *CABIOS* 6: 625-634.

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. *International Journal of Genome Research* 1: 195-226.